EYE TRACKING: A PROMISING MEANS OF TRACING, EXPLAINING, AND PREVENTING THE EFFECTS OF DISPLAY CLUTTER IN REAL TIME

by

Nadine Marie Moacdieh

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2015

Doctoral Committee:

Professor Nadine B. Sarter, Chair
Associate Professor Eytan Adar
Professor Yili Liu
Associate Professor Bernard J. Martin

© Nadine Marie Moacdieh

2015

DEDICATION

To Pep and Mem

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help and support of a number of people. First and foremost, I would like to thank my advisor, Dr. Nadine Sarter, for her constant guidance and support throughout these five years. I am infinitely grateful for the time that she devoted to help me grow as a researcher and her encouragement and motivation every step of the way. The lifelong lessons in perseverance and dedication that I learned from her will stand me in good stead in my future career.

I would like to thank the members of my dissertation committee, Dr. Eytan Adar, Dr. Yili Liu, and Dr. Bernard Martin, for their time, support, and valuable feedback during the course of this dissertation.

I would also like to thank Dr. Benjamin Bassin, Dr. David Somand, and Dr. Travis Ganje at the University of Michigan (UM) Emergency Department (ED) for allowing me to develop parts of this project in the ED environment. I am very grateful for their assistance in developing medical scenarios and for granting me access to the ED environment. I would similarly like to extend my thanks to Cindy Trafford for her help in coordinating our meetings, as well as all the medical personnel in the ED for their participation in the related experiments.

I am grateful to Rackham Graduate School, the Industrial and Operations Engineering (IOE) department, and the Federal Aviation Administration for providing financial support during my time at UM.

I am also indebted to the amazing staff in the IOE department for all their help. In particular, I would like to thank Olof (Mint) Minto, Eyvind Claxton, Chris Konrad, and Chuck Woolley for providing technical assistance on countless occasions. Candy Ellis, Amy Warhaft, and Rick Krause have likewise been extremely helpful. In addition, I am very thankful to Matt Irelan, Tina Sroka, Gwen Brown, Liz Fisher, Wanda Dobberstein, and Rod Capps for making everything operate smoothly and always being ready to provide assistance. I would also like to thank all the College of Engineering undergraduate and graduate students who participated in my experiments. In particular, I am extremely grateful to Vernnaliz Carrasquillo and Yadrianna Acosta-Sojo for their willingness to participate in my experiments and their patience throughout.

I can never thank the current and previous members of the THInC Lab enough for everything they have done throughout these years. The programmers and undergraduate students who were part of the lab all provided invaluable support. I am particularly grateful to Joe Phillips for all his help in different projects, as well as Noah Klugman, Khevna Shah, and Chihiro Jimbo. Particular thanks goes to Nick Cruz for his amazing ability to grasp programming languages and work so meticulously. As for the PhD students in the THInC Lab – Huiyang Li, Sara Lu Riggs, Brandon Pitts, Julie Prinet, and Yuzhi Wan – they have been a never-ending source of support, feedback, laughs, comfort, and friendship. I do not think I could have asked for a more amazing group of people to share a lab with. Whether flying to conferences together, walking to the Dude for coffee, finding surprises at my desk when I least expect them, playing soccer, receiving comforting reminders to *press on,* or learning about Chinese cuisine, the times that we shared helped make my years at UM truly memorable.

Many heartfelt thanks as well to Joelle Hayek, Sandy Elkhoury, and Noura Farra – friends who were far away but nonetheless close in spirit at all times.

Last but not least, I would like to thank my amazing family for their constant love, support, and prayers. I especially want to thank my parents, Pep and Mem, my brothers, Emile, Sami, and Munir, my grandparents Teta Salwa, Teta Souad, and Jeddo Mounir (who I know is very happy and proud at this moment), and my uncles, aunts, and cousins both in the US and Lebanon, especially Aida, Freddy, Grace, Layla, Leena, Neda, Ramzi, Randa, and Tante Samia. My family has been with me every step of the way and this journey would not have been possible without them. I am infinitely grateful for their messages of encouragement, loving emails, surprise packages, and long Skype calls, as well as their timely reminders of what is most important… *Soli Deo gloria.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

Introduction

The amount of digital data available today is unprecedented and projected to grow 50-fold by 2020 (Gantz & Reinsel, 2012). Increasingly, people are faced with the problem of separating important, task-relevant information from a large quantity of unnecessary, irrelevant data on their displays (Cukier, 2010; Doyon-Poulin, Robert, & Ouellette, 2012). This struggle has led to concerns about display clutter, a loosely-defined concept that most often refers to excess data. Clutter is detrimental to visual search and the extraction of information in a variety of displays, from websites (e.g., Grahame, Laberge, & Scialfa, 2004) and maps (e.g., Hegarty, De Leeuw, &, Bonura, 2008) to flight deck displays (Alexander, Stelzer, Kim, & Kaber, 2008) and electronic medical records (Singh, Spitzmueller, Petersen, Sawhney, & Sittig, 2013). More specifically, clutter leads to delays in searching for critical or task-relevant information as well as misses, or failures to notice important information on a display.

In some domains, such delays and misses can pose a significant threat to the efficiency and safety of operations, as evidenced by numerous incidents and accidents. For example, in 1979, when a stuck-open valve at the Three Mile Island nuclear facility allowed large amounts of nuclear reactor coolant to escape, the operators were unable to quickly diagnose the problem and prevent a partial nuclear meltdown (President's Commission on the Accident at Three Mile Island, 1979). They subsequently cited the large amount of data in the control room as one of the main reasons why they could not quickly locate the source of the failure (Lusted, 2012). The crash of Air France Flight 447 in 2009 is another example of clutter-induced difficulties (Bureau d'Enquetes et d'Analyses, 2011). Following an unexpected autopilot disconnect, incorrect responses of the pilot resulted in an aerodynamic stall from which the flight crew did not recover. The immense quantity and density of data and alarms that were presented during the event rendered it extremely difficult to identify the underlying cause of the problem (Creedy, 2011). Finally, in 2009, in the medical domain, display clutter contributed to the transplant of a kidney infected with hepatitis (Hamill, 2011). None of the physicians involved noticed the alert to the positive hepatitis C test within their medical displays, an oversight that was later attributed to the large amount of irrelevant and fragmented data in the system (McMullen, 2011). These breakdowns in attention and performance highlight the importance of addressing the problem of clutter, especially in domains where the issue is compounded by high stress. Stress, by itself, is known to cause significant performance decrements (Staal, 2004) and may interact with clutter to exacerbate its effects on attention and information acquisition.

The overall goals of this dissertation were thus to 1) analyze and better understand the effects of clutter, in combination with stress, on information acquisition, 2) detect these effects of clutter as early as possible in real time, and 3) trigger real-time display adjustments to prevent performance breakdowns . The following sections will provide an overview of the current knowledge base in clutter and highlight gaps in that literature. Eye tracking will then be introduced as a promising approach to trace these effects of clutter and provide the basis for detecting and overcoming the problem of clutter in real time.

The Problem of Display Clutter

Although researchers have yet to reach consensus on a precise definition of clutter, there is agreement that it can significantly degrade performance in a variety of domains. Clutter has been shown to cause distraction (Beck, Trenchard, van Lamsweerde, Goldstein, & Lohrenz, 2012) and uncertainty (Bravo & Farid, 2008; Ewing, Woodruff, & Vickers, 2006; Lohrenz, Layne, Edwards, Gendron, & Bradley, 2006; Schons & Wickens, 1993), as well as increase memory load (Westerbeek & Maes, 2011). More importantly, the concern about clutter relates to its detrimental effects on visual search (e.g., Beck, Lohrenz, & Trafton, 2010; Henderson, Chanceaux, & Smith, 2009; Neider & Zelinsky, 2011) and noticing (Bravo & Farid, 2006; Ewing, Woodruff, & Vickers, 2006; Peng, Ward, & Rundensteiner, 2004; Schons & Wickens, 1993; Yeh & Wickens, 2001).

A Primer on Visual Search and Noticing

People rely on visual input to guide their everyday actions, from searching for a piece of clothing to noticing a flashing alert. Yet what guides visual selection remains open to debate (Theeuwes, 2010). As a result, the topic of visual attention capture has been the subject of extensive research over many decades (e.g., Jonides & Gleitman, 1972; Neisser, 1967; Treisman & Gelade, 1980). A complete review of this literature is outside the scope of this dissertation (for a thorough review of attention capture, see Theeuwes (2010)). However, some of the fundamental concepts related to visual selection and noticing need to be introduced to explain how clutter can disrupt these processes.

A major concept related to attention capture is top-down (or endogenous) and bottom-up (or exogenous) control of attention (Burnham, 2007; Theeuwes & Belopolsky, 2010). Top-down control refers to voluntary, goal-driven attention allocation, whereas bottom-up control refers to automatic and involuntarily attention capture by some salient feature in the environment (Rauschenberger, 2003). Salience here refers to features such as unique color, luminance changes, or abrupt onsets (e.g., Reynolds & Chelazzi, 2004). Both top-down and bottom-up attention control play a significant role in information acquisition, although the extent of their respective contribution and interaction remains a source of discussion (Theeuwes, 2010).

The distinction between top-down and bottom-up processing serves to define the two processes of interest in this research: visual search and noticing. Visual search can be defined as the process of actively looking for and locating a given target whose location is unknown (Treisman, 1991). Noticing, on the other hand, refers to "the conscious registration of some event" (Schmidt, 1995; p. 29) without being given a specific target. As such, visual search is an example of top-down attention allocation whereas noticing relies more strongly on bottom-up attention capture.

Several theories of visual search have been proposed (see Eckstein (2011) or Wolfe (2003) for reviews). According to one widely accepted view, namely Feature Integration Theory (FIT; Treisman & Gelade, 1980), visual search follows a two-stage process. It starts with a very short *preattentive* stage during which salient items, such as one red shape surrounded by green shapes, capture attention in a bottom-up fashion, leading to very fast search if the salient item matches the target. Otherwise, a second *attentive* stage will follow during which serial processing of all items takes place until the target is found. Here, the user's goals and knowledge of the target inform the search process in a top-down fashion. This second stage tends to take longer and is a function of the number of items, also referred to as the set size (Palmer, 1994), that need to be processed. The so-called *set size effect* is defined as the increase in search time as the set size increases (i.e., the slope of search time vs. set size is positive; Wolfe, 1994). Several theories of visual search have been developed since the emergence of FIT (e.g., Itti & Koch, 2001; Wolfe, 1994), but the basic assumption of a two-step process remains largely in place.

In addition to set size, several other factors can affect response time in a visual search task. Crowding is one such factor (e.g., Levi, 2008; Pelli, Palomares, & Majaj, 2004; Tullis, 1983; Vlaskamp & Hooge, 2006). It refers to the close spacing between a target and surrounding distractors which makes it difficult to isolate the features of the target (van den Berg, Cornelissen, & Roerdink, 2009; Whitney & Levi, 2011). Another factor that leads to increased search time is a high degree of similarity between the target and surrounding distractors (e.g., Findlay & Gilchrist, 2005; Henderson, 2007).

Finally, clutter – the focus of this dissertation – is another factor that has been consistently linked to decrements in visual search and noticing (e.g., Bravo & Farid, 2006; Neider & Zelinsky, 2011). However, the lack of an agreed-upon definition of clutter represents one challenge for making progress towards a better understanding of the phenomenon and its effects on attention and performance. Therefore, the development of a comprehensive and ergonomics-oriented definition of clutter was a necessary early step in this dissertation research. This was achieved by means of a systematic review of the literature on clutter that was published in the *Human Factors* journal (Moacdieh & Sarter, 2014).

Literature Review of Clutter

The literature review of clutter consisted of two sections. The first addressed the problem of obtaining a comprehensive definition of clutter that would be of most use to the human factors community. The second section described the merits and disadvantages of the different approaches to measuring clutter. The focus here will be on the first part, and several of the findings related to definitions of clutter will be reported.

This review was conducted using an iterative approach. As a first step, a search was performed in several scientific databases (Google Scholar, ScienceDirect, Web of Knowledge, JSTOR, ACM Digital Library, IEEE Xplore, and SciVerse Scopus), using key words and phrases such as clutter, data/information overload, and display/visual density. The goal was to identify seminal papers and the most relevant studies related to display clutter. As a next step, the same search was performed in a number of prominent human factors and applied psychology journals or conference proceedings (e.g., *Human Factors*, *Proceedings of the Human Factors and Ergonomics Society Annual Meetings*, *Attention, Perception, and Psychophysics*, *Journal of Experimental Psychology: Human Perception and Performance* and *Applied*, *Vision Research*, and *Journal of Vision*). In this case, the search was limited to publications during the last 15 years, with the goal to get a complete overview of the most recent work on display clutter.

Given the deliberately broad nature of this review, articles from a wide variety of domains, such as pattern recognition, optics, and signal processing were included. Articles were incorporated in the review if they evaluated clutter in digital displays, including maps, natural scenes, cockpit displays, television screens, etc. Other forms of clutter, such as roadside clutter, were excluded. Research on related concepts such as complexity (Donderi, 2006; Fadiran, Molnar, & Kaplan, 2006) was not included unless the concept was explicitly used in connection with display clutter. The search process initially yielded 269 papers. Of these papers, 76 matched all of the inclusion criteria and were ultimately used in the review to obtain a better definition of clutter.

Defining display clutter. The difficulty with agreeing on a definition of clutter is partly due to the fact that the phenomenon has been studied in many different domains, ranging from aviation (Kaber et al., 2013), radar systems (e.g., McKenzie, Wong, & Gibbins, 2013), and healthcare (e.g., Hammond et al., 2012) to marketing (Rotfeld, 2006) and website design (Grahame et al., 2004). The resulting disagreement about what constitutes clutter is reflected in the different approaches to measuring it, which include image processing, performance evaluation, and/or subjective evaluation as assessment techniques. In the case of image processing algorithms, clutter is calculated based on pixel-based display characteristics, such as luminosity or contrast (e.g., Bravo & Farid, 2008; Chang, Zhang, Liu, Yang, & Li, 2010; Lohrenz, Trafton, Beck, & Gendron, 2009; van den Berg et al., 2009). For example, Rosenholtz, Li, and Nakano (2007) developed a measure of clutter based on using three different metrics: feature congestion, edge density, and subband entropy. Feature congestion refers to the idea that the more cluttered a display is, the more difficult it would be to add an item that would be able to capture attention. First, clutter maps are developed separately for color, texture, and orientation, and then they are combined into a single clutter map and provide a scalar measure of clutter. Edge density is based on the idea by Mack and Oliva (2004) that clutter is related to the number of edges in a display. In this case, a filter calculates the density of edge pixels as a percentage of the total number of pixels to give a measure of clutter. Finally, the subband entropy method creates a clutter map and a scalar measure of clutter based on the assumption that clutter is inversely proportional to the amount of redundancy in an image. This measure was later found to be highly related to clutter around a target or local clutter (Asher, Tolhurst, Troscianko, & Gilchrist, 2013). Together, the three techniques provide an overall estimate of display clutter.

On the other hand, performance evaluation relies on comparing people's performance using displays that have (supposedly) different levels of clutter (e.g., Beck et al., 2010; Wickens, Nunes, Alexander & Steelman, 2005; Yamani & McCarley, 2011; Yeh, Brandenburg, Wickens, & Merlo, 2000). Performance measures that have been calculated include search time and error rate (Beck et al., 2012; Grahame et al., 2004), as well as a number of domain-specific measures. For example, in aviation research, Kaber et al. (2013) measured clutter in a flight simulator based on pilots' tracking error. In the medical domain, Zeng, Cimino, and Zou (2002) asked physicians to retrieve specific patient information and calculated the search time and accuracy of the responses to determine the level of clutter in medical displays. Finally, subjective assessment requires users to rate or rank the amount of clutter perceived in displays, thus relying on people's judgment to determine whether a display is cluttered (e.g., Alexander, Stelzer, Kim, & Kaber, 2008; Kaufmann & Kaber, 2010; Kim & Sundar, 2010). Thus clutter is a multi-faceted construct that can only be understood by considering several dimensions, as outlined in the following sections.

Display-based clutter dimensions. Density is the most prominent display-based clutter dimension, where clutter is viewed as the presence of a large number of objects within a display (e.g., Clay, 1993; Horrey & Wickens, 2004; Kroft & Wickens, 2002; Mack & Oliva, 2004; Tufte, 1983; Tullis, 1983; Ververs & Wickens, 1998). More precisely, it is not the number but rather the density of display entities – "too much data on too small an area" (Ellis & Dix, 2007; p.1216) – that results in clutter (Coco & Keller, 2009; van den Berg et al., 2009). The display density perspective of clutter thus extends the traditional notion of set size, where simple shapes and blank backgrounds are used, to more real-life objects (Rosenholtz et al., 2007). The nature of the “objects” on these displays varies widely, depending on the application domain. For example, the objects may be words or graphics on a webpage (Grahame et al., 2004), symbols or marks on a map (e.g., Lohrenz, Trafton, Beck, & Gendron, 2009; Yeh & Wickens, 2001), sensor readings in an infrared display (Wang & Zhang, 2011), icons on an airplane cockpit display (S.-H. Kim & Kaber, 2009), words in an electronic health record (Hammond et al., 2012), or the point sources in a radar display (McKenzie, Wong, & Gibbins, 2013). Since it may not always be possible to clearly enumerate individual entities within an image or display, definitions of this kind typically have applicability only within a particular domain. The view of clutter as display density is also somewhat oversimplified and incomplete; increasing the number or density of objects may increase clutter, but other factors – such as proper organization of these items – may mitigate the performance effects of density (Doyon-Poulin et al., 2012).

The second display-based dimension of clutter – layout – emphasizes this point. Here, the poor organization of display entities, not merely their number or density, is highlighted (for a review of display organization, see Nielsen (1993), Shneiderman (1998), or Wickens, Hollands, Banbury, and Parasuraman (2013)). This perspective has been reflected in numerous clutter definitions (e.g., Bravo & Farid, 2008; Doyon-Poulin et al., 2012; Rosenholtz, Li, Mansfield, & Jin, 2005; Tufte, 1991; van den Berg et al., 2009). For example, Peng, Ward, and Rundensteiner (2004; p.89) refer to clutter as “a disordered collection of graphical entities” or “the opposite of structure”. Factors that fall under the broad category of display organization include the lack of logical or conceptual grouping of items (Wickens & Carswell, 1995), the absence of symmetry (Oliva, Mack, Shrestha, & Peeper, 2004), the degree to which the target is obscured or masked

(Alexander, Stelzer, Kim, & Kaber, 2008; Bravo & Farid, 2004; Chu, Yang, & Li, 2012; Toet, 2010; Xing, 2007), the presence of high entropy (or lack of predictability) within a display (Rosenholtz, Li, & Nakano, 2007), and the variability in color, luminance, contrast, and orientation in a display (Lohrenz et al., 2009; Rosenholtz et al., 2005). All of these factors contribute to poor guidance to the target, which, in turn, leads to poor search performance. It is important to note that too little variability (e.g., a monochrome display) can also be detrimental, as unique colors help quickly identify objects (e.g., Kahneman & Treisman, 1984; Shontz et al., 1971; Yantis,1993), and color coding can be very useful for guiding visual search (e.g., Fisher & Tan, 1989).

Another layout-based aspect of clutter, which was mentioned earlier, is crowding, i.e., close spacing between a target and surrounding distractors. Crowding is related to local, as opposed to global, clutter (e.g., Beck, Lohrenz, & Trafton, 2010; Ewing et al., 2006). While global clutter refers to the amount or organization of information across an entire display, local clutter focuses on the amount and organization of information surrounding an important small area which includes the target.

Target-background or target-distractor similarity is the third major display-based dimension of clutter. Proponents of this perspective view clutter as the degree of similarity between characteristics of the background, such as color, and those of the target (Christ, 1975; Duncan & Humphreys, 1989; Liao, Wu, & Sheu, 2013; Wolfe, Oliva, Horowitz, Butcher, & Bompas, 2002). This perspective of clutter features prominently in the radar and Automatic Target Recognition (ATR) domains, where the main concern is to the detection of a target against background noise. In these domains, clutter has been defined as “objects or features in the scene that are similar to the desired target” (Chu et al., 2012; p. 067003-1), the “quantity of background signatures similar to the target" (H. Camp, Moyer, & Moore, 2010; p. 76620A-1), or the "signal caused by the background objects resembling a target. The more target-like objects or attributes in the background, the higher the clutter level" (He, Zhang, Liu, & Chang, 2008; p.5534).

Towards a performance-based, ergonomics-oriented definition of clutter. The above views of clutter are all based on features of the display, independent of the user and task. For example, from a display density perspective, an additional item, whether task relevant or not, will still contribute to clutter in the same way. However, the task relevance of display objects can have considerable bearing on whether they negatively affect performance (Alexander et al., 2008; Barbu, Lohrenz, & Layne, 2006; Doyon-Poulin et al., 2012; Horrey & Wickens, 2004; Lohrenz, Layne, Edwards, Gendron, & Bradley, 2006; Rosenholtz et al., 2007). User-based factors such as workload (Naylor, 2010) may also influence the perception and effects of clutter. Consider the example of a Primary Flight Display (PFD) on a modern airplane. At any given time, a number of task irrelevant items may be present on this display but, during routine operations, this does not necessarily result in adverse performance effects for an experienced pilot who imposes structure on the display in a top-down fashion. However, during high stress or workload, or with an inexperienced pilot, these irrelevant elements may distract or slow down search for critical information.

From a human factors and ergonomics perspective, it is arguably these performance and attentional effects of clutter that are of primary importance. Understanding and predicting these effects requires consideration of the interaction between display-based and user-based factors. Researchers that share this view have developed more ecologically-oriented definitions of clutter. For example, emphasizing the importance of the current task, clutter can be defined as “unwanted or unnecessary information” (Lohrenz et al., 2009; p.90), “redundant information” (Ahlstrom, 2005; p. 90), or “an abundance of irrelevant information” (Doyon-Poulin et al., 2012; p.2D1-2). A more comprehensive definition is that of Rosenholtz et al. (2005), who describe clutter as “the state in which excess items, or their representation or organization, lead to a degradation of performance at some task” (p. 761). Another similar definition is that of Alexander et al. (2008), who define clutter as the “unintended effect of displaying visual imagery that may obscure or confuse other information or that may not be relevant to the task at hand” (p. 1180). Combining these different definitions, clutter in this dissertation will be defined as “**the presence of performance and attentional costs that result from the interaction between high data density, poor display organization, and an abundance of irrelevant information**” (Moacdieh & Sarter, 2014; p. 65).

One important question is how user-based factors interact with clutter to affect attention and performance. In particular, it is critical to examine the role of factors that 1) are encountered in the data-rich, complex environments of interest in this line of research and 2) are known to degrade performance by themselves and may exacerbate the effects of clutter. While some of these factors, such as workload (e.g., Naylor, 2010), have been studied in the context of clutter, others, such as stress, have not received as much attention and will be explored more carefully in this research.

The Negative Effects of Stress on Attention and Performance

Stress has been identified as a hazard in numerous complex domains, such as aviation (Samel, Vejvoda, & Maas, 2004), medicine (Arora et al., 2010; Sexton, Thomas, & Helmreich, 2000), the military (Guznov, Matthews, Funke, & Dukes, 2011), and driving (Westerman & Haigney, 2000). Concerns about stress stem from its reported effects on attention and performance. These effects include attentional narrowing, a phenomenon where a person's visual field shrinks due to stress, making it difficult to detect cues in the peripheral visual field (Chajut & Algom, 2003; Cohen, 1980; Easterbrook, 1959; Janelle, Singer, & Williams, 1999; Murata, 2004; Salas, Driskell, & Hughes, 1996). More generally, stress makes it difficult to detect task-relevant information, increasing the amount of attention allocated to distractors or task-irrelevant data (Braunstein-Bercovitz, 2003; Braunstein-Bercovitz, Dimentman-Ashkenazi, & Lubow, 2001). Other reported effects of stress include decreased vigilance (Helton & Russell, 2011), depleted working memory (Ashcraft & Kirk, 2001), and reduced sampling of the environment (Hancock & Weaver, 2003).

As with clutter, there is no agreed-upon definition of stress which is reflected in how different experiments have manipulated and measured stress. Examples range from time pressure (Stokes, Kemper, & Marsh, 1992; Van Galen & van Huygevoort, 2000) and high mental workload (Recarte & Nunes, 2000, 2003) to heat (Vasmatzidis, Schlegel, & Hancock, 2002) and sleep deprivation (Baranski, Gill, McLellan, Moroz, Buguet, & Radomski, 2002). The difficulty with defining stress stems from its close relation to several other concepts such as anxiety and arousal (Tepas & Price, 2001). Anxiety differs from stress by being a more affective condition, or an "aversive emotional and motivational state" (Eysenck et al., 2007; p. 336). On the other hand, arousal, which is defined as the activation level of the central nervous system (Razmjou, 1996), has been strongly linked and even equated with stress, particularly in early research studies (Broadhurst, 1957; Duffy, 1941, 1957; Easterbrook, 1959; Razmjou, 1996; Yerkes and Dodson, 1908).

A different perspective has been adopted by researchers who view stress as a complex, multi-faceted construct that is based on the interaction between the environment and the person (e.g., Gaillard, 2001; Hancock & Warm, 1989; Matthews et al., 2002). For example, Hancock and Warm (1989) believe stress to be a function of three factors: environmental stressors, cognitive adaptation mechanisms, and physical or performance effects. Similarly, Wofford and Daly (1997) agree that there are three components to the stress response, namely physiological, psychological, and behavioral responses. In addition, McGrath (1976) views stress as the interaction between physical and/or psychological demands, a person's perception of how well they can cope, and their perception of how important these demands are. This definition, which will be adopted in this dissertation, highlights the importance of motivation in bringing about the feeling of stress (see also Welford, 1973; Lovallo, 1997). In other words, a condition can be termed stressful when factors such as time pressure and workload, combined with motivation to perform to the best of one's abilities, bring about negative physical and cognitive effects.

Some research has shown that stress may affect how people perceive clutter. For example, in two related flight simulator studies, results showed that higher stress (severe weather conditions during a simulated flight) elicited higher clutter ratings (Alexander et al., 2012; Kim et al., 2011). However, more behavior- and performance-based, rather than subjective data is needed to better understand the effects of, and interaction between, clutter and stress. Such data is needed also to develop means of detecting and counteracting clutter. The next section will detail the proposed means of achieving these goals

Eye Tracking-based Adaptive Interfaces as a Means of Overcoming the Effects of Clutter

To account for varying degrees of clutter, stress, and task demands, a context-sensitive – as opposed to static – approach to display design is needed. Interfaces may be adjusted in response to changes in an operator's state, environmental factors, or at an operator's request (e.g., Dorneich et al, 2003; Hancock & Scallen, 1996; Kaber & Riley, 1999; Moray, Inagaki, & Itoh, 2000; Sarter , 2007; Schmorrow & Kruse, 2004; Parasuraman, Mouloua, & Molloy, 1996; Rouse, 1988). Two approaches to context-sensitive display (and overall system) adjustments have been discussed: 1) *adaptable* and 2) *adaptive* design (Barnes et al., 2006; Miller et al., 2005; Parasuraman, Bahri, Deaton, Morrison, & Barnes, 1992; Sarter, Billings & Woods, 1997).

The adaptable approach puts operators in charge of deciding when and how to change the display (e.g., Opperman, 1994; Scerbo, 2001). This approach has a number of advantages, namely more perceived control over the system, better situation awareness, and increased acceptance of the system (Miller & Parasuraman, 2007; Wickens, 1994). However, an adaptable approach also has a major disadvantage: it adds an extra interface management task for the operator, thus increasing workload in often already demanding circumstances (Kirlik, 1993; Wiener, 1989). Also, operators may not always be the best judges of their own needs and abilities (Andre & Wickens, 1995). For example, one study examined the benefits of an adaptable interface for emergency management (Shen, Carswell, Santhanam, & Bailey, 2011). Participants in the experiment could change the type of display provided between plan view, elevation view, and 3D view, with each display being best suited for a particular task. Most participants did not choose the correct display for their tasks, and performance improved when they were provided with help to select the best display.

Alternatively, an adaptive approach could be adopted where the system is responsible for deciding when and how to change the display (Inagaki, 2003; Kaber, Wright, Prinzel, & Clamann, 2005; Keeble & Macredie, 2000). For example, in the study by Cummings, Brzezinski, and Lee (2007) on the design of an unmanned aerial vehicle control display, an intelligent algorithm identified upcoming high-workload situations and highlighted these conditions for operators to alert them when several critical tasks were predicted to occur at the same. Adaptive systems have been shown to be less acceptable to users who perceive a loss of control (Shneiderman, 1997; Wickens, 1994). However, adaptive systems present a major advantage in that they do not add to a user's workload during critical times (Hou, Kobierski, & Brown, 2007; Parasuraman et al., 1998). For example, Bailey et al. (2006) found that mental workload decreased with an adaptive, as opposed to an adaptable approach to a failure detection task. This advantage of adaptive systems seems particularly desirable in the high-stress and time-constrained situations that are the focus of this dissertation, where operators' workload levels are usually already high. Support for this notion is provided by a flight simulator study by Olson and Sarter (2000) which showed that pilots generally preferred to have control over flight deck automation (i.e., an adaptable design); however, in high-stress cases, involving time pressure and high workload, they preferred to have the automation make critical decisions while still retaining the right to make adjustments afterwards (an adaptive approach).

The adaptive approach to context-sensitive design is thus adopted in this dissertation. In order to implement an adaptive display system, three components or mechanisms are needed:

1. A sensing mechanism that measures or traces user and/or environmental conditions in real time

2. A control algorithm that decides when intervention or change is needed in the display and what that change should be
3. An adjustment mechanism that triggers the appropriate display adjustment(s).

Sensing Mechanism: Eye Tracking as a Promising Means of Tracing the Effects of Clutter

One critical component of any adaptive system is a means of sensing the state of the user and/or environment in order to determine when interventions are needed. In the case of creating an adaptive display for clutter reduction, this translates into the need for a mechanism to trace the effects of clutter on attention allocation in real time. Techniques that are traditionally used to measure clutter are not suited for this goal. Image processing is of limited use as it does not reflect how user- and environment-based factors like stress interact with design to shape performance and attention. Measures such as response time and accuracy assess performance, but they do not trace the underlying attentional mechanisms continually and in real time. Moreover, relying on response time and accuracy for display adaptations involves waiting until the completion of a task, by which time it may be too late to adjust the display. Similarly, subjective evaluation using questionnaires and interviews cannot be used in real time. Other mechanisms discussed in the literature on adaptive systems include the use of critical events (e.g., Barnes & Grossman, 1985) and models of cognition (e.g., Rouse, Geddes, & Curry, 1987; see Parasuraman et al. (1992) or Stanney (2004) for reviews). Such techniques are of limited use as they do not take into account the state of the operator at each point in time but rather rely on matching previously-seen patterns.

Using a physiological measure helps overcome this problem. This approach has given rise to the field of augmented cognition (Stanney et al., 2004). Numerous physiological measures – and sometimes combinations of these measures – have been used, including EEG (Freeman et al., 1999), fNIRS (Izzetoglu et al., 2005), galvanic skin response (Critchley, 2002), heart rate (Lee et al., 2005), blood pressure (Boucsein & Backs, 2000), and eye movements or characteristics (Fuchs, Jones, & Hale, 2007; Iqbal, Zheng, & Bailey, 2004; Pomplun & Sunkara, 2003).

Eye movements as a promising sensing mechanism. Recording eye movements, or the location of gaze, appears to be a very promising approach for developing adaptive displays to overcome clutter. Eye movements can be obtained using an eye tracker, a device that uses infrared light to trace where people are looking at on a display (see Duchowski (2007) or Poole and Ball (2006) for a detailed review). Eye trackers can be desktop-mounted, meaning that they are placed by the display to be tracked, or head-mounted, in which case they are worn by the person either as a headpiece or as a device similar to glasses. In both cases, eye trackers output raw eye location data known as points of regard (POR) that indicate where a person is looking in the display. PORs are typically recorded at 50 Hz or higher and are expressed in x and y coordinates (Munn, Stefano, & Pelz, 2008). In turn, the PORs can be used to determine the two basic components of eye tracking research, *fixations* and *saccades*. Fixations, which are characterized by location and duration, are formed from spatially stable PORs and it is during this time that visual processing takes place (Findlay, 2004; see Figure 1.1). The rapid eye movements between successive fixations are called saccades, during which time visual processing is usually suppressed (Yarbus, Haigh, & Riggs, 1967). A scanpath is the path of a sequence of fixations and saccades, and it provides a means to visualize eye movements (Noton

& Stark, 1971). Finally, an area of interest (AOI) is an experimenter-defined region of the display on which analysis of eye tracking data is performed. Together, these components have been used as the building blocks for eye tracking research.

There are several techniques that can be used to identify fixations and saccades from the raw gaze points (Salvucci & Goldberg, 2000). In this research, a set of consecutive gaze points constituted a fixation if there were at least six gaze points and they were within a two-degree visual angle radius (Goldberg & Kotbal, 1999). With an eye tracker that sampled at 60 Hz, this meant that the minimum duration for fixations was 100 ms. This method is classified as a dispersion algorithm by Salvucci and Goldberg (2000). Any other gaze points that were not classified as fixations were grouped together as saccades, meaning that it was not possible to detect any smaller microsaccades that may have been present. However, for the purposes of this research, these smaller saccades can be considered negligible and do not have any significant influence on results.

As for challenges related to using eye tracking, these include the high cost and setup/analysis time (Jacob & Karn, 2003) and the correct selection of parameters for fixation calculation (Inhoff & Radach, 1998). However, the benefits of using eye tracking to avoid delays and misses in data-rich, safety-critical domains outweigh these issues, and hardware developments in the future may alleviate many of these problems (Pavlas, Lum, & Salas, 2012).

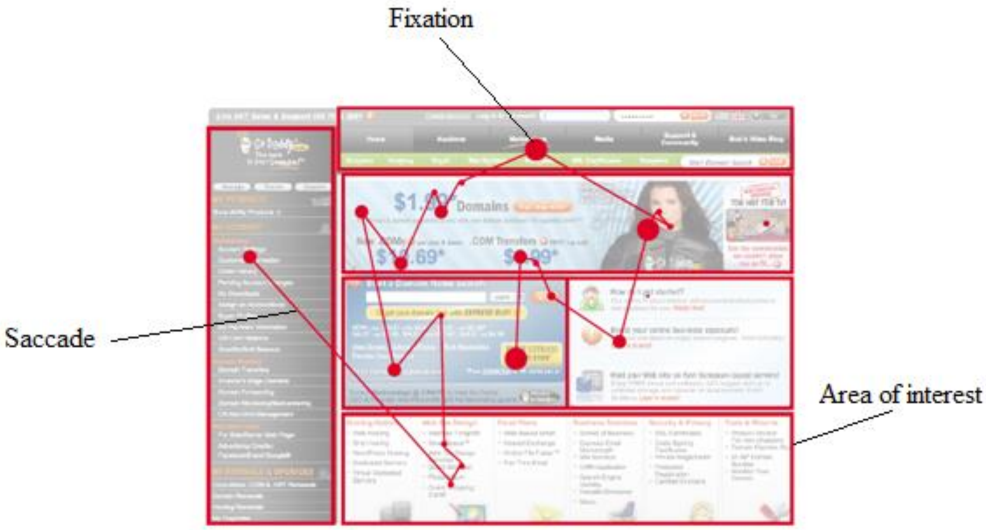


Figure 1.1: Fixations are usually depicted as a circle whose diameter is proportional to fixation duration. Saccades are represented as lines between two successive fixations, while areas of interest are typically drawn as rectangles. All of the fixations and saccades together create the scanpath (adapted from Bonigala (2009))

Benefits of using eye tracking. Eye tracking has been used extensively in human factors research, including studies in aviation (e.g., Alexander & Wickens, 2005; Ellis, Kramer, Shelton, Arthur, & Prinzel, 2011; Schnell, Kwon, Merchant, & Etherington, 2004), driving (Di Stasi, Contreras, Candido, & Cantena, 2011; Liang , Reyes, & Lee, 2007), website design (Katsanos, Tselios, & Avouris, 2010; DeWitt, 2010), and medicine (Chetwood et al., 2012; Marquard, Jo, Henneman, Fisher, & Henneman, 2012). There are several benefits to using eye tracking that have encouraged its use in research, as described below.

Eye tracking as a real-time, non-invasive tool. Eye tracking provides a number of advantages as a sensing mechanism in an adaptive display. First and most importantly, eye tracking data can be obtained in real time. Such data has been used to detect driver distraction (e.g., Liang et al., 2007) and sleepiness (Jin et al., 2013), evaluate user learning (Kardan & Conati, 2012), measure workload (Durkee, Geyer, Pappada, Ortiz, & Galster, 2013), select areas of a display (Kumar, Paepcke, & Winograd, 2007), and provide support to non-native English speakers by displaying the meanings of difficult words (Hyrskykari, 2006). Desktop-mounted eye trackers are also completely non-invasive, in contrast to many other sensing mechanisms, such as EEG.

Eye tracking and attention. The use of eye tracking is based on the assumption that the location of a person's gaze can be considered a stand-in for the locus of attention, a theory known as the eye-mind hypothesis (Just & Carpenter, 1978). Some researchers (e.g., Anderson, Bothell, & Douglass, 2004) claim that there could be a dissociation or lag between attention and the location of gaze. Others propose that attention can sample different areas of the display at a higher rate than eye movements (e.g., Horowitz & Wolfe, 2003). However, for most human factors applications, eye tracking can still provide a relatively good estimate of the location of attention (Goldberg & Wichansky, 2003; Rayner, 1998; Zelinsky, 2008).

The use of eye tracking has proven invaluable in the study of attention and visual search (e.g., Findlay & Gilchrist, 2005; Trukenbrod & Engbert, 2007; Williams & Pollatsek, 2007). At the most basic level, the correlation between the number of saccades and search time has been used to prove that eye movements are necessary for visual search (e.g., Zelinsky & Sheinberg, 1997). In addition, eye tracking has been used to test and validate models of eye movements during visual search (e.g., Zelinksy, 2008) and to create salience maps (maps that indicate the regions of the screen most likely to attract attention; e.g., Chukoskie, Snider, Mozer, Krauzlis, & Sejnowski, 2013; Henderson, Brockmole, Castelhano, & Mack, 2007; Judd, Ehinger, Durand, & Torralba, 2009; Parkhurst, Law, & Niebur, 2002). The salience of abrupt visual onsets has also been verified using eye tracking, with saccades to suddenly appearing distractors observed during visual search for a target (e.g., Theeuwes et al., 1998). Other eye tracking experiments have examined the concept of inhibition of return (Klein, 2000) or the tendency to avoid already visited areas of a display (e.g., Beck, Peterson, Boot, Vomela, & Kramer, 2006; Geyer, von Muhlenen, & Muller, 2007). The above studies show that, in contrast to performance outcome measures such as response time, an eye tracker is a process-oriented tool that can trace the shifts and degradations of attention at a fine-grained level of analysis.

In addition, eye tracking data can reflect user- and task-based influences on attention control, such as experience (Beck et al., 2012; Konstantopoulos, Chapman, & Crundall, 2010). And, importantly for the purposes of this dissertation, there is evidence that eye tracking can capture the effects of stress on attention (e.g., Di Nocera, Camilli, & Terenzi, 2007).

Eye tracking and clutter. To a limited extent, researchers have used eye tracking to explore the effects of clutter on attention and performance (e.g., Beck et al., 2010; Beck et al., 2012; Grahame, Laberge, & Scialfa, 2004; Neider & Zelinsky, 2011; Zhu & Sun, 2012). For example, research on large set size has led to the conclusion that the higher or more dense the number of objects on a display, the slower the search as the number of fixations and fixation duration increase, while the mean saccade amplitude decreases (e.g., Bertera & Rayner, 2000; Greene & Rayner, 2001; Vlaskamp & Hooge, 2006).

These types of experiments all used very simple, artificial stimuli, such as letters and shapes. Researchers in human factors and other applied disciplines have attempted to build on such studies to determine whether they could be generalized to more complex displays. For example, some studies showed a significant increase in the number of fixations in more cluttered aeronautical charts and websites (e.g., Beck et al., 2010, 2012; Grahame et al., 2004).

Calculating the number of fixations on added objects or distractors (Hegarty, De Leeuw, & Bonura, 2008; Beck et al., 2010) as well as the amount of time spent on task-relevant items (Fabrikant, Hespana, & Hegarty, 2010) has also been used to determine the effects of clutter, Longer fixation times and thus lower fixation rates have been linked to higher levels of clutter (Beck et al., 2010; Henderson et al., 2009; Zhu & Sun, 2012), as have the latency of the first saccade (Henderson et al., 2009; Zhu & Sun, 2012; Zelinsky, 2001), scanpath length (sum of saccade amplitudes; Goldberg, Stimson, Lewenstein, Scott, & Wichansky, 2002); scanpath ratio (scanpath length divided by the length of the shortest path from the starting point to the target (Neider & Zelinsky, 2011)), and final saccade amplitude (length of the last saccade before target detection can give an indication of how easily noticeable the target is in peripheral vision (e.g., Henderson, Weeks, & Hollingsworth, 1999)).

Eye tracking has also been used to study users' search strategies, although there is no firm agreement on the matter. In one case, a "coarse-to-fine" search strategy was observed where participants tried to quickly extract as much as they could from the display before resorting to slower, more deliberate search (Beck et al., 2012). Henderson et al. (2009), on the other hand, found that users tend to search in areas of high clutter first, which may be because these were regions of high salience as well.

Research gaps related to clutter and eye tracking. Despite the promise of eye tracking and the studies conducted to date, there are numerous gaps in the literature on clutter and eye tracking that need to be resolved before an eye-tracking based adaptive display can be developed. First, none of the eye movement metrics mentioned above have been collected and analyzed in real time. Rather, they all reflect eye movements over an entire task period. In rare cases, one fixation at a time has been analyzed, but a moving window of time to calculate the metrics has never been used in the context of clutter. In addition, these metrics may not fully capture all of the effects of clutter on attention which are best described in terms of whether they are related to the *location/spread*, *directness,* and *duration* of eye movements. Location metrics depend only on fixation coordinates; they show whether clutter causes a dispersion of eye movements across the display, thus preventing the user from focusing on important information. Increased spread suggests increased coverage of sampling of different areas of the display, which could occur with a large amount of irrelevant data and poor guidance to the target. Directness measures differ in that the sequence of fixations is taken into account; these measures can indicate whether clutter made search less ordered or systematic. Directness measures help show how efficiently users reached the target destination, which, in turn, can provide insight into whether there was strong guidance to the target or whether elements of the display were distracting. Finally, duration measures indicate how long a person looked at a particular area and relate clutter primarily to the difficulty extracting information from the display or the perceived importance of the information.

The most frequently used eye movement metrics that have been used to date do not fully capture the three aspects of spread, directness, and duration. Spread metrics, such as the number of fixations in various AOIs, and duration metrics, such as mean fixation duration, are most commonly used (e.g., Beck et al., 2010, 2012). However, there are other metrics in each category that may provide valuable insight on the different aspects of clutter. For example, convex hull area and spatial density are two other location/spread measures which can indicate how dispersed attention is. Convex hull area refers to the minimum convex area which contains the set of fixation points (Goldberg & Kotval, 1999); it relates to a person's useful field of view (Wickens & McCarley, 2008). Spatial density, on the other hand, consists of placing a grid of cells on the display and then dividing the number of cells containing at least one gaze point by the total number of cells (Goldberg & Kotval, 1999). Few studies employ measures of directness, such as the ratio of transitions (placing a grid of cells on the display and then dividing the number of cells containing a transition either to or from another cell by the total number of cells) and the number of backtracks (a saccade or gaze angle between two successive saccades that is greater than 90 degrees; Goldberg & Kotval, 1999).

Another limitation of eye tracking research is that several of the metrics employed to date rely on areas of interest, or AOIs, to be defined in the display by the experimenter. This means that results are display-dependent and also highly reliant on the accuracy of the eye tracker. Conversely, in this line of research, the proposed metrics were all display-independent, making them more generalizable and applicable across displays. In addition, the focus was on metrics that would lend themselves to real time processing, meaning that averages or rates were considered where necessary. The metrics proposed for this research are described in Table 1.1. One of the goals of this research was then to determine 1) which of the above metrics best trace the effects of clutter on the spread, directness, and duration of attention and 2) which metrics are best suited for use in real time to create an adaptive display.

Table 1.1: Summary of the eye movement metrics proposed for this research

Eye tracking metric	**Notes**
Location metrics	
Convex hull area (pixels^2)	Minimum convex area which contains the fixation points (Hegarty et al., 2008)
Spatial density	The number of grid cells containing gaze points divided by the total number of cells (Cowen, Ball, & Delin, 2002)
Nearest neighbor index (NNI)	The ratio between (1) the average of the observed minimum distances between points and (2) the mean random distance expected if the distribution were random (see Di Nocera et al. (2007) for more details); higher values indicate more dispersion of fixations
Directness metrics	
Scanpath length per second (pixels/sec)	The total scanpath length per second
Mean saccade amplitude (pixels)	Mean of all the saccades within a defined time period
Backtrack rate (/sec)	A backtrack is defined as an angle between two saccades that is greater than 90 degrees (Goldberg & Kotval, 1999)
Rate of transitions (/sec)	Rate of transitions between = grid cells (as defined for spatial density; Goldberg & Kotval, 1999)
Duration metrics	
Mean fixation duration (sec)	Mean duration of all fixations within a defined time period

Adaptation Algorithm: Logistic Regression

The second important component of any adaptive system is a means of determining when adaptations or changes are needed. This subsystem analyzes the data collected by the sensing mechanism (in this case, eye tracking data) and determines whether the system and/or user is currently in a desirable (i.e., no change needed) or non-desirable (i.e., changes needed) state. At the simplest level, researchers can compare the different measures collected in real time and evaluate them or compare them to some threshold (e.g., Prinzel et al., 2000). For example, Pope et al. (1995) determined whether manual or automatic mode was needed based on whether individual EEG values that indicated task engagement were increasing or decreasing. Alternatively, a modeling or machine learning approach can be used in order to determine the state of the system based on combinations of the input features. Techniques that have been used include artificial neural networks (e.g., Russell, 2005), genetic programming (Bergstrom et al., 2000), Naïve Bayes (Mokhtar, Abdullah, & Zin, 2011), support vector machines (SVM; e.g., Liang et al., 2004), and logistic regression (Barr et al., 2008; Ratwani, McCurry, & Trafton, 2008; Steichen, Garenini, & Conati, 2013).

Which one of these techniques is preferable to use is still open to debate. Evidence from studies that have compared multiple techniques is inconclusive. For example, Liang et al., (2004) compared SVM and logistic regression for detecting driver distraction based on eye movements, and found that SVM outperformed logistic regression. On the other hand, Steichen et al. (2013) tested SVM, decision trees, multilayer perception, and logistic regression and found that logistic regression had the highest accuracy among all. Several researchers have pointed out that logistic regression is a good option for real-time analysis given that it is light-weight and efficient for online processing, as well as the fact that it provides insight into the importance of different features (Kozma, Klami, & Kaski, 2009). In this case, it would help identify the best eye movement metrics to use. Other advantages of logistic regression include its robustness, which would be particularly useful in this research where unequal sample sizes and a small data set size are expected (Blom, Paradis, & Duncan, 2012). The combination of these various advantages led to the decision to use this technique in the present research.

Display Adjustments: Highlighting and Shortcut Panel

The third and final component of an adaptive system consists of the actual changes or adaptations that are triggered. Several display design techniques have been used to date to ’declutter’ an interface. In general, these techniques aim to increase the salience of important information. They include highlighting the important information (O’Hara et al., 2007), cuing the important information (Yeh, Merlo, Wickens, & Brandenburg, 2003), dimming, lowlighting, or increasing the transparency of irrelevant data (Kroft & Wickens, 2001; St. John Smallman, Manes, Feher, & Morrison, 2004), or removing unnecessary data (Butichibabu, Grayhem, Hansman, & Chandra, 2012; Yeh & Wickens, 2001). Only few of the aforementioned approaches have been implemented in real time. Exceptions include Barker et al. (2004), who obscured the unnecessary detail on a navigation map in real time when high user cognitive load was detected (based on a number of physiological measures). In addition, St. John et al. (2004) used an algorithm to determine the threat level of aircraft on a tactical display. The ones considered to be less threatening were dimmed, which led to faster detection of the threatening aircraft. None of these approaches were based on eye movements.

It is notable that decluttering techniques typically focus on the excess data aspect of clutter; addressing poor display organization has received considerably less attention. In one of few studies exploring this approach, it was shown that it may be beneficial to change the location of information to bring it close to where it is needed at any given time (Sukaviriya & Foley, 1990). This may be particularly useful in the medical domain, where the few attempts to address clutter have often involved redesigning displays so that relevant information is better grouped (Samal et al., 2011; Zeng et al., 2002).

Implementing the appropriate display adaptation technique is non-trivial, as excessive or poorly-designed adaptations may make the system more difficult and frustrating to use (Rothrock, Koubek, Fuchs, Haas, & Salvendy, 2002). One important consideration is that the addition of any new information must not result in large switching costs, or constant switching of attention (Gopher, Armony, & Greenspan, 2000) which can lead to stress and worse performance, especially during difficult tasks (McFarlane, 2002). Another consideration is the need to minimize task interruptions (Iqbal, Adamczyk, Zheng, & Bailey, 2005). As with any form of automation, it is also necessary to prevent overtrust in what may not be a completely reliable system (St. John et al., 2004). One promising means of preventing these problems is to develop adaptations that support at-a glance monitoring, in which case few attentional resources are required to process the display (Kaber, Riley, Tan, & Endsley, 2001). In such cases, operators would rely on what is known as preattentive processing, where mental processing can occur automatically and early on without consuming cognitive resources (Treisman, 1986). Ambient cues in a user's peripheral field, for example, are one approach that would support preattentive processing and minimally interfere with one's current task (Arroyo, Stelker, & Stouffs, 2002).

For this research, the display adjustments were the last phase of the project and represented an initial test of whether physicians would find such display adjustments useful or distracting. Two types of adjustments were selected, one from the display density group (highlighting) and one from the display organization group (a shortcut panel the provides a suggested location). Highlighting was selected given that it could be easily implemented in the context of an EMR task and would be similar to some of the cues that ED physicians already have. As for the shortcut panel, it was inspired in part by the interviews with physicians, where they pointed out the need to immediately see critical values.

Summary

Display clutter can lead to significant performance decrements in visual search and noticing tasks, particularly during stressful situations. The overall goal of this dissertation was to use eye tracking to detect, explain, and counteract the effects of clutter on attention and performance. The specific aims were to:

1. Identify and empirically validate a set of eye movement metrics that can capture and explain the effects of clutter (both high density and poor organization) and stress on information acquisition, based on underlying attention processes.
2. Determine which of these metrics are best suited to tracing the effects of clutter in real time and use these metrics to build models to predict the effects of clutter prior to significant performance decrements.
3. Implement, test, and evaluate the acceptability of real-time display adjustments to prevent breakdowns in attention allocation and information acquisition.

The effects of clutter in both simple, highly-controlled settings (using static images or a graphics program) and in a more complex, data-rich environment (simulated emergency department (ED) electronic medical records (EMRs)) were explored as part of this research. This combination of research settings allowed for the systematic manipulation of clutter and its constituents in some cases, as well as for examining the acceptability of findings to more real-world contexts. Ultimately, this research is expected to contribute to increased efficiency and safety of operations in a variety of domains, such as website design, medicine, and aviation.

The following chapter, Chapter 2, will provide an introduction to the ED environment and analyze the problem of clutter and stress in this particular domain. Chapters 3, 4, 5, and 6 will then describe a series of experiments that were conducted in order to accomplish the goals of this dissertation. Finally, Chapter 7 will summarize the lessons learned and suggest future work that is needed in this field of research.

References

Ahlstrom, U. (2005). Work domain analysis for air traffic controller weather displays. *Journal of Safety Research*, *36*(2), 159-169.

Amy L. Alexander , David B. Kaber , Sang-Hwan Kim , Emily M. Stelzer , Karl Kaufmann & Lawrence J. Prinzel III (2012) Measurement and Modeling of Display Clutter in Advanced Flight Deck Technologies, The International Journal of Aviation Psychology, 22:4, 299-318

Alexander, A. L., Stelzer, E. M., Kim, S. H., & Kaber, D. B. (2008). Bottom-up and top-down contributors to pilot perceptions of display clutter in advanced flight deck technologies. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1180-1184). New York City, NY: Sage Publications.

Alexander, A. L., & Wickens, C. D. (2005). Synthetic vision systems: Flightpath tracking, situation awareness, and visual scanning in an integrated hazard display. In *Proceedings of the 13th International Symposium on Aviation Psychology*.

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye Movements Do Not Reflect Retrieval Processes Limits of the Eye-Mind Hypothesis. *Psychological Science*, *15*(4), 225-231.

Andre, A.D. (2001). The value of workload in the design and analysis of consumer products. In P.A. Hancock, & P.A. Desmond, Stress, workload, and fatigue (pp. 373-382). Mahwah, NJ: L. Erlbaum.

Andre, A. D., & Wickens, C. D. (1995). When users want what's not best for them. *Ergonomics in Design: The Quarterly of Human Factors Applications*,*3*(4), 10-14.

Arora, S., Ashrafian, H., Davis, R., Athanasiou, T., Darzi, A., & Sevdalis, N. (2010). Emotional intelligence in medicine: a systematic review through the context of the ACGME competencies. *Medical education*, *44*(8), 749-764.

Arroyo, E., Selker, T., & Stouffs, A. (2002). Interruptions as multimodal outputs:Which are the less disruptive? In *Proceedings of the IEEE 4th International Conference on Multimodal Interfaces* (*ICMI '02,* pp. 479–483). Los Alamitos, CA: IEEE Computer Society.

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of experimental psychology: General, 130*(2), 224.

Asher, M. F., Tolhurst, D. J., Troscianko, T., & Gilchrist, I. D. (2013). Regional effects of clutter on human target detection performance. *Journal of vision*, *13*(5), 25.

Bailey, N. R., Scerbo, M. W., Freeman, F. G., Mikulka, P. J., & Scott, L. A. (2006). Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(4), 693-709.

Baldassi, S., Megna, N., & Burr, D. C. (2006). Visual clutter causes high-magnitude errors. *PLoS biology, 4*(3), e56.

Baranski, J. V., Gil, V., McLellan, T. M., Moroz, D., Buguet, A., & Radomski, M. (2002). Effects of modafinil on cognitive performance during 40 hr of sleep deprivation in a warm environment. *Military Psychology, 14*(1), 23.

Barbu, C., Lohrenz, M., & Layne, G. (2006). Intelligent electronic navigational aids: a new approach. In *Proceedings of the 5th International Conference on Machine Learning and Applications* (ICMLA '06) (pp. 109-116). Orlando, FL: IEEE.

Barker, R. A., Edwards, R. E., O'Neill, K. R., & Tollar, J. R. (2004). *DARPA improving warfighter information intake under stress—Augmented cognition concept validation experiment (CVE) analysis report for the Boeing team* (Contract NBCH030031). Arlington, VA: Defense Advanced Research Projects Agency.

Barnes, M. J.; Grossman, J. (1985). The Intelligent Assistant for Electronic Warfare Systems. NWC TP 5885. U.S. Naval Weapons Center: China Lake, CA.

Barnes, M., R. Parasuraman, and K. Cosenzo. 2006. Adaptive automation for military robotic systems. In NATO Technical Report RTOTR-HFM-078 uninhabited military vehicles: Human factors issues in augmenting the force. 420–440.

Barr, D. J. (2008). Analyzing 'visual world'eyetracking data using multilevel logistic regression. *Journal of memory and language*, *59*(4), 457-474.

Beck, M. R., Lohrenz, M. C., & Trafton, J. G. (2010). Measuring search efficiency in complex visual search tasks: global and local clutter. *Journal of Experimental Psychology: Applied; Journal of Experimental Psychology: Applied*, *16*(3), 238-250.

Beck, M. R., Peterson, M. S., Boot, W. R., Vomela, M., & Kramer, A. F. (2006). Explicit memory for rejected distractors during visual search. *Visual Cognition*, *14*(2), 150-174.

Beck, M. R., Trenchard, M., van Lamsweerde, A., Goldstein, R. R., & Lohrenz, M. (2012). Searching in clutter: Visual attention strategies of expert pilots. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1411-1415). Boston, MA: Sage Publications.

Bednarik, R., Vrzakova, H., & Hradis, M. (2012, March). What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In*Proceedings of the symposium on eye tracking research and applications* (pp. 83-90). ACM.

Bergström, A., Jaksetic, P., & Nordin, P. (2000, January). Enhancing information retrieval by automatic acquisition of textual relations using genetic programming. In *Proceedings of the 5th international conference on Intelligent user interfaces* (pp. 29-32). ACM.

Blom, E., Paradis, J., & Duncan, T. S. (2012). Effects of input properties, vocabulary size, and L1 on the development of third person singular–s in child L2 English. *Language Learning*, *62*(3), 965-994.

Bonigala, M. (2009, October 2). Top ten website design mistakes. Retrieved from http://www.spellbrand.com/top-ten-website-design-mistakes

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). ACM.

Boucsein, W., & Backs, R. W. (2000). Engineering psychophysiology as a discipline: Historical and theoretical aspects. *Engineering psychophysiology. Issues and applications*, 3-30.

Booth, R., & Sharma, D. (2009). Stress reduces attention to irrelevant information: Evidence from the Stroop task. *Motivation and Emotion*, *33*(4), 412-418.

Braunstein-bercovitz, H. (2003). Does stress enhance or impair selective attention? The effects of stress and perceptual load on negative priming.*Anxiety, stress, and coping*, *16*(4), 345-357.

Braunstein-Bercovitz, H., Dimentman-Ashkenazi, I., & Lubow, R. E. (2001). Stress affects the selection of relevant from irrelevant stimuli. *Emotion*, *1*(2), 182.

Bravo, M. J., & Farid, H. (2004). Recognizing and segmenting objects in clutter. *Vision research*, *44*(4), 385-396.

Bravo, M. J., & Farid, H. (2006). Object recognition in dense clutter. *Attention, Perception, & Psychophysics, 68*(6), 911-918.

Bravo, M. J., & Farid, H. (2008). A scale invariant measure of clutter. *Journal of Vision, 8*(1):23, 1-9

Broadhurst, P. L. (1957). Emotionality and the Yerkes-Dodson law. *Journal of Experimental Psychology, 54*(5), 345.

Brookhuis, K. A., & de Waard, D. (2001). Assessment of drivers'workload: performance and subjective and physiological indexes. *Stress, workload and fatigue*.

Bureau d'Enquetes et D'Analyses. (2012). Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro–Paris. *Ministère de l'Écologie. du Dévéloppement durable, des Transports et du Logement, Paris*.

Burnham, B. R. (2007). Displaywide visual features associated with a search display's appearance can mediate attentional capture. Psychonomic Bulletin Review, 14(3), 392−422.

Butchibabu, A., Grayhem, R., Hansman, R. J., & Chandra, D. (2012, October). Evaluating a de-cluttering technique for NextGen RNAV and RNP charts. In*Digital Avionics Systems Conference (DASC), 2012 IEEE/AIAA 31st* (pp. 2D2-1). IEEE.

Byun, H., & Lee, S. W. (2002). Applications of support vector machines for pattern recognition: A survey. In *Pattern recognition with support vector machines* (pp. 213-236). Springer Berlin Heidelberg.

Camp, H. A., Moyer, S., & Moore, R. K. (2010). Comparing masked target transform volume (MTTV) clutter metric to human observer evaluation of visual clutter. In *Proceedings of SPIE Defense, Security, and Sensing* (pp, 76620A-1-12). Orlando, FL: International Society for Optics and Electronics (SPIE).

Campbell, C., & Ying, Y. (2011). Learning with support vector machines.*Synthesis Lectures on Artificial Intelligence and Machine Learning*, *5*(1), 1-95.

Chajut, E., & Algom, D. (2003). Selective attention improves under stress: implications for theories of social cognition. *Journal of personality and social psychology*, *85*(2), 231.

Christ, R. E. (1975). Review and analysis of color coding research for visual displays. *Human Factors*, *17*(6), 542-570.

Chu, X., Yang, C., & Li, Q. (2012). Contrast-sensitivity-function-based clutter metric. *Optical Engineering, 51*(6), 067003-1-6.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in cognitive sciences*, *4*(5), 170-178.

Chang, H., Zhang, J., Liu, X., Yang, C., & Li, Q. (2010). Color image clutter metrics for predicting human target acquisition performance. In *Proceedings of the 6th International Conference Wireless Communications Networking and Mobile Computing (WiCOM)* (pp. 1-4). Chengdu, China: IEEE.

Chetwood, A. S., Kwok, K. W., Sun, L. W., Mylonas, G. P., Clark, J., Darzi, A., & Yang, G. Z. (2012). Collaborative eye tracking: a potential training tool in laparoscopic surgery. Surgical Endoscopy, 26(7), 2003–2009.

Chukoskie, L., Snider, J., Mozer, M. C., Krauzlis, R. J., & Sejnowski, T. J. (2013). Learning where to look for a hidden target. *Proceedings of the National Academy of Sciences*, *110*(Supplement 2), 10438-10445.

Clay, M. C. (1993). Key cognitive issues in the design of electronic displays of instrument approach procedure charts. Department of Transportation Report DOT-VNTSC-FAA-93-18.

Creedy, S. (2011, June). Information overload and cockpit confusion triggers training call. *The Australian.* Retrieved from http://www.theaustralian.com.au/business/aviation/information-overload-and-cockpit-confusion-triggers-training-call/story-e6frg95x-1226080851027

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Critchley, H. D. (2002). Book review: electrodermal responses: what happens in the brain. *The Neuroscientist*, *8*(2), 132-142.

Coco, M. I, & Keller, F. (2009). The impact of visual information on reference assignment in sentence production. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci)*. Amsterdam, Netherlands: Cognitive Science Society.

Cohen, S. (1980). Aftereffects of stress on human performance and social behavior: a review of research and theory. *Psychological bulletin*, *88*(1), 82.

Cowen, L., Ball, L. J., & Delin, J. (2002). An eye movement analysis of web page usability. In *People and Computers XVI-Memorable Yet Invisible* (pp. 317-335). Springer London.

Cukier, K. (2010). *Data, data everywhere: A special report on managing information*. Economist Newspaper

Cummings, M. L., Bruni, S., Mercier, S., & Mitchell, P. J. (2007). *Automation architecture for single operator, multiple UAV command and control*. MASSACHUSETTS INST OF TECH CAMBRIDGE.

Cummings, M. L., Brzezinski, A. S., & Lee, J. D. (2007). The impact of intelligent aiding for multiple unmanned aerial vehicle schedule management.

DeWitt, A. J. (2010). Examining the order effect of website navigation menus with eye tracking. *Journal of Usability Studies*, *6*(1), 39-47.

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, *1*(3), 271-285.

Di Stasi, L. L., Contreras, D., Cándido, A., Cañas, J. J., & Catena, A. (2011). Behavioral and eye-movement measures to track improvements in driving skills of vulnerable road users: First-time motorcycle riders. *Transportation research part F: traffic psychology and behaviour*, *14*(1), 26-35.

Dorneich, M. C., Whitlow, S. D., Ververs, P. M., & Rogers, W. H. (2003, October). Mitigating cognitive bottlenecks via an augmented cognition adaptive system. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on* (Vol. 1, pp. 937-944). IEEE.

Doyon-Poulin, P., Robert, J., & Ouellette, B. (2012). Review of visual clutter and its effects on pilot performance: A new look at past research. In *Proceedings of the 2012 IEEE/AIAA 31st Digital Avionics Systems Conference (DASC)* (pp. 2D1-1-11). Williamsburg, VA. IEEE.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5), 352-359.

Duffy, E. (1941). An explanation of "emotional" phenomena without the use of the concept "emotion". *The Journal of General Psychology*, *25*(2), 283-293.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological review*, *96*(3), 433.

Durkee, K., Geyer, A., Pappada, S., Ortiz, A., & Galster, S. (2013). Real-Time Workload Assessment as a Foundation for Human Performance Augmentation. In *Foundations of Augmented Cognition* (pp. 279-288). Springer Berlin Heidelberg.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological review*, *66*(3), 183.

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5), 14.

Ellis, G., & Dix, A. (2007). A taxonomy of clutter reduction for information visualisation. *Visualization and Computer Graphics, IEEE Transactions on,13*(6), 1216-1223.

Ellis, K. K., Kramer, L. J., Shelton, K. J., Arthur, J. T., & Prinzel, L. J. (2011, September). Transition of attention in terminal area NextGen operations using synthetic vision systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 55, No. 1, pp. 46-50). SAGE Publications.

Ewing, G. J., Woodruff, C. J., & Vickers, D. (2006). Effects of 'local' clutter on human target detection. *Spatial Vision*, *19*(1), 37-60.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion*, *7*(2), 336.

Fabrikant, S. I., Hespanha, S. R., & Hegarty, M. (2010). Cognitively inspired and perceptually salient graphic displays for efficient spatial inference making.*Annals of the Association of American Geographers*, *100*(1), 13-29.

Falk, J. L., & Bindra, D. (1954). Judgment of time as a function of serial position and stress. *Journal of Experimental Psychology*, *47*(4), 279.

Findlay, J. M. (2004). Eye scanning and visual search. In J. M. Henderson & F. Ferreira (Eds.), The interface of language, vision, and action: eye movements and the visual world. New York: Psychology Press.

Findlay, J. M., & Gilchrist, I. D. (2005). Eye guidance and visual search. In G. Underwood, Cognitive processes in eye guidance, 282-295. New York: Oxford University Press, USA

Fisher, D. L., & Tan, K. C. (1989). Visual displays: The highlighting paradox. *Human Factors*, *31*(1), 17-30.

Freeman, F. G., Mikulka, P. J., Prinzel, L. J., & Scerbo, M. W. (1999). Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. *Biological psychology*, *50*(1), 61-76.

Fuchs, S., Jones, D. L. & Hale, K. S. (2007, October). *Approaches for bias mitigation during imagery analysis.*Paper presented at Augmented Cognition International 2007, Baltimore, MD.

Gaillard, A. W. K. (2001). Stress, workload, and fatigue as three biobehavioral states: A general overview. In P.A. Hancock, & P.A. Desmond (Eds.), Stress, workload, and fatigue. Mahwah, NJ: L. Erlbaum

Gantz, J. & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. In *Proceedings of IDC iView, IDC Anal. Future*

Geyer, T., Von Mühlenen, A., & Müller, H. J. (2007). What do eye movements reveal about the role of memory in visual search?. *The Quarterly Journal of Experimental Psychology*, *60*(7), 924-935.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631-645.

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., & Wichansky, A. M. (2002, March). Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on Eye tracking research & applications* (pp. 51-58). ACM.

Goldberg, H. J., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In J. Hyönä, R. Radach, & H. Deubel (Eds.), The mind's eye: Cognitive and applied aspects of eye movement research (pp. 493-516). Amsterdam: Elsevier

Gopher, D., Armony, L., & Greenspan, Y. (1998). Switching tasks and attention policies and the ability to prepare for such shifts. Paper presented at the 10th Congress of the European Society for Cognitive Psychology, Jerusalem, Israel.

Grahame, M., Laberge, J., & Scialfa, C. T. (2004). Age differences in search of web pages: The effects of link size, link number, and clutter. *Human Factors*, *46*(3), 385-398.

Guznov, S., Matthews, G., Funke, G., & Dukes, A. (2011). Use of the RoboFlag synthetic task environment to investigate workload and stress responses in UAV operation. *Behavior research methods*, *43*(3), 771-780.

Hamill, S. (2011, July). Entire UPMC transplant team missed hepatitis alert. *Pittsburgh Post-Gazette.* Retrieved from http://www.post-gazette.com/news/health/2011/07/10/Entire-UPMC-transplant-team-missed-hepatitis-alert/stories/201107100235

Hammond, K. W., Efthimiadis, E. N., & Laundry, R. (2012). Efficient de-identification of electronic patient records for user cognitive testing. In *Proceedings of the 45th Hawaii International Conference on System Sciences* (pp. 2771-2778). Maui, HI: IEEE.

Hancock, P. A., & Scallen, S. F. (1996). The future of function allocation.*Ergonomics in Design: The Quarterly of Human Factors Applications*, *4*(4), 24-29.

Hancock, P. A., & Szalma, J. L. (2003). Operator stress and display design.*Ergonomics in Design: The Quarterly of Human Factors Applications*, *11*(2), 13-18.

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. Human Factors, 31, 519–537

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*,*52*, 139-183.

He, G., Zhang, J., Liu, D., & Chang, H. (2008). Clutter metric based on the Cramer-Rao lower bound on automatic target recognition. *Applied optics*, *47*(29), 5534-5540.

Hegarty, M., De Leeuw, K., & Bonura, B. (2008, November). What do spatial ability tests really measure. In Proceedings of the 49th Meeting of the Psychonomic Society. Chicago, IL.

Helton, W. S., & Russell, P. N. (2011). Working memory load and the vigilance decrement. *Experimental brain research*, *212*(3), 429-437.

Henderson, J. M. (2007). Regarding scenes. Current Directions in Psychological Science, 16, 219–222.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye movements: A window on mind and brain*, 537-562.

Henderson, J. M., Chanceaux, M., & Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, *9*(1), 32.

Henderson, J.M., Weeks, P.A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during scene viewing. Journal of Experimental Psychology: Human Perception and Performance, 25, 210–228.

Horrey, W. J., & Wickens, C. D. (2004). Driving and side task performance: The effects of display clutter, separation, and modality. *Human Factors, 46*(4), 611-624.

Horowitz, T. S., & Wolfe, J. M. (2003). Memory for rejected distractors in visual search? Visual Cognition, 10, 257–287.

Hou, M., Kobierski, R. D., & Brown, M. (2007). Intelligent adaptive interfaces for the control of multiple UAVs. *Journal of Cognitive Engineering and Decision Making*, *1*(3), 327-362.

Hyrskykari, A. (2006). Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading. *Computers in human behavior*,*22*(4), 657-671.

Inagaki, T. (2003). Adaptive automation: Sharing and trading of control.*Handbook of cognitive task design*, *8*, 147-169. Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. In E. Dykstra-Erickson & K.Manfred Tscheligi (Eds.), *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (pp. 1477–1480). New York: ACM Press.

Inhoff, A. W., & Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005, April). Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 311-320). ACM.

Izzetoglu, M., Izzetoglu, K., Bunce, S., Ayaz, H., Devaraj, A., Onaral, B., & Pourrezaei, K. (2005). Functional near-infrared neuroimaging. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, *13*(2), 153-159.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, *2*(3), 4.

Janelle, C. M., Singer, R. N., & Williams, A. M. (1999). External distraction and attentional narrowing: Visual search evidence. *Journal of Sport & Exercise Psychology*.

Jin, L., Niu, Q., Jiang, Y., Xian, H., Qin, Y., & Xu, M. (2013). Driver sleepiness detection system based on eye movements variables. *Advances in Mechanical Engineering*, *2013*.

Joint Planning and Development Office (JPDO). (2010). Concept of operations for the Next Generation Air Transportation System. Version 3.1, 10 April 2010. Washington, DC: Federal Aviation Administration.

Jonides, J., & Gleitman, H. (1972). A conceptual category effect in visual search: O as letter or as digit. *Perception & Psychophysics*, *12*(6), 457-460.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009, September). Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on* (pp. 2106-2113). IEEE.

Just, M. A., & Carpenter, P. A. (1978). Inference processes during reading: Reflections from eye fixations. In J. W. Senders, D. F. Fisher, & R. A. Monty (Eds.), Eye movements and the higher psychological functions. Hillsdale, N.J.: Erlbaum.

Kaber, D. B., Alexander, A. L., Stelzer, E. M., Kim, S. H., Kaufmann, K., & Hsiang, S. (2008). Perceived clutter in advanced cockpit displays: measurement and modeling with experienced pilots. *Aviation, space, and environmental medicine*, *79*(11), 1007-1018.

Kaber, D. B., & Riley, J. M. (1999). Adaptive automation of a dynamic control task based on secondary task workload measurement. *International journal of cognitive ergonomics*, *3*(3), 169-187.

Kaber, D. B., Riley, J. M., Tan, K. W., & Endsley, M. R. (2001). On the design of adaptive automation for complex systems. *International Journal of Cognitive Ergonomics*, *5*(1), 37-57.

Kaber, D. B., Wright, M. C., Prinzel, L. J., & Clamann, M. P. (2005). Adaptive automation of human-machine system information-processing functions.*Human Factors: The Journal of the Human Factors and Ergonomics Society*,*47*(4), 730-741.

Kardan, S., & Conati, C. (2012). Exploring gaze data for determining user learning with an interactive simulation. In *User Modeling, Adaptation, and Personalization* (pp. 126-138). Springer Berlin Heidelberg.

Kim, S.-H., & Kaber, D. B. (2009). Assessing the effects of conformal terrain features in advanced head-up displays on pilot performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 36-40). San Antonio, TX: Sage Publications.

Kim, S. H., Prinzel, L. J., Kaber, D. B., Alexander, A. L., Stelzer, E. M., Kaufmann, K., & Veil, T. (2011). Multidimensional measure of display clutter and pilot performance for advanced head-up display. *Aviation, space, and environmental medicine*, *82*(11), 1013-1022.

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an" aid" can (and should) go unused. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(2), 221-242.

Kaber, D., Kim, S.-H., Kaufmann, K., Alexander, A., Steltzer, E., & Hsiang, S. (2008). *Modeling the effects of HUD visual properties, pilot experience and flight scenario on a multi-dimensional measure of clutter*. Hampton, VA: Langley Research Center.

Kaber, D. B., Naylor, J. T., Gil, G. H., Pankok, C., & Kim, S.- H. (2013). Influence of flight domain and cockpit display dynamics on pilot perceived clutter. *Journal of Aerospace Information Systems*, *10*(12), 550-559.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. *Varieties of attention*, *1*, 29.

Katsanos, C., Tselios, N., & Avouris, N. (2010). Evaluating website navigability: validation of a tool-based approach through two eye-tracking user studies. *New review of Hypermedia and Multimedia*, *16*(1-2), 195-214.

Kaufmann, K. A., & Kaber, D. B. (2010). The influence of individual differences in perceptual performance on pilot perceptions of head-up display clutter. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 70-74). San Francisco, CA: Sage Publications.

Keeble, R. J., & Macredie, R. D. (2000). Assistant agents for the world wide web intelligent interface design challenges. *Interacting with Computers*, *12*(4), 357-381.

Kim, N. Y., & Sundar, S. S. (2010). Relevance to the rescue: Can “smart ads” reduce negative response to online ad clutter? *Journalism & Mass Communication Quarterly*, *87*(2), 346-362.

Klein, R. M. (2000). Inhibition of return. *Trends in cognitive sciences*, *4*(4), 138-147.

Konstantopoulos, P., Chapman, P., & Crundall, D. (2010). Driver's visual attention as a function of driving experience and visibility. Using a driving simulator to explore drivers’ eye movements in day, night and rain driving.*Accident Analysis & Prevention*, *42*(3), 827-834.

Kozma, L., Klami, A., & Kaski, S. (2009, November). GaZIR: gaze-based zooming interface for image retrieval. In *Proceedings of the 2009 international conference on Multimodal interfaces* (pp. 305-312). ACM.

Kroft, P., & Wickens, C. D. (2001). *The display of multiple geographical data bases: Implications of visual attention*. Technical Report ARL-01-2/NASA-01-2). Savoy, IL: Aviation Research Laboratory.

Kroft, P., & Wickens, C. D. (2002). Displaying multi-domain graphical database information: An evaluation of scanning, clutter, display size, and user activity. *Information Design Journal, 11*(1), 44-52.

Kumar, M., Winograd, T., & Paepcke, A. (2007, April). Gaze-enhanced scrolling techniques. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems* (pp. 2531-2536). ACM.

Lee, J. D., Hoffman, J. D., & Hayes, E. (2004, April). Collision warning design to mitigate driver distraction. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 65-72). ACM.

Lee, C. K., Yoo, S., Park, Y. J., Kim, N., Jeong, K., & Lee, B. (2005). Using neural network to recognize human emotions from heart rate variability and skin resistance. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference* (pp. 5523–5525). Los Alamitos, CA: IEEE Computer Society

Levi, D. M. (2008). Crowding: An essential bottleneck for object recognition. A mini-review. Vision Research, 48, 635–654.

Liang, Y., Reyes, M. L., & Lee, J. D. (2007). Real-time detection of driver cognitive distraction using support vector machines. *Intelligent Transportation Systems, IEEE Transactions on*, *8*(2), 340-350.

Liao, M. J., Wu, Y., & Sheu, C. F. (2013). Effects of perceptual complexity on older and younger adults' target acquisition performance. *Behaviour & Information Technology,* doi: 10.1080/0144929X.2013.847974.

Lohrenz, M. C., Layne, G. L., Edwards, S. S., Gendron, M. L., & Bradley, J. T. (2006). Feature clustering to measure clutter in electronic displays. In *Proceedings of the 12th Industry, Engineering, and Management Systems conference*. Cocoa Beach, FL: The Association for Industry, Engineering, and Management Systems.

Lohrenz, M. C., Trafton, J. G., Beck, M. R., & Gendron, M. L. (2009). A model of clutter for complex, multivariate geospatial displays. *Human Factors, 51*(1), 90-101.

Lovallo, W. R. (1997). Stress and health: Biological and psychological interactions. Thousand Oaks, CA: Sage

Lusted, M. (2012). *The Three Mile Island Nuclear Disaster (Essential Events).* Essential Library

Mack, M.L., & Oliva, A. (2004). Computational estimation of visual complexity. In *Proceedings of the 12th Annual Object, Perception, Attention, and Memory Conference*. Minneapolis, Minnesota.

Markowetz, F. (2003). Classification by Support Vector Machines, Max-PlanckInstitute for Molecular Genetics- Computational Molecular Biology, Berlin, pp. 12.

Marquard, J. L., Jo, J., Henneman, P. L., Fisher, D. L., & Henneman, E. A. (2012). Can Visualizations Complement Quantitative Process Analysis Measures? A Case Study of

Nurses Identifying Patients Before Administering Medications. *Journal of Cognitive Engineering and Decision Making*, 155534341245768

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L., Huggins, J., Gilliland, K., et al. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress and worry. Emotion, 2, 315–340

McFarlane, D. C. (2002). Comparison of four primary methods for coordinating the interruption of people in human-computer interaction, *Human-Computer Interaction, 17*, 63–139.

McGrath, J. E. (1976). Stress and behavior in organizations. *Handbook of industrial and organizational psychology*, *1351*, 1396.

McKenzie, M., Wong, S., & Gibbins, D. (2013). An empirical evaluation of infrared clutter for point target detection algorithms. In F. A. Sadjadi, F. & A. Mahalanobis (Eds.), *Proceedings of SPIE Vol. 874409: Automatic Target Recognition XXIII*. Baltimore, MD: International Society for Optics and Photonics (SPIE). doi:10.1117/12.2015668.

McMullen, R. (2011, August). Badly Designed Electronic Medical Records Can Kill You. Fast Code Design. Retrieved from http://www.fastcodesign.com/1664763/badly-designed-electronic-medical-records-can-kill-you

Miller, C. A., Funk, H., Goldman, R., Meisner, J., & Wu, P. (2005, July). Implications of adaptive vs. adaptable UIs on decision making: Why “automated adaptiveness” is not always the right answer. In *Proc. of the 1st inter. conf. on augmented cognition*.

Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control.*Human Factors: The Journal of the Human Factors and Ergonomics Society*,*49*(1), 57-75.

Moacdieh, N., & Sarter, N. (2014). Display clutter: A review of definitions and measurement techniques. *Human Factors*. doi: 0018720814541145.

Mogg, K., Bradley, B. P., & Hallowell, N. (1994). Attentional bias to threat: Roles of trait anxiety, stressful events, and awareness. *The Quarterly Journal of Experimental Psychology*, *47*(4), 841-864.

Mokhtar, R., Abdullah, S. N. H. S., & Zin, N. A. M. (2011, June). Classifying modality learning styles based on Production-Fuzzy Rules. In *Pattern Analysis and Intelligent Robotics (ICPAIR), 2011 International Conference on* (Vol. 1, pp. 154-159). IEEE.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*(1), 44.

Munn, S. M., Stefano, L., & Pelz, J. B. (2008, August). Fixation-identification in dynamic scenes: Comparing an automated algorithm to manual coding. In*Proceedings of the 5th symposium on Applied perception in graphics and visualization* (pp. 33-42). ACM.

Murata, A. (2004). Foveal task complexity and visual funneling. *Human Factors, 46*(1), 135–141.

Naylor, J. (2010). The influence of dynamics, flight domain and individual flight training & experience on pilot perception of clutter in aviation displays (Masters thesis). Available from http://repository.lib.ncsu.edu/ir/handle/1840.16/6103

Neider, M. B., & Zelinsky, G. J. (2011). Cutting through the clutter: Searching for targets in evolving complex scenes. *Journal of vision*, *11*(14), 7.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton. Century, Crofts.

NextGen Integration and Implementation Office. (2010). NextGen Implementation Plan. Washington, DC: Federal Aviation Administration.

Nielsen, J.(1993). *Usability engineering*. Academic Press, Boston, MA.

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision research*, *11*(9), 929-IN8.

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: detecting the snake in the grass. *Journal of experimental psychology: general*, *130*(3), 466.

O'Hara, K., Mitchell, A. S., & Vorbau, A. (2007). Consuming video on mobile devices. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM, New York, NY, 857-866. doi:http://doi.acm.org/10.1145/1240624.1240754

Oliva, A., Mack, M. L., Shrestha, M., & Peeper, A. (2004). Identifying the perceptual dimensions of visual complexity of scenes. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 1041-1046). Chicago, IL: Cognitive Science Society.

Olson, W. A., & Sarter, N. B. (2000). Automation management strategies: Pilot preferences and operational experiences. *The International Journal of Aviation Psychology*, *10*(4), 327-341.

Oppermann, R. (1994). Adaptively supported adaptability. *International Journal of Human-Computer Studies*, *40*(3), 455-472.

Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision research*, *34*(13), 1703-1721.

Parasuraman, R., Bahri, T., Deaton, J. E., Morrison, J. G., & Barnes, M. (1992). *Theory and design of adaptive automation in aviation systems*. Catholic Univ Of America Washington Dc Cognitive Science Lab.

Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *38*(4), 665-679.

Parasuraman, R., Warm, J. S., & See, J. E. (1998). Brain systems of vigilance. In R. Parasuraman (Ed.), The attentive brain (pp. 221–256). Cambridge, MA: MIT Press

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, *42*(1), 107-123.

Pavlas, D., Lum, H., & Salas, E. (2012). How to Build a Low-Cost Eye-Tracking System. *Ergonomics in Design: The Quarterly of Human Factors Applications*,*20*(1), 18-23.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. Journal of Vision, 4(12), 12.

Peng, W., Ward, M. O., & Rundensteiner, E. A. (2004). Clutter reduction in multi-dimensional data visualization using dimension reordering.In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS)* (pp. 89-96). Austin, TX: IEEE.

Phillips, R. J., & Noyes, L. (1982). An investigation of visual clutter in the topographic base of a geological map. *The Cartographic Journal*, *19*(2), 122-132.

President's Commission on the Accident at Three Mile Island. (1979). *The need for change, the legacy of TMI: report of the President's Commission on the Accident at Three Mile Island*. The Commission.

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2000). A closed-loop system for examining psychophysiological measures for adaptive task allocation. *The International journal of aviation psychology*, *10*(4), 393-410.

Pomplun, M., & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of HCI International 2003* (Vol. 3, pp. 542–546). Mahwah, NJ: Erlbaum.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research.*Encyclopedia of human computer interaction*, *1*, 211-219.

Pope, A.T., Bogart, E.H., Bartolome, D.S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. Biological Psychology 40, 187–195.

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2008, April). Predicting postcompletion errors using eye movements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 539-542). ACM.

Rauschenberger, R. (2003). Attentional capture by auto- and allo-cues. Psychonomic Bulletin & Review, 10(4), 814−842.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.

Razmjou, S. (1996). Mental workload in heat: toward a framework for analyses of stress states. *Aviation, space, and environmental medicine*.

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied*, *6*(1), 31.

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, *9*(2), 119.

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. Annual Review of Neuroscience, 27, 611−647.

Rosenholtz, R., Li, Y., Mansfield, J., & Jin, Z. (2005). *Feature congestion: a measure of display clutter.* In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 761-770). Portland, OR: ACM Press.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision, 7*(2), 1-22. doi: 10.1167/7.2.17.

Rothrock, L., Koubek, R., Fuchs, F., Haas, M., & Salvendy, G. (2002). Review and reappraisal of adaptive interfaces: toward biologically inspired paradigms.*Theoretical Issues in Ergonomics Science*, *3*(1), 47-84.

Rouse, W. B. (1988). Adaptive aiding for human/computer control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *30*(4), 431-443.

Rouse, W. B., Geddes, N. D., & Curry, R. E. (1987). An architecture for intelligent interfaces: Outline of an approach to supporting operators of complex systems. *Human-Computer Interaction*, *3*(2), 87-122.

Russell B. (2005). Neural network applications in geophysics. Hampson-Russell Software, a Veritas Company, Calgary, Canada, 1-3

Salas, E., Driskell, J. E., & Hughes, S. (1996). The study of stress and human performance. *Stress and human performance(A 97-27090 06-53), Mahwah, NJ, Lawrence Erlbaum Associates, Publishers, 1996,*, 1-45.

Salem, S., Halford, C., Moyer, S., & Gundy, M. (2009). Rotational clutter metric. *Optical Engineering, 48*(8), 086401-086401-086411.

Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications* (pp. 71-78). ACM.

Samel, A., Vejvoda, M., & Maass, H. (2004). Sleep deficit and stress hormones in helicopter pilots on 7-day duty for emergency medical services.*Aviation, space, and environmental medicine*, *75*(11), 935-940

Samal, L., Wright, A., Wong, B. T., Linder, J. A., & Bates, D. W. (2011). Leveraging electronic health records to support chronic disease management: the need for temporal data views. *Informatics in primary care*, *19*(2), 65-74.

Sarter, N. B. (2005, September). Graded and multimodal interruption cueing in support of preattentive reference and attention management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, No. 3, pp. 478-481). SAGE Publications.

Sarter, N. (2007). Coping with complexity through adaptive interface design. In*Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments* (pp. 493-498). Springer Berlin Heidelberg.

Sarter, N. B., Billings, C. E., & Woods, D. D. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed., pp. 1926–1943). New York: Wiley.

Scerbo, M. (2001). Adaptive automation. *Neuroergonomics: The brain at work*, 239-252.

Schmorrow, D. D., and Kruse. A. A. (2004). Augmented Cognition. In W.S. Bainbridge (Ed.), Berkshire Encyclopedia of Human Computer Interaction. Great Barrington, MA: Berkshire Publishing Group, 54-59.

Schnell, T., Kwon, Y., Merchant, S., & Etherington, T. (2004). Improved flight technical performance in flight decks equipped with synthetic vision information system displays. *The international Journal of Aviation psychology*, *14*(1), 79-102.

Schons, V., & Wickens, C. D. (1993). *Visual separation and information access in aircraft display layout* (Technical Report ARL-93-7/NASA-A 3 I-93-1). Savoy, IL: University of Illinois, Aviation Research Lab.

Schmidt, R. W. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. W. Schmidt (Ed.), Attention and awareness in foreign language learning (pp. 1–63). Honolulu, HI: University of Hawai‘i, National Foreign Language Resource Center.

Sexton, J. B., Thomas, E. J., & Helmreich, R. L. (2000). Error, stress, and teamwork in medicine and aviation: cross sectional surveys. *Bmj*, *320*(7237), 745-749.

Shen, M., Carswell, M., Santhanam, R., & Bailey, K. (2012). Emergency management information systems: Could decision makers be supported in choosing display formats?. *Decision Support Systems*, *52*(2), 318-330.

Shneiderman, B. (1997). Designing information-abundant web sites: issues and recommendations. *International Journal of Human-Computer Studies*, *47*(1), 5-29.

Shneiderman, B., Plaisant, C., Cohen, M., & Jacobs, S. (2009). *Designing the user interface*: *Strategies for Effective Human-Computer Interaction (5th Edition).* Upper Saddle River, New Jersey: Prentice Hall.

Shontz, W. D., Trumm, G. A., & Williams, L. G. (1971). Color coding for information location. *Human Factors*, *13*(3), 237-246.

Singh, H., Spitzmueller, C., Petersen, N. J., Sawhney, M. K., & Sittig, D. F. (2013). Information overload and missed test results in electronic health record–based settings. *JAMA internal medicine*, *173*(8), 702-704.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *30*(4), 445-459.

St John, M., Manes, D. I., Smallman, H. S., Feher, B., & Morrison, J. G. (2004). *An intelligent threat assessment tool for decluttering naval air defense displays* (No. TR-1915). Space And Naval Warfare Systems Center San Diego Ca.

Staal, M. A. (2004). Stress, cognition, and human performance: A literature review and conceptual framework.

Stanney, K. M. (2004): Handbook of virtual environments. Mahwah, NJ: Lawrence Erlbaum Associates.

Stanney, K., Samman, S., Reeves, L., Hale, K., Buff, W., Bowers, C., & Lackey, S. (2004). A paradigm shift in interactive computing: Deriving multimodal design principles from behavioral and neurological foundations.*International Journal of Human-Computer Interaction*, *17*(2), 229-257.

Steichen, B., Carenini, G., & Conati, C. (2013, March). User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 317-328). ACM.

Stokes, A. F., Kemper, K. L., & Marsh, R. (1992). *Time-stressed flight decision making: A study of expert and novice aviators*.

Stokes AF, Kite K (2001). On grasping a nettle and becoming emotional. In *Stress, workload, and fatigue*. Edited by Hancock PA, Desmond PA. Mahwah, NJ: L. Erlbaum; 2001.

Sukaviriya, P., & Foley, J. D. (1990, August). Coupling a UI framework with automatic generation of context-sensitive animated help. In *Proceedings of the 3rd annual ACM SIGGRAPH symposium on User interface software and technology* (pp. 152-166). ACM.

Tepas, D. I., & Price, J. M. What is stress and what is fatigue. *Stress, workload, and fatigue. Mahwah, NJ: L. Erlbaum*.

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection.*Acta psychologica*, *135*(2), 77-99.

Theeuwes, J., & Belopolsky, A. (2010). Top–down and bottom–up control of visual selection: controversies and debate. In V. Coltheart (Ed.), Tutorials in visual cognition (pp. 67−92). New York: Psychology Press.

Theeuwes, J., Kramer, A. F., Hahn, S., & Irwin, D. E. (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects.*Psychological Science*, *9*(5), 379-385.

Toet, A. (2010). Structural similarity determines search time and detection probability. *Infrared Physics & Technology, 53*(6), 464-468.

Treisman, R. (1986). Identification of a protein-binding site that mediates transcriptional response of the c-fos gene to serum factors. *Cell*, *46*(4), 567-574.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

Trukenbrod, H. A., & Engbert, R. (2007). Oculomotor control in a sequential search task. *Vision research*, *47*(18), 2426-2443.

Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information* (Vol. 2). Cheshire, CT: Graphics press.

Tullis, T. S. (1983). The formatting of alphanumeric displays: A review and analysis. *Human Factors, 25*(6), 657-682.

Tullis, T. S. (1988). Screen design. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 377–411). Amsterdam: North-Holland.

van den Berg, R., Cornelissen, F. W., & Roerdink, J. (2009). A crowding model of visual clutter. *Journal of Vision, 9*(4), 1-11. doi: 10.1167/9.4.24.

van Galen, G. P., & van Huygevoort, M. (2000). Error, stress and the role of neuromotor noise in space oriented behaviour. *Biological psychology, 51*(2), 151-171.

Vapnik V. (1995), "The nature of statistical learning theory," Springer-Verlag, New-York.

Vasmatzidis, I., Schlegel, R. E., & Hancock, P. A. (2002). An investigation of heat stress effects on time-sharing performance. *Ergonomics, 45*(3), 218-239.

Ververs, P. M., & Wickens, C. D. (1998). Head-up displays: Effect of clutter, display intensity, and display location on pilot performance. *The International Journal of Aviation Psychology, 8*(4), 377-403.

Vlaskamp, B. N., & Hooge, I. T. C. (2006). Crowding degrades saccadic search performance. Vision Research, 46, 417–425.

Wang, X., & Zhang, T. (2011). Clutter-adaptive infrared small target detection in infrared maritime scenarios. *Optical Engineering, 50*(6), 067001-12.

Welford, A. T. (1973). Stress and performance. *Ergonomics, 16*(5), 567-580.

Westerbeek, H.G.W., & Maes, A. (2011). Referential scope and visual clutter in navigation tasks. In K. van Deemter, A. Gatt, R. van Gompel, & E.J. Krahmer (Eds.), *Proceedings of the Workshop on the Production of Referring Expressions (PRE-CogSci 2011)* (pp. 1-6). Boston, Massachusetts: Cognitive Science Society.

Westerman, S. J., & Haigney, D. (2000). Individual differences in driver stress, error and violation. *Personality and Individual Differences, 29*(5), 981-998.

Whitney, D., & Levi, D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends in cognitive sciences,15*(4), 160-168.

Wickens, C. (1994). Designing for situation awareness and trust in automation. In Proceedings of the IFAC Conference. Baden-Baden, Germany, pp. 174-179.

Wickens, C. D., & Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *The International Journal of Aviation Psychology, 19*(2), 182-199.

Wickens, C. D., Alexander, A. L., Horrey, W. J., Nunes, A., & Hardy, T. J. (2004). Traffic and flight guidance depiction on a synthetic vision system display: The effects of clutter on performance and visual attention allocation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 218-222). New Orleans, LA: Sage Publications.

Wickens, C. D., Ambinder, M. S., Alexander, A. L., & Martens, M. (2004). The role of highlighting in visual search through maps. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp 1895-1899). New Orleans, LA: Sage Publications.

Wickens, C. D., & Carswell, C. M.. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors, 37*(3), 473-494.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering Psychology and Human Performance* (4th Edition). Pearson Education, Inc., New Jersey.

Wickens, C. D., & McCarley, J. S. (2008). Applied attention theory. Boca Raton, FL: Taylor & Francis.

Wiener, E. L. (1989). Human factors of advanced technology (glass cockpit) transport aircraft.

Williams, C. C., & Pollatsek, A. (2007). Searching for an O in an array of Cs: Eye movements track moment-to-moment processing in visual search.*Perception & psychophysics*, *69*(3), 372-381.

Wofford, J. C., & Daly, P. S. (1997). A cognitive-affective approach to understanding individual differences in stress propensity and resultant strain.*Journal of Occupational Health Psychology*, *2*(2), 134.

Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision research*, *34*(9), 1187-1195.

Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in cognitive sciences*, *7*(2), 70-76.

Wolfe, J. M., Oliva, A., Horowitz, T. S., Butcher, S. J., & Bompas, A. (2002). Segmentation of objects from backgrounds in visual search tasks. *Vision research*, *42*(28), 2985-3004.

Xing, J. (2007). *Information complexity in air traffic control displays* (DOT/FAA/AM-07/26). Technical Report to the Federal Aviation Administration. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.8718&rep=rep1&type=pdf.

Yamani, Y., & McCarley, J. S. (2010). Visual search asymmetries within color-coded and intensity-coded displays. *Journal of Experimental Psychology: Applied*, *16*(2), 124.

Yantis, S. (1993). Stimulus-driven attentional capture. *Current Directions in Psychological Science*, *2*(5), 156-161.

Yarbus, A. L. (1967). *Eye movements and vision* (Vol. 2, No. 5.10). L. A. Rigss (Ed.). New York: Plenum press.

Yeh, M., Brandenburg, D., Wickens, C. D., & Merio, J. (2000). *Up or down? A comparison of helmet mounted display and hand held display tasks with high clutter imagery* (No. ARL-00-11/FED-LAB-00-3). Illinois University at Urbana-Champaign, Savoy Aviation Research Lab. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA436770

Yeh, M., Merlo, J. L., Wickens, C. D., & Brandenburg, D. L. (2003). Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(3), 390-407.

Yeh, M., & Wickens, C. D. (2001). Attentional filtering in the design of electronic map displays: A comparison of color coding, intensity coding, and decluttering techniques. *Human Factors*, *43*(4), 543-562.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology and psychology*,*18*(5), 459-482.

Zelinsky, G.J. (2001). Eye movements during change detection: Implications for search constraints, memory limitations, and scanning strategies. Perception & Psychophysics, 63, 209-225.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition.*Psychological review*, *115*(4), 787.

Zelinsky, G. J., & Sheinberg, D. L. (1997). Eye movements during parallel–serial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(1), 244.

Zeng, Q., Cimino, J. J., & Zou, K. H. (2002). Providing concept-oriented views for clinical data using a knowledge-based system: An evaluation. *Journal of the American Medical Informatics Association*, *9*, 294-305. doi: 10.1197/jamia.M1008

Zhu, B., & Sun, X. (2012). Effects of Superposition on Oculomotor Guidance and Target Recognition. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (pp. 333-337). Boston, MA: Sage Publications. doi: 0.1177/1071181312561077

Zikopoulos, P. C.; Deroos, D.; Parasuraman, K.; et al. (2013). Harness the Power of Big Data: The IBM Big Data Platform. New York: McGraw-Hill.

Zuschlag, M. (2004). Quantification of visual clutter using a computational model of human perception: An application for head-up displays. In *Proceedings of the Human Performance, Situation Awareness and Automation II Conference.* Daytona Beach, FL.

CHAPTER 2

Analyzing the Problem of Clutter in the Emergency Department

The focus of this dissertation is display clutter in data-rich, stressful, and safety-critical domains where operators need to find information quickly and reliably. The hospital emergency department (ED) is a perfect example of such a domain. Physicians in the ED largely rely on electronic medical records (EMRs) to obtain the patient information they need to perform medical diagnoses. It is critical to ensure that this information can be extracted without delay and that no relevant information is overlooked. However, concerns have been raised regarding display clutter and data overload in current EMRs (Singh, Spitzmueller, Petersen, Sawhney, & Sittig, 2013; Van Vleck, Stein, Stetson, & Johnson, 2007).

These concerns were echoed by physicians at the ED of the University of Michigan (UM) Health System, where a new EMR system was recently implemented. Two physicians and one resident at the UM ED agreed to collaborate on this research study in an effort to address problems with clutter they are encountering.

The first part of this chapter will explain the choice of application domain based on a review of literature that shows that EMR clutter can be detrimental to performance, particularly in the stressful and high-tempo ED environment. The second part of this chapter will focus specifically on the EMR currently used at the UM ED. That section will describe preliminary steps in this dissertation research which served to understand how the UM ED operates, what role the EMR plays is in this environment, and what clutter-related problems have been experienced by ED physicians to date.

High Clutter and High Stress: The Case for Selecting ED EMRs

The two main factors that make medical EMR displays and the ED environment good choices for this research are 1) the reported high levels of clutter in EMRs and 2) the high stress that characterizes EDs.

Clutter in EMRs

EMRs (also known as electronic health records or EHRs) are now used by more than 50% of professionals and nearly 80% of hospitals, a 100% increase from 2012 (US Department of Health and Human Services, 2013). EMRs have grown in prominence primarily as a result of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, which pledged monetary incentives to medical providers who demonstrate meaningful use of EMRs (Blumenthal, 2010). Thus, the adoption of EMRs is now less of a choice and more of a necessity for medical providers.

EMR systems typically provide a wide range of functionalities, including computerized physician order entry (CPOE), decision support (such as notifications of possible drug interactions), and the storage and presentation of patient medical data. This medical data includes laboratory results, medical/surgical history, physician/nurse notes, allergy information, vital signs, and medication lists, among others. Physicians refer to this information in order to perform accurate diagnoses and identify patient treatment options. When the HITECH act was enacted, the expectation was that these EMR functions would contribute to increased efficiency and safety of operations in hospital environments (Blumenthal & Glaser, 2007).

However, current EMRs have failed to fully support this goal. A recent study estimated the number of deaths related to preventable medical errors at 440,000 per year (James, 2013), up from the 98,000 that was estimated fifteen years ago by the seminal Institute of Medicine “To Err Is Human” study (Kohm, Corrigan, & Donaldson, 1999). It seems that the introduction of EMRs has done little to prevent the proliferation of medical errors. Instead, the growth in EMR use has been paralleled by an unanticipated increase in EMR-related errors or adverse events. In 2011, more than 1,000 adverse events associated with EMRs were reported to the Pennsylvania Patient Safety Authority (PPSA), twice the number reported in the previous year (PPSA, 2012). One study highlighted how EMRs and other health informatics technology “can create new hazards in the already complex delivery of health care” (Institute of Medicine, 2012). Increasingly, there is a growing realization that errors involving EMRs are not due to a lack of training or negligence on the part of the doctors but rather stem from poor EMR display design (Karsh, Weinger, & Abbott, 2010). This poor display of information can lead to missing or misinterpreting critical data, which, in turn, can lead to incorrect or delayed medical care (Farley et al., 2013). Examples of medical errors that have been linked to a large amount of task-irrelevant and/or poor display organization include giving a patient an incorrect drug and performing surgery on the wrong patient (Maya, 2009). Design-related errors thus present a considerable threat to the efficient operation of hospitals and to the health and well-being of patients.

One of the most prominent issues related to EMR design is display clutter. Several factors contribute to this problem. Advances in technology have led to the availability of large amounts of medical data (Zeldes & Baum, 2011), and federal regulations require the collection and storage of this data in order for a medical provider to qualify for meaningful use incentives (Blumenthal & Tavenner, 2010). In addition, the aging population means that many patients have an extensive medical history that is captured by the EMR (Ludwick & Doucette, 2009). Also, medical decision support systems – algorithms designed to aid physicians in making choices and detecting conflicts such as drug-drug interactions – have also led to even more information being presented in EMRs in the form of alerts and warnings (Singh et al., 2013). Together these factors have raised concerns that there is currently an excess and possibly harmful amount of data in EMRs (Hammond, Efthimiadis, & Laundry, 2012; Weir & Nebeker, 2007; Zeng, Cimino, & Zou, 2002; Zhang, Pakhomov, McInnes, & Melton 2011). The poor organization of this data (Ahmed, Chandra, Herasevich, Gajic, & Pickering 2011) and the fact that much of it is irrelevant to physicians' tasks (Bobillo, Delgado, & Gómez-Romero 2008) have added to concerns about clutter.

Empirical studies show performance effects such as increased EMR use time and difficulty finding relevant information (Duftschmid et al., 2013; Murphy, Reis, Sittig, & Singh 2012), as well as errors when extracting information (Zeng et al., 2002). Physicians report higher workload ratings (Ahmed et al., 2011), a higher likelihood of missing more data (Singh et al., 2013), an inability to obtain the "big picture" of a patient's medical condition (Tully et al., 2013), and difficulty separating information from noise (van Vleck et al., 2007). One researcher likened using the current cluttered EMRs to carrying out a complicated "information treasure hunt" (Eiblang, 2009).

The concern about EMR clutter and the selection of EMRs as the representative complex domain for this research thus seem to be justified. Moreover, given the stipulations of the HITECH act, the reliance on EMRs will continue to grow despite any current design issues with EMRs. There is currently little oversight and guidance for the design of EMRs, and the little information available does not emphasize patient safety (Hoffman & Podgurski, 2011). To inform the design of better EMRs and maintain the efficiency and safety of hospital operations, it is therefore crucial to better understand EMR clutter and support physicians in their medical diagnoses.

Hospital EDs: The Problem of Stress

The issue of timely and accurate diagnoses is particularly important in the ED, which is characterized by time pressure, risk, and criticality (Kachalia et al., 2007; Levin et al., 2006). Physicians in the ED are required to diagnose, rapidly and accurately, a patient's condition in order to administer the proper care. Delays or errors in finding and interpreting data in ED EMRs need to be avoided at all costs as they can lead to severe consequences for patient health (Croskerry et al., 2004).

Several inherent facets of the ED make errors "incredibly common" (Clark, 2013). For example, patient overcrowding and high physician workload (Trzeciak & Rivers, 2003), contribute to considerable time constraints. Moreover, patients that present to the ED often have high-acuity and life-threatening conditions, meaning that medical procedures would need to be performed under significant time pressure (Kachalia et al., 2007). In contrast to other types of care, ED physicians are usually unfamiliar with patients and their medical history, some of which may be missing (Fordyce et al., 2003). In addition, frequent interruptions in the ED can make it difficult for physicians to resume and complete interrupted tasks in a timely fashion (Coiera, Jayasuriya, Hardy, Bannan, & Thorpe 2002). The 24-hour operation of the ED also necessitates frequent shift changes and handovers, a process which has been implicated in numerous accidents in other complex domains (Brazier & Pacitti, 2008). The timing and length of shifts can also lead to decreased levels of awareness as a result of disrupted circadian cycles (Kuhn, 2001).

All of these factors contribute to making the ED a highly stressful environment (Levin et al., 2006) where it is critical that clutter is minimized to support the information needs of physicians. The next step in this research was to examine these tasks and information needs in the specific ED under study, the UM ED.

Analyzing and Understanding the Application Domain

Several steps were taken to fully characterize and understand the tasks and workflow in the ED, the affordances of the EMR used, and physicians' perspectives on clutter in their EMRs. These steps consisted of observations in the ED, a survey deployed among ED medical personnel, and interviews with ED physicians.

ED Observations: Learning About EMR Functions and Their Use

Prior to beginning the observations, I completed formal EMR training offered by the UM ED in order to understand the basic functioning and affordances of the system. I also obtained

HIPAA (Health Insurance Portability and Accountability Act) clearance to be able to access patient data. A total of nine days of observations took place and consisted of shadowing two collaborating physicians and the collaborating resident. The observation sessions included busy morning, late night, and weekend shifts to get a comprehensive impression of ED operations. This provided a detailed understanding of the how physicians use the EMR to perform their diagnoses.

The current EMR. The EMR used in the UM ED consists of several pages that provide a wide array of functions for physicians (note that, due to copyright restrictions, it is not possible to show any of the EMR screens). All medical personnel in the ED use this same EMR. The medical personnel includes attending physicians (fully trained medical doctors who are the highest authority in the ED), physician assistants (i.e., they have obtained a license to practice medicine under a physician's supervision), residents (i.e., they have obtained their medical degree and are training in emergency medicine), and nurses.

The focus here will be on the use of the EMR to extract medical information, as opposed to the functions related to information entry (e.g., placing orders, documenting encounters). The main pages that are used for information extraction are:

- *Chart Review*: physicians can use this page to access all the previous notes that document patients' prior medical encounters, including visits to the ED or their own medical care providers. The notes are all displayed in a list; clicking on any entry will show the contents of the note.
- *Clinical Snapshot*: this page provides a summary of all of a person's main medical data, including the current triage note, vital signs, allergy information, and ordered medications. There are many tabs that make up this page; notable among them is the

Medical History page, which physicians use to learn about a patient's previous complaints and surgical procedures. In addition, the Medical History page contains a record of the patient's social history (e.g., smoker or non-smoker) and family history. All of the entries are displayed in a list. The Medications tab is another important one in the Clinical Snapshot, and it has a list of the medications that patients are currently taking and have taken in the past. Finally, the Triage Summary tab contains the triage note and measured vital signs (the process of triage will be discussed later).

- *Imaging/EKG results*: this page allows physicians to view the results of the medical imaging tests that were performed, such as x-ray results.
- *Orders*: physicians use this page to view the details of any orders (e.g., blood tests, medications, laboratory procedures, etc.) that have been placed. The orders are all displayed as a list.
- *Patient Timeline*: usually filled in by the nurses attending to the patient, this page documents, in a list shown in chronological order, all the steps that have been taken since the patient presents to the ED. Examples of entries in the list could be the patient's recorded vital signs or test that was taken.
- *Laboratory Results:* this page is used to view the results of any laboratory tests that were ordered, such as the results of a blood test. The values are displayed in a table, with abnormal results in red.

Workflow in the ED. The main goal of emergency medicine is to diagnose a patient's condition and determine the necessary course of action. This can be as simple as recommending an over-the-counter medication and discharging the patient, or it might require performing lifesaving procedures and admitting patients to hospital. While it is impossible to establish one fixed procedure that is adopted for all patients, the typical sequence of operations in the UM ED tends to follow the order depicted in Figure 2.1. The different main stages are described below:

1. *New patient in ED*. Patients first go to the *triage* area, where they are given a quick assessment by a triage nurse. The patient's vital signs, such as heart rate and blood pressure, are measured and he/she describes the problem to the nurse. The nurse then writes a triage note – a brief description of the problem – and enters that note and the vital signs into the system. The nurse also assigns an acuity rating to this person, with 1 being the highest acuity and 3 the lowest; this suggests to the physician the order that patients need to be attended to. Depending on how many people there are in the ED at a time, a person with an acuity rating of 3 may wait a long time to be seen if there are several people with higher acuity.
2. *Gather information about patient.* Based in part on the acuity level, the physician will select a patient to attend to next (note that physicians are typically taking care of multiple patients at the same time). The physician reads the triage note and then, depending on the particular patient case, the physician could look up the patient's medical history, prior medications, older visits to the ED, etc.
3. *Make initial hypothesis*. After reading the triage note and some or all of the patient's medical history, the physician can then formulate an initial hypothesis of the patient's condition. It is impossible to make any final diagnosis, however, without seeing the patient.
4. *Visit patient and perform physical exam*. The physician visits the patient in his or her room and, depending on the situation, may do one or more physical exams to try and

identify the patient’s medical condition. During this time, the physician also asks the patient and/or present family members for any additional information not found in the EMR, or to clarify/confirm the information, as it may not always be current.

5. *Gather information about patient.* Following the physical exam, the physician may once again choose to consult the EMR in light of the new knowledge obtained. For example, the patient may have mentioned some medical procedure in the past, so the physician will look up the relevant note for that procedure, if available.
6. *Revise hypothesis*. Following from Stages 4 and 5, the physician can now modify the original hypothesis and come up with a more accurate, albeit still tentative, diagnosis. In rare cases, the physician may reach a final conclusion at this point, but typically not before some action (Stage 7) has been performed.
7. *Execute course of action*. Based on the physician’s revised hypothesis, he or she will decide what laboratory tests need to be ordered, medications administered, or exams performed. While these tests/procedures are being carried out there is usually nothing else a physician can do for that patient but wait for the results. The physician can be notified of the results in several different ways. There is a notification in the EMR that indicates, next to each person’s name, when results are available to be viewed in the system. Very often, however, the nurses or other physicians will inform the attending physician of the availability of the results in the EMR. At that point, the physician is considered back at Stage 5 (gathering information), after which the hypothesis or suggested diagnosis is updated once more. Stages 5, 6, and 7 are repeated until the physician decides on a final diagnosis and the patient is either discharged or admitted to hospital.

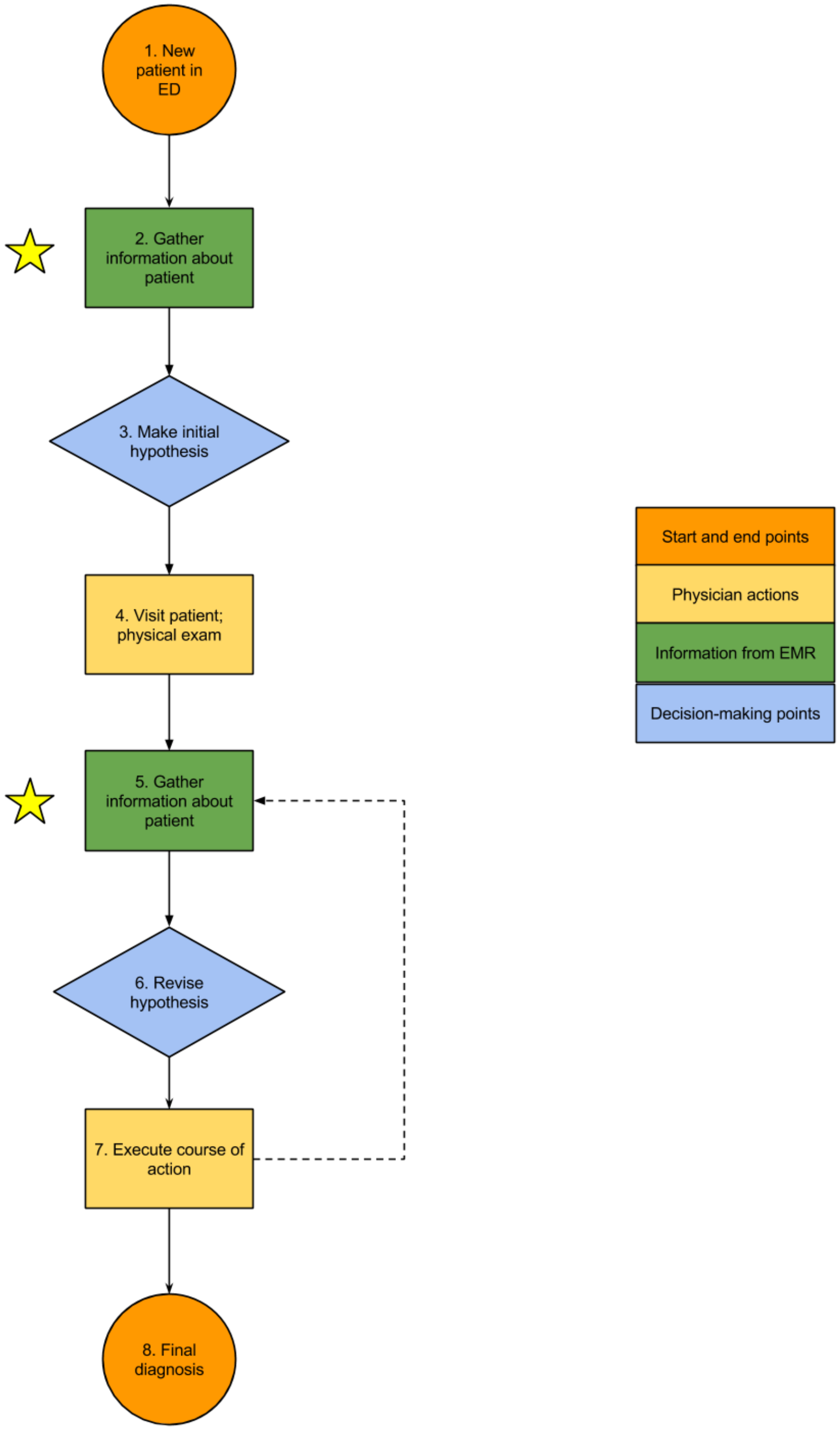

1. New patient in ED
2. Gather information about patient
3. Make initial hypothesis
4. Visit patient; physical exam
5. Gather information about patient
6. Revise hypothesis
7. Execute course of action
8. Final diagnosis
Start and end points
Physician actions
Information from EMR
Decision-making points

Figure 2.1: Outline of the workflow in the ED; the stages of interest for this research (marked with a star) are the stages that involve extracting information from the EMR.

Note that the process outlined above is rarely followed exactly. When highly-critical and time-sensitive patients present to the ED (e.g., suspected stroke, myocardial infarction, serious car accident resulting in trauma, etc.), the patient is rushed to a dedicated area known as the resuscitation bay, and an auditory alert will inform all physicians that a critical patient requires their attention. In such a case, physicians may choose to bypass Stage 2 and immediately see the patient (Stage 4). They can then try to gather information from family members and other sources while attending to the patient.

Based on the general workflow, the stages that are of most interest in this research are Stages 2-3 and Stages 4-5, the two sequences where physicians gather information from the ED in order to formulate or modify a diagnosis. Stages 4-5 differ from Stages 2-3 in that the physician now has additional information, both from the patient and from any laboratory/imaging results that may be available. EMR clutter may cause both of these information gathering procedures to take too much time, preventing physicians from spending more time with the patient or attending to other patients. Also, misses may occur at either stage; they may be particularly detrimental in Stage 4-5 as the physician approaches the final diagnosis.

In summary, the ED observations made it possible to understand the workflow in the ED and the role of the EMR in that context. Each of the EMR pages plays an important role at different points during a physician's assessment of a patient and several of these pages seem to suffer from clutter, both in terms of excess irrelevant data and poor organization (see Table 2.1). Particularly in the case of older patients with an extensive medical history, it became more difficult for physicians to locate what they needed. In one case, a physician using the Patient Timeline page to determine whether orthostatic measurements had been performed on a patient changed his mind three times before finally reaching a final answer. In another case, it was difficult to sift through all the text from previous medical notes when searching for the medicine a patient is taking. Physicians were also observed to scroll and click a lot, a source of additional wasted time when there was excess data. The Laboratory Results page was also often filled with results shown in red, which indicated abnormalities, meaning that it was hard to determine what information was really important. However, it was difficult to gauge from these observations what physicians thought of their EMRs, which was the focus of the next step in this research.

Table 2.1: EMR pages that seemed to be most affected by clutter

EMR page	**Potential clutter problems**	
	Excess data	***Poor organization***
Medical History	• Large volume of data in the case of older patients • Unnecessary code names for medical conditions are appended to each entry	• Data not organized in any logical fashion • Inconsistent font size • Inefficient use of space, with all entries in a long list; this leads to a lot of scrolling
Patient Timeline	• Large volume of data entered, including input from nurses, physicians, and automated systems • Much of the data is required for billing or meaningful use criteria but is not relevant for physicians	• Different font sizes • Inconsistent formatting of entries, with some in narrative form and others as short points
Laboratory Results	• Large quantity of results displayed in one page, potentially including values over several years	• Inconsistent formatting (e.g., some entries may be right-justified and others left-justified) • Red entries indicating problematic values are not always relevant and make it easier to miss important ones

Survey to Assess Physicians' Opinions on EMR Clutter

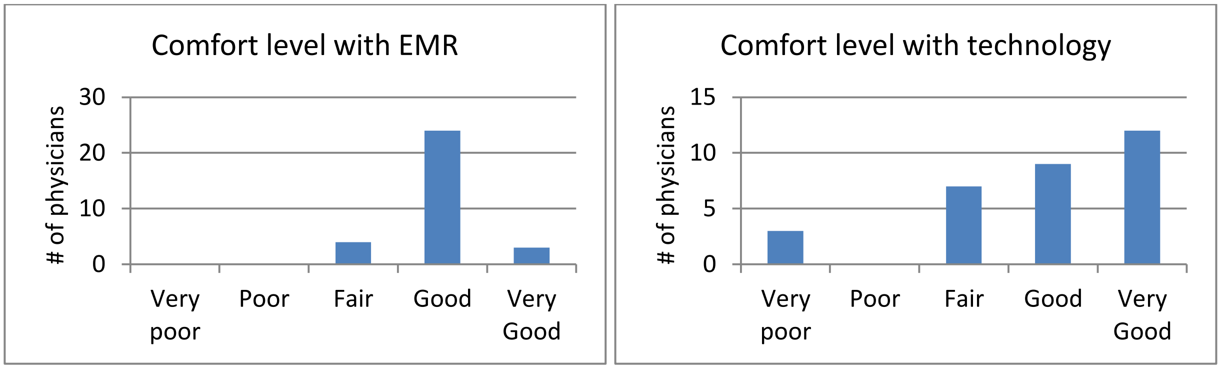
The goal of this next step was to complement the anecdotal evidence from the observations with subjective data on physicians' impressions of clutter. This survey was conducted in collaboration with Dr. Travis Ganje, a resident at the UM ED. The survey was classified as exempt research (Institutional Review Board (IRB) ID: #HUM00082309).

Methods. An anonymous survey was developed using Qualtrix. It was administered electronically to all 52 emergency medicine residents at the University of Michigan ED, with 31 residents responding. Of the residents that responded, 25 were between the age of 21 and 30, and 6 were between the age of 31 and 40. All of the responders were using the ED EMR at the time of the study.

The complete list of survey questions can be seen in Appendix A. First, residents answered questions about the types of EMRs they have used in the past. They then rated their comfort level with the EMR and technology in general on a 5-point Likert scale, with 1 being Very Poor and 5 being Very Good. On a similar 5-point scale, they provided ratings for how satisfied they were with their EMR (1: very dissatisfied; 5 very satisfied) and how important they believed the display of information was in general and for safety and efficiency (1: not important at all; 5: extremely important). They were also asked whether, based on their experience, they thought visual display design contributed to preventing or causing errors (1: strongly disagree; 5: strongly agree).

Next, a number of open-ended questions asked residents to indicate how they thought the display of information could be improved to benefit efficiency and safety. Finally, residents were presented with 26 screenshots from different pages in the EMR. For the first 15 images, they had to respond to the question “How much information is in this image?”, and for the next (different) 11 screenshots they had to respond to “How much clutter do you perceive?” In both cases, answers were provided using a slider from 0 to 100, with 100 referring to high information content or clutter.

Results. In general, residents indicated that they were comfortable with the EMR and with technology in general (see Figure 2.2). They were also largely satisfied with the EMR overall. Residents acknowledged the importance of good visual representation of data in general: 16 of 31 residents indicated that visual data representation was both “extremely important” for safety and for efficiency. The importance that physicians attach to the design of visual data was also reflected in their responses to the final two questions. When asked whether, in their experience, display design has helped avoid an error, 16 of 31 residents agreed or strongly agreed. Similarly, when asked whether they have experienced a case where design has led to an error, 14 of the 31 residents agreed or strongly agreed.



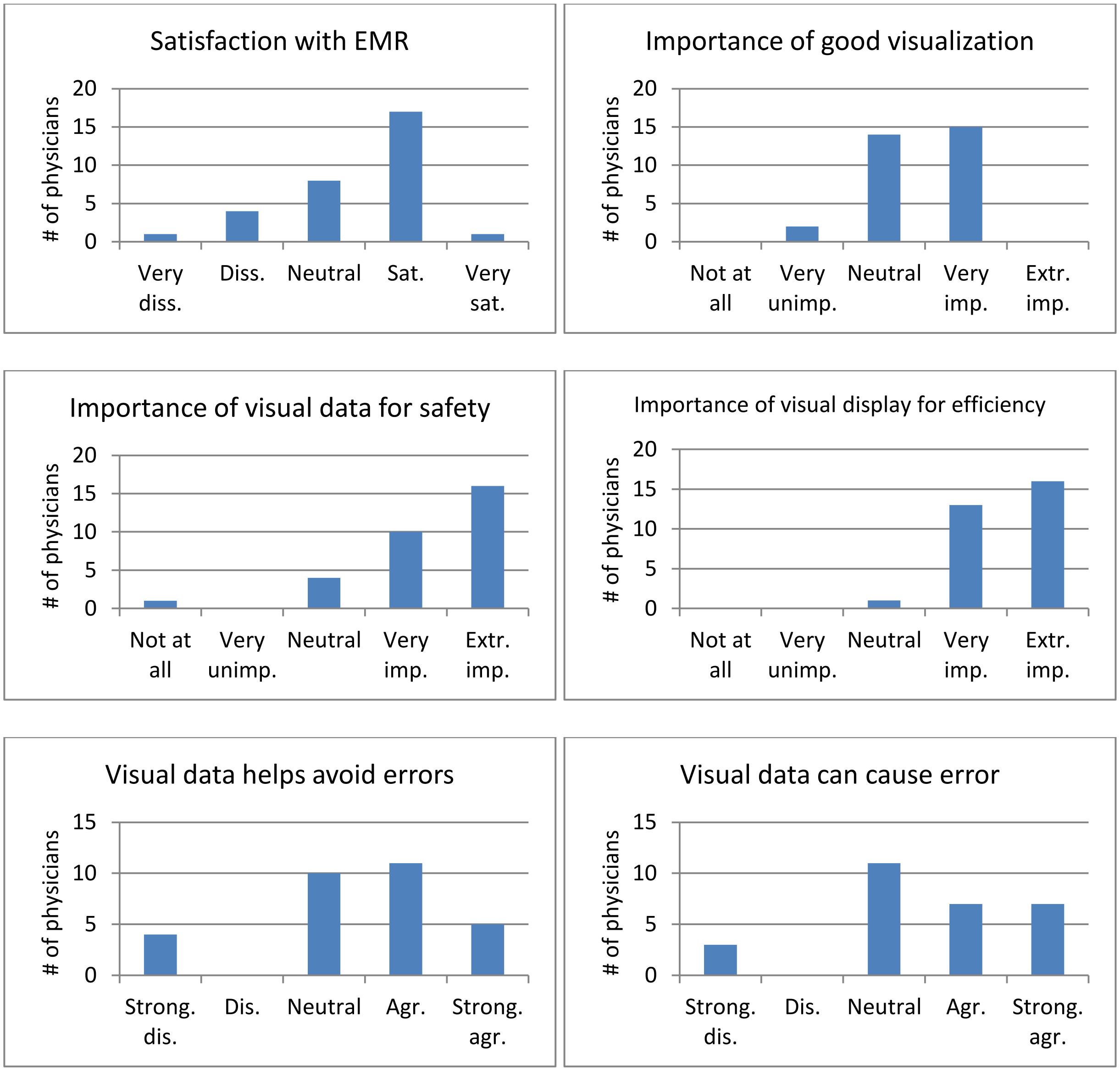


Figure 2.2: Results of the initial questions asked in the survey

The results of residents' rating of information and clutter on various EMR pages can be seen in Figure 2.3, as well as Tables 2.2 and Table 2.3. In order to assess the level of clutter on each page, the three clutter algorithms created by Rosenholtz et al. (2007), feature congestion (FC), subband entropy (SE), and edge density (EdD), were calculated (see Chapter 1). The results were then compared to residents' ratings of information content and clutter. For the ratings of amount of information, the Pearson's correlation was 0.78, 0.6, and 0.53 for FC, SE, and EdD, respectively (all $p < 0.05$). As for the ratings of clutter, the correlations were higher at 0.86, 0.87, and 0.73 for FC, SE, and EdD, respectively (all $p < 0.05$).

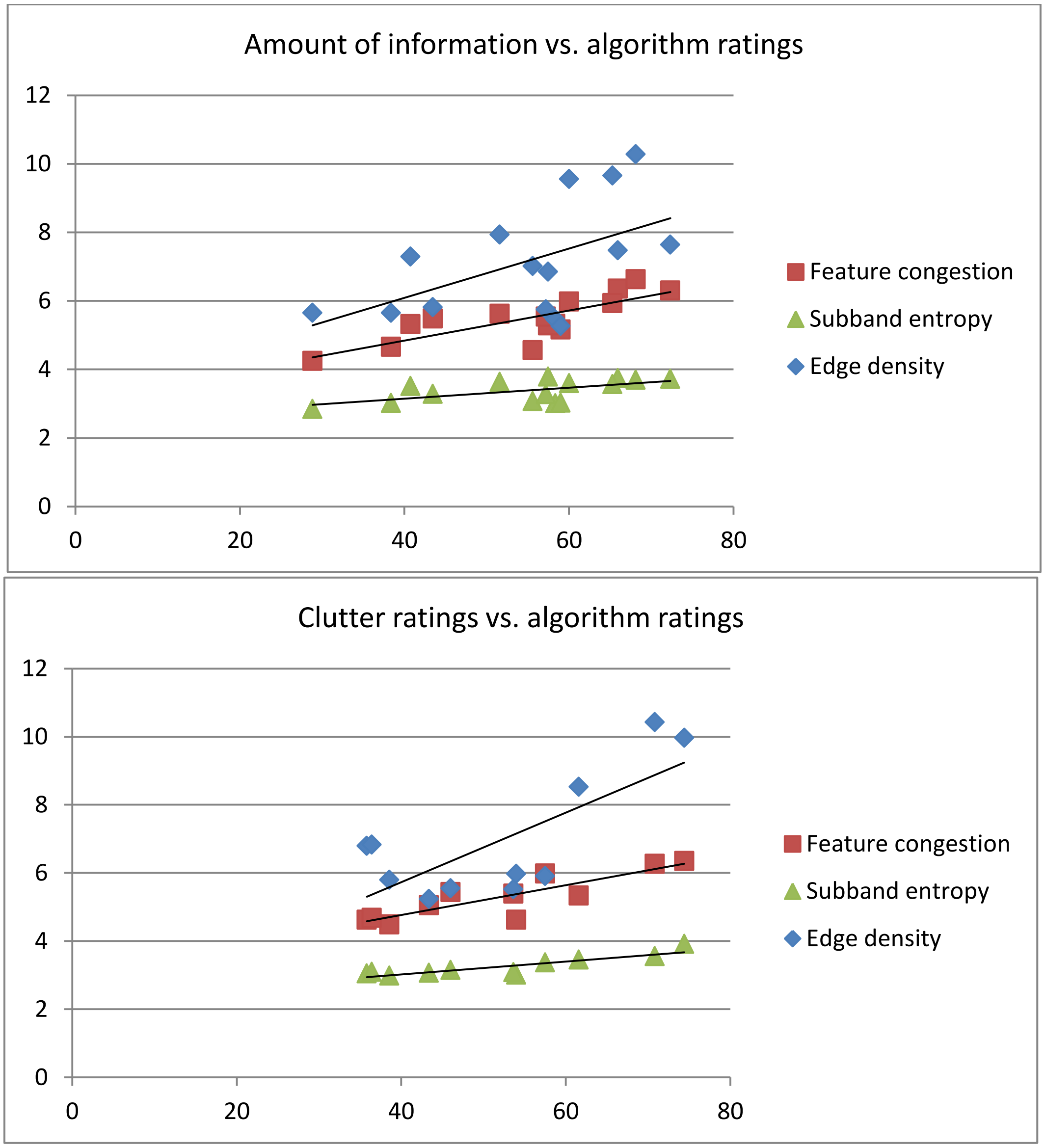


Figure 2.3: Ratings of information and clutter, plotted against the output from the three image processing algorithms of Rosenholtz et al. (2007): feature congestion (FC), subband entropy (SE), and edge density (EdD)

Table 2.2: Results for analysis of ratings of amount of information: coefficient of determination (R^2) and associated level of significance (*p*), as well as the Pearson correlation coefficient (*r*) and associated p value (*p*)

Ratings of amount of information				
	Regression values		**Correlation values**	
	***R*²**	***p***	***r***	***p***
FC	0.61	<0.001	0.78	<0.001
SE	0.36	0.0175	0.6	0.0175
EdD	0.28	0.04	0.53	0.04

Table 2.3: Results for analysis of ratings of clutter: coefficient of determination (R^2) and associated level of significance (*p*), as well as the Pearson correlation coefficient (*r*) and associated p value (*p*)

Ratings of clutter				
	Regression values		**Correlation values**	
	***R*²**	***p***	***r***	***P***
FC	0.74	<0.001	0.86	<0.001
SE	0.75	<0.001	0.87	<0.001
EdD	0.54	0.0097	0.73	0.0097

Finally, when asked about improvements to the EMR and about changes that would improve safety and efficiency, residents gave a wide range of responses. Four residents mentioned that information needs to be easier to access and find. Seven physicians mentioned that there is a need to reduce clutter or remove irrelevant information from the screen. Several of the residents highlighted the useless (i.e., not relevant to the current situation) information that is often displayed, such as self-populating physician notes, as examples of excess, irrelevant data. Furthermore, three residents mentioned the need to consolidate and organize data better such that relevant information is placed together.

Discussion. The residents who responded can be considered to be representative of ED physicians, if slightly more proficient with the EMR given their relatively young age and recent training. The survey served a number of purposes. First, it confirmed that physicians place a lot of importance on the design of visual information, even though their work is primarily based on physical exams and procedures. They believe display design has a significant effect on their work and can lead to errors if not done correctly. Second, the results suggested that physicians see a considerable need for improvements to the EMR in terms of the amount and organization of data, specifically removing irrelevant, automatically-generated information and grouping relevant information. Finally, the results from the ratings of amount of information and clutter were telling. The R^2 values regarding the amount of information were lower than the values for the clutter ratings (0.61, 0.36, and 0.21 for FC, SE, and EdD, respectively as compared to 0.75, 0.74, and 0.54). This suggests that it is not just the quantity of information that factors into physicians' perception of clutter but that other factors, such as the color variation and organization, play a role as well. This is confirmed by the fact that color variation and entropy were more significantly correlated with clutter ratings and had higher R^2 values than edge density which most closely represents the amount of data or objects on a screen. While subjective impressions of clutter are important, it is ultimately the performance decrements in information acquisition that will determine whether the EMR displays are problematic and require changes.

Physician Interviews: More Detailed Insight into the Effects of EMR Clutter

The final step in this familiarization and exploration stage of the dissertation research was to conduct interviews with physicians to identify their specific information needs (i.e., what exactly they need to know at different points in time) and determine whether clutter is a significant problem for them. The interviews also helped with the design of realistic ED patient scenarios in subsequent experiments on EMR clutter.

Methods. Three physicians who work in the UM ED were interviewed separately for around one hour each. They were asked to walk through different typical ED medical scenarios (see Table 2.4 for an example; the full list of questions can be seen in Appendix B). Some of the patient cases were based on material taken from the Harvard Medical School Simulation Casebook (Howard, Siegelman, Guterman, Hayden, & Gordon, 2011). Physicians read each case out loud and then had to answer the same two questions. The first question was what information they needed from the EMR at that point in time, and the second question was how they would use the EMR to obtain that information. All responses were recorded. By answering these two questions, physicians indicated where they struggled while using the EMR and, in particular, where clutter contributes significantly to their problems.

Table 2.4: Sample question used for the physician interviews (the full list can be seen in Appendix B)

Sample triage note	**Questions (same for all scenarios)**
A 75-year-old male presents to the ED feeling dizzy and with double vision. He is not speaking coherently and is being supported by his wife and son. His wife says he seems particularly weak on his right side.	1. What information do you need to know at this point to be able to proceed with your diagnosis? Please try to list these in (relative) order of importance. 2. How would you proceed with the EMR in order to obtain the information that you need? Please walk me through the steps you are likely to take.

In addition, physicians were asked a number of more open-ended questions related to clutter. Specifically, they were asked to comment on the amount of information in the EMR, how the amount of information affects how they extract information, and what suggestions they have for ways to reduce clutter.

Results. The interview sessions were recorded and then transcribed and summarized to identify common themes. The main problems with clutter that were identified repeatedly by different physicians were:

- *Excess irrelevant and/or redundant information*: much of the data in the EMR is not important to the physicians. Examples include what is known as pertinent negatives, where the EMR presents a list of common illnesses/problems that the patient does *not* currently have. This often causes confusion and delays and takes up unnecessary space. Another example are nursing notes, which often contain fixed information that nurses have to gather about patients, such as a pain rating. Several of these are yes/no answers or generic replies, and which are of no use to physicians. However, physicians are forced to scroll through them in order to get to more relevant information.
- *Poor salience of important information*: Critical information is not given due prominence. Measures such as a person's potassium level, blood pressure, heart rate, and non-negative troponin levels are highly time-sensitive and physicians need to be aware and alerted to this data as soon as possible. However, in the current EMR design, this information is not always easily visible unless physicians deliberately visit the relevant page or are alerted by a colleague.
- *Excess alerts*: the irony – and challenge to designers – is that the attempts to have important information stand out has largely led to an overload of visual alerts or "alert fatigue". The EMR currently highlights in red laboratory results that are outside

certain ranges. It is not uncommon to find laboratory results where almost all entries are in red. Physicians admitted to many times ignoring these values, as they so often are meaningless; whether a person's value is abnormal or not could depend on a lot of different factors apart from whether it falls within a certain range. However, the physicians acknowledged that these nuisance alarms occasionally lead to missing important information that would alter the course of action for a patient.

- *Highly fragmented information*: because of the way the EMR is structured and set up, many times information that is related is not presented in close proximity, a violation of the proximity-compatibility principle (Wickens & Carswell, 1995). Physicians frequently need to click through several different pages in order to complete a diagnosis of a patient, often having to search through many entries in each page, especially for people with a long and complicated medical history. In addition, entries are not always ordered or entered properly. For example, the medical history is listed in the order in which data are entered by a medical provider, meaning that they are neither in alphabetical nor chronological order. In some cases, it is also difficult to find certain particularly rare types of information. People on a narcotics contract are one typical case: in such situations, people are limited in their narcotics intake, and it will be indicated in the EMR that this patient is on such a contract. However, it is often difficult for the physician to find an actual copy of that contract, as there is no clear link to this information. Another challenge can be to aggregate information or quickly get a complete picture of a person's medical history. One pertinent example is the amount of radiation a person has received over the course of their life: this is critical information for physicians who want to prescribe radiation treatment.

Currently, acquiring this information this requires physicians to go through the EMR, navigate through pages related to surgical history, notes, etc., and manually estimate how much radiation the patient has received.

Discussion and Conclusion

This phase of the dissertation research served to understand physicians' information needs and ascertain to what extent clutter represents a problem for them. The observations in the ED helped understand the workflow and requirements of physicians at different stages of a patient's visit. This, in turn, proved extremely valuable in subsequent studies for developing realistic experiment scenarios. The observations also provided a preview of some of physicians' struggles with clutter. These struggles were then confirmed more formally by means of the survey deployed among ED residents. The survey showed that physicians are aware of the problem of clutter and that several factors, such as color variation, play a role in their perception of the phenomenon.

Finally, the interviews with ED physicians provided additional evidence of the struggles that physicians experience with clutter. The interview results confirmed that display clutter was not just a matter of aesthetic preference but a potential hazard to patient safety within the ED. Several physicians gave concrete examples of when they missed information among a large volume of poorly organized data. Overcoming the issue of clutter would thus greatly benefit safety and efficiency in the ED, and physicians seemed to readily welcome EMR design improvements aimed at reducing clutter. Discussing specific scenarios with the physicians also provided ideas for subsequent simulated patient scenarios.

References

Ahmed, A., Chandra, S., Herasevich, V., Gajic, O., & Pickering, B. W. (2011). The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. *Critical care medicine, 39*(7), 1626-1634.

Blumenthal, D. (2010). Launching HIteCH. *New England Journal of Medicine,362*(5), 382-385.

Blumenthal, D., & Glaser, J. P. (2007). Information technology comes to medicine. *The New England journal of medicine*, *356*(24), 2527.

Blumenthal, D., & Tavenner, M. (2010). The "meaningful use" regulation for electronic health records. *New England Journal of Medicine*, *363*(6), 501-504.

Bobillo, F., Delgado, M., & Gómez-Romero, J. (2008). Representation of context-dependant knowledge in ontologies: A model and an application. *Expert Systems with Applications*, *35*(4), 1899-1908.

Brazier, A., & Pacitti, B. (2008). Improving shift handover and maximising its value to the business. *Loss Prevention Bulletin*, *204*, 21.

Coiera, E. W., Jayasuriya, R. A., Hardy, J., Bannan, A., & Thorpe, M. E. (2002). Communication loads on clinical staff in the emergency department.*The Medical Journal of Australia*, *176*(9), 415-418.

Croskerry, P., Shapiro, M., Campbell, S., LeBlanc, C., Sinclair, D., Wren, P., & Marcoux, M. (2004). Profiles in patient safety: medication errors in the emergency department. *Academic emergency medicine*, *11*(3), 289-299.

Duftschmid, G., Rinner, C., Kohler, M., Huebner-Bloder, G., Saboor, S., & Ammenwerth, E. (2013). The EHR-ARCHE project: Satisfying clinical information needs in a Shared Electronic Health Record System based on IHE XDS and Archetypes. *International journal of medical informatics*, *82*(12), 1195-1207.

Eibling, D. (2009, August 29). Missing the point in the design of electronic medical records. Retrieved March 10, 2014, from http://maya.com/blog/missing-the-point-in-the-design-of-electronic-medical-records

Farley, H. L., Baumlin, K. M., Hamedani, A. G., Cheung, D. S., Edwards, M. R., Fuller, D. C., & Pines, J. M. (2013). Quality and Safety Implications of Emergency Department Information Systems. *Annals of Emergency Medicine*.

Fordyce, J., Blank, F. S., Pekow, P., Smithline, H. A., Ritter, G., Gehlbach, S., ... & Henneman, P. L. (2003). Errors in a busy emergency department. *Annals of emergency medicine*, *42*(3), 324-333.

Hammond, K. W., Efthimiadis, E. N., & Laundry, R. (2012, January). Efficient De-identification of Electronic Patient Records for User Cognitive Testing. In*HICSS* (pp. 2771-2778).

Hoffman, S., & Podgurski, A. (2011). Meaningful use and certification of health information technology: what about safety?. *The Journal of Law, Medicine & Ethics*, *39*(s1), 77-80.

Howard Z, Siegelman J, Guterman E, Hayden EM, Gordon JA (2011). *Simulation Casebook.* The Gilbert Program in Medical Simulation, Harvard Medical School Casebook. Retrieved from http://mycourses.med.harvard.edu/ResUps/GILBERT/pdfs/HMS_7607.pdf.

Institute of Medicine (US) Committee on Patient Safety and Health Information Technology. (2012). *Health IT and patient safety: building safer systems for better care*. National Academies Press.

James, J. T. (2013). A new, evidence-based estimate of patient harms associated with hospital care. *Journal of patient safety*, *9*(3), 122-128.

Karsh, B. T., Weinger, M. B., Abbott, P. A., & Wears, R. L. (2010). Health information technology: fallacies and sober realities. *Journal of the American medical informatics Association*, *17*(6), 617-623.

Kachalia, A., Gandhi, T. K., Puopolo, A. L., Yoon, C., Thomas, E. J., Griffey, R., & Studdert, D. M. (2007). Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. *Annals of emergency medicine*, *49*(2), 196-205.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (1999). To err is human: Building a safer health system. Committee on Health Care in America. Institute of Medicine.

Kuhn, G. (2001). Circadian rhythm, shift work, and emergency medicine.*Annals of emergency medicine*, *37*(1), 88-98.

Levin, S., France, D. J., Hemphill, R., Jones, I., Chen, K. Y., Rickard, D., ... & Aronsky, D. (2006). Tracking workload in the emergency department. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 526-539.

Ludwick, D. A., & Doucette, J. (2009). Primary care physicians' experience with electronic medical records: barriers to implementation in a fee-for-service environment. *International journal of telemedicine and applications*, *2009*, 2.

Maya (2009). Missing the point in the design of medical records. Retrieved from http://maya.com/blog/missing-the-point-in-the-design-of-electronic-medical-records

Murphy, D. R., Reis, B., Sittig, D. F., & Singh, H. (2012). Notifications received by primary care practitioners in electronic health records: a taxonomy and time analysis. *The American journal of medicine*, *125*(2), 209-e1.

Pennsylvania Patient Safety Authority 2012 Annual Report (April 30, 2013). Retrieved from http://patientsafetyauthority.org/PatientSafetyAuthority/Documents/Annual%20Report%202012%20.pdf

Singh, H., Spitzmueller, C., Petersen, N. J., Sawhney, M. K., & Sittig, D. F. (2013). Information overload and missed test results in electronic health record–based settings. *JAMA internal medicine*, *173*(8), 702-704.

Trzeciak, S., & Rivers, E. P. (2003). Emergency department overcrowding in the United States: an emerging threat to patient safety and public health.*Emergency medicine journal*, *20*(5), 402-405.

Tully, M. P., Kettis, Å., Höglund, A. T., Mörlin, C., Schwan, Å., & Ljungberg, C. (2013). Transfer of data or re-creation of knowledge–Experiences of a shared electronic patient medical records system. *Research in Social and Administrative Pharmacy*, *9*(6), 965-974.

University of Michigan Health System (2015). Hospital Profiles. Retrieved from http://medicine.umich.edu/dept/emergency-medicine/education-residency/hospital-profiles

US Department of Health and Human Services: Doctors and hospitals' use of health IT more than doubles since 2012. (2013, May 22). Retrieved from http://www.hhs.gov/news/press/2013pres/05/20130522a.html

Van Vleck, T. T., Stein, D. M., Stetson, P. D., & Johnson, S. B. (2007). Assessing data relevance for automated generation of a clinical summary. In*AMIA Annual Symposium Proceedings* (Vol. 2007, p. 761). American Medical Informatics Association.

Weir, C. R., & Nebeker, J. R. (2007). Critical issues in an electronic documentation system. In *AMIA Annual Symposium Proceedings* (Vol. 2007, p. 786). American Medical Informatics Association.

Zeldes, N., & Baum, N. (2011). Information overload in medical practice. *J Med Pract Manage*, *26*(5), 314-6.

Zeng, Q., Cimino, J. J., & Zou, K. H. (2002). Providing Concept-oriented Views for Clinical Data Using a Knowledge-based System An Evaluation. *Journal of the American Medical Informatics Association*, *9*(3), 294-305.

Zhang, R., Pakhomov, S., McInnes, B. T., & Melton, G. B. (2011). Evaluating measures of redundancy in clinical texts. In *AMIA Annual Symposium Proceedings* (Vol. 2011, p. 1612). American Medical Informatics Association.

CHAPTER 3

Analyzing the Effects of Clutter: Selecting Eye Tracking Metrics to Capture Observed Performance Decrements

Prior to developing real-time solutions for detecting and overcoming clutter, it was necessary to determine first 1) which eye movement metrics are most sensitive to the effects of clutter on attention during visual search and noticing, and 2) whether these eye movement metrics can help explain any related performance decrements. More specifically, the goal was to establish whether performance decrements can be attributed to increased dispersion of attention (as evidenced by *spread* metrics), inefficient search (*directness* metrics), and/or increased time to process particular areas of the display (*duration* metrics). As detailed earlier, the proposed metrics in each of these groups are all display-independent, meaning that they do not depend on the location of the target or on areas of interest (AOIs) on the display. This implies that, in contrast to more commonly used metrics, they can be applied across displays and domains.

Two experiments were performed where clutter was the main independent variable. In particular, the focus was on data density, the most prominent aspect of clutter in the adopted definition; the organization aspect of clutter was examined more carefully and systematically in subsequent studies. In the two experiments described in this chapter, participants performed one or both of two tasks, visual search and noticing. As described in Chapter 1, visual search refers to the process of actively looking for a given target (Treisman, 1991), whereas noticing involves a bottom-up process related to registering critical events that were not defined in advance (Schmidt, 1995; p. 29). The dependent measures in all cases were performance measures (response time and miss rate), eye tracking, and subjective data (namely, ratings of clutter).

The first, largely exploratory, experiment was conducted using simple illustrations and student participants performing a search task. Then, the second experiment, which will be examined in more detail, involved medical personnel conducting both search and noticing tasks in a simulation of the emergency department (ED).

Experiment 1: Effects of Data Density on Visual Search in a Simple, Highly-Controlled Environment

In this first experiment, the objective was to explore whether eye tracking is, at all, a promising means of tracing the effects of clutter during visual search. Low and high-clutter illustrations from the *Where's Waldo* book series by Handford (1997a, 1997b) were used, all of which contained the same target (Waldo) that participants had to search for. There is precedence for using this series of images for visual search experiments (e.g., Andrews & Coppola, 1999; Klein & McInnes, 1999; Otero-Milan, Troncoso, Macknik, Serrano-Pedraza, & Martinez-Conde, 2008). In this case, the illustrations were chosen because a) they made for a task that student participants were familiar with and could easily perform without excessive training and b) they reduced the possibility that a factor such as experience might affect performance on the search tasks. The goal was to ensure – to the extent possible – that only the deliberate manipulation of clutter was causing changes in eye movements and attention allocation.

Methods

Participants. The participants in this study were 33 engineering students from the University of Michigan (UM). Their average age was 21.9 years (*SD* = 4.41). Participants had self-reported normal or corrected to normal vision; contact lenses were allowed but glasses were not, due to the limitations of the eye tracker. The eye tracking data of 11 participants was rejected because of eye discrimination and calibration issues, resulting in a participant count of 22 (average age: 20.8 years, *SD* = 3.01) for the eye tracking results. All participants gave informed consent and received $15 upon completion of the experiment. This study was approved by the UM Health Sciences and Behavioral Sciences Institutional Review Board (HSBS-IRB; ID: #HUM00051160).

Visual display. As mentioned previously, the experimental stimuli in this study were digital versions of images from the *Where's Waldo* book series (Handford, 1997a, 1997b). The target, Waldo, was present in each image, and subtended around 4 degrees vertically and 1 degree horizontally. The timing of image presentation and the recording of all keyboard and mouse input were controlled by the Matlab Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

Experiment setup. The images were displayed on the full screen of a 19-inch monitor with a resolution of 1440x900. The viewing distance was approximately 50 cm and the lighting conditions were the same for all participants. Eye movements were recorded using an ASL Eye-Trac D6 infrared-based, desktop-mounted eye tracker (sampling rate: 60 Hz; accuracy: less than 1 degree visual angle; precision: 0.5 degrees visual angle). The eye tracker was placed directly in front of the computer monitor such that the front of the eye tracker was around 5 cm from the screen. The participants were seated at a distance of around 50 cm from the screen, making for a visual angle of around 25 degrees in the horizontal direction and 20 degrees in the vertical directions. Calibration took place at the start of the experiment using a 9-point grid. The duration of the calibration procedure varied across participants but generally took around 5 minutes.

Experiment design. A total of 12 different images were used. Of these images, 6 were classified as high in clutter and 6 were classified as low in clutter, based on the density of depicted people and objects in each image. This classification was based on the perception of the experimenters and was not quantified; rather it was validated using subjective feedback following the experiment. Participants were divided into two groups, Group 1 and Group 2, both of which saw the same set of 6 low-clutter and 6 high-clutter images. The 12 images were presented in a random but fixed random order to both groups, with the order of Group 2 the reverse of Group 1.

The dependent measures were the search time to find the target (Waldo), the number of misses (where a miss is defined as an incorrect answer or a failure to answer within the given time limit, as explained later), and the eye movement metrics. A debriefing questionnaire was also administered at the end of the experiment. As part of this questionnaire, participants had to rate the amount of clutter they perceived (scale: 0-10, with 10 being the highest perceived clutter level) in sample images from the experiment.

Experiment procedure. After signing the consent form and being informed about the purpose of the experiment, participants were shown an image of the target (Waldo) and allowed to look at it until they felt they could remember it well enough to search for it. Next, participants were given a practice search task in which they had to search for Waldo in a given display. They needed to successfully find Waldo in this practice task, after which the experiment could begin.

Participants were instructed to press the “up” arrow on the keyboard as soon as they located Waldo, at which point response time was recorded. Also, participants had to use the mouse to click on the assumed location of Waldo, which was used to calculate error rate. Participants were given a maximum search time of 120 seconds for each image before the trial timed out. A white screen with crosshairs in the middle was presented for five seconds between two pictures, and participants were instructed to look at the crosshairs for the duration of their appearance. Participants were instructed to perform the task as fast and accurately as possible. After completing the experiment, participants filled out the debriefing questionnaire.

Results

Subjective clutter ratings. Participants rated the amount of clutter in a sample of experiment images using a scale from 0 (least cluttered) to 10 (most cluttered). A Mann-Whitney U test revealed that the subjective ratings of the images were significantly affected by clutter ($U = 3149$, $p < 0.01$), with a mean rating of 3.8 (*SD* = 1.91) for low-clutter images and 7.6 (*SD* = 1.71) for high-clutter images.

Search time and miss rate. A repeated-measures analysis of variance (ANOVA) with listwise deletion was used to analyze the effects of clutter on performance. Mean search time for the target significantly increased for cluttered displays ($F(1, 32) = 18.7$, $p < 0.001$; $\eta_p^2 = 0.37$), from 18.5 (*SD* = 10.1) seconds in low-clutter images to 41.8 (*SD* = 29.9) seconds in high-clutter images. Similarly, the mean miss rate significantly increased, with a mean of 0.02 errors (*SD* = 0.054) per person in the low-clutter case, compared to a mean of 0.62 errors (*SD* = 0.20) per person in high-clutter images ($F(1, 32) = 286$, $p < 0.001$, $\eta_p^2 = 0.9$).

Eye tracking results. The effects of clutter on the eye tracking metrics were also analyzed using a repeated-measures ANOVA (see Table 3.1 and Figure 3.1). Convex hull area increased from 118,524 (*SD* = 30,000) pixels^2 to 165,611 (*SD* = 25,000) while spatial density increased from 0.44 (*SD* = 0.12) to 0.71 (*SD* = 0.10). The mean saccade amplitude decreased from 89.1 (*SD* = 13.2) pixels in low clutter to 77.7 (*SD* = 9.3) in high clutter and the rate of transitions increased from 0.16 (*SD* = 0.068) in low clutter to 0.30 (*SD* = 0.058) in high clutter. Finally, mean fixation duration significantly increased from 0.56 (*SD* = 0.12) seconds in low clutter to 0.61 (*SD* = 0.10) seconds in high clutter.

Table 3.1: Eye tracking results. Note: NNI (nearest neighbor index) was not calculated here

Eye tracking metric	**Low clutter mean (*SD*)**	**High clutter mean (*SD*)**	**Main effects of clutter**
		Spread metrics	
Convex hull area (pixels^2)	118,524 (30,000)	165,611 (25,000)	*F*(1, 21) = 85.4, *p* < 0.001
Spatial density*	0.44 (0.12)	0.71 (0.10)	*F*(1, 21) = 128.9, *p* < 0.001
		Directness metrics	
Scanpath length per second (pixels/sec)	257 (309)	608 (758)	Not significant
Mean saccade amplitude (pixels)	89.1 (13.2)	77.7 (9.3)	*F*(1, 21) = 43.9, *p* < 0.001
Backtrack rate (/sec)	0.42 (0.079)	0.42 (0.060)	Not significant
Rate of transitions* (/sec)	0.16 (0.068)	0.30 (0.058)	*F*(1, 21) = 151.4, *p* < 0.001)
		Duration metrics	
Mean fixation duration (sec)	0.56 (0.12)	0.61 (0.10)	*F*(1, 21) = 5.8, *p* = 0.025

*The display was divided into 117 cells (9x13) of approximately 50x50 pixels each

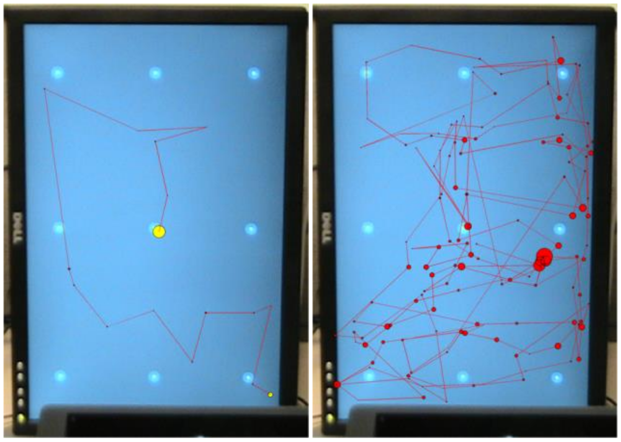


Figure 3.1: Sample fixation patterns for one participant corresponding to low clutter (left) and high clutter (right) images. Note the relatively higher spatial density (i.e., more areas of the display containing at least one fixation point), higher fixation duration (larger diameter of fixation circles), and shorter saccade amplitude in the case of high clutter

Experiment 1 Discussion

The significant increase in subjective ratings for the high-clutter images confirms that the manipulation of clutter was successful. As expected, high clutter led to longer response times and higher error rates. From the eye tracking results, it appears that a number of spread, directness, and duration eye movement metrics can be used to trace and explain the performance effects of clutter. The spread metrics (namely, spatial density and convex hull area) indicate a larger dispersion of fixations in the case of high clutter, suggesting increased uncertainty about target location and a lack of bottom-up capture. The lower mean saccade amplitude, a directness metric, implies that fixations were more closely spaced together, which may signify a more careful and systematic search for the target. At the same time, however, the higher rate of transitions reveals that participants moved back and forth more often between larger areas of the screen, pointing to a less efficient search (Goldberg & Kotval, 1999). Finally, the increased mean fixation duration indicates that it took participants more time to extract information from items on highly-cluttered images. Together, these effects explain the longer response time observed in case of high clutter.

This study does have limitations in that the amount of data was not systematically manipulated. Also, the amount of occlusion of the target (Waldo) was not the same for the different displays. These limitations were overcome in Experiment 2, which built on the initial, promising findings from Experiment 1.

Experiment 2: Effects of Data Density and Stress on Visual Search and Noticing in ED EMRs

The next step was to test whether all or some of the above eye movement metrics would prove useful and diagnostic of clutter in a more complex environment. Moreover, the question was whether they could reflect interactions of clutter with user-based factors, such as workload or stress, which are common in such complex environments. The main goals of this second experiment were to 1) analyze the effects of clutter on visual search and noticing performance and determine which eye movement metrics can be used to detect the effects of clutter on information acquisition, and 2) to examine to what extent stress – a major performance-shaping factor in domains like the ED – and task difficulty interact with clutter and exacerbate its effects. The application domain in this case was Electronic Medical Records (EMRs) in the ED.

Methods

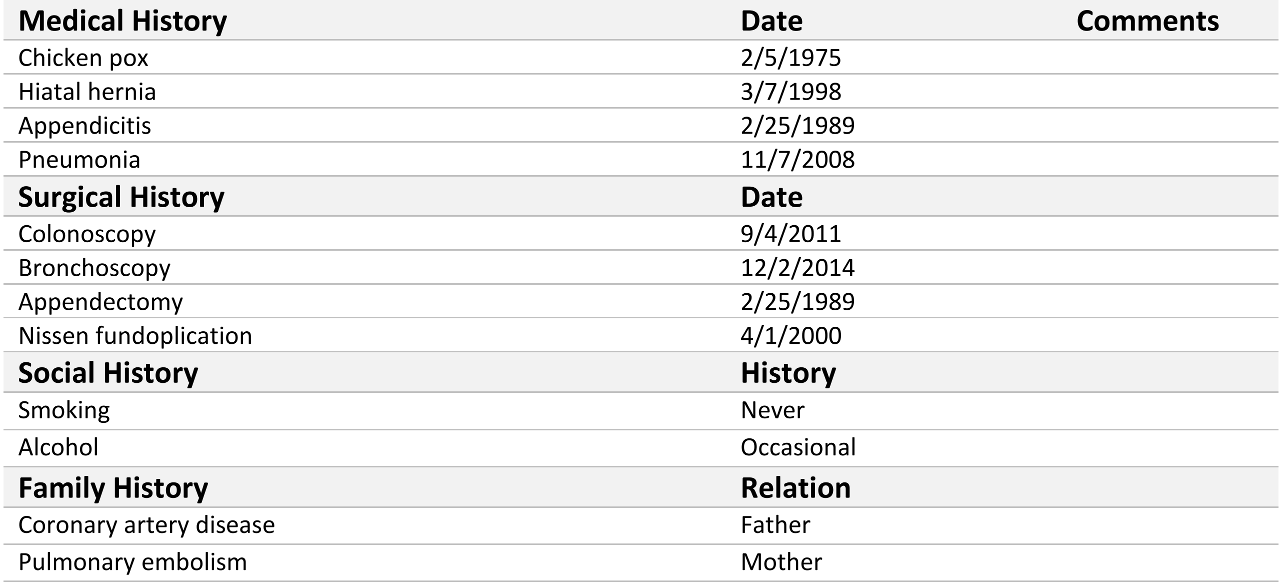
Participants. The participants in this experiment were 15 medical practitioners (seven female and eight male). They included 11 residents, one staff physician, and three physician assistants who were employed in the University of Michigan ED. Their average age was 30.1 years (*SD* = 3.3), and their average years of experience in this ED was 4.3 years (*SD* = 4.0). Their average experience using the EMR was 2.0 years (*SD* = 0.75). Participants had self-reported normal or corrected to normal vision; contact lenses were allowed but glasses were not, due to the limitations of the eye tracker. All participants gave informed consent and received $125 for their time. This study was approved by the UM HSBS-IRB (ID: #HUM00078246).

Simulated EMR. Three pages of the EMR that is currently being used by all research participants were replicated. Following the discussions and observations with ED physicians, these three pages were were identified as being particularly cluttered due to poor organization and high data density. The pages were created using Windows Presentation Foundation (WPF) with XAML and C# and they were populated with made-up but realistic patient data:

1. *Medical History page* (see Figure 3.2; copyright restrictions mean that the actual page cannot be shown): this page displays the patient's medical, surgical, social, and family history. One important problem with this page is that the information is not presented in any logical order within each group; rather, it is listed in the order in which different medical providers entered that data.
2. *Laboratory Results page*: results from current and prior laboratory tests are displayed in a table, with abnormal results indicated in red. This page tends to include a lot of distracting information in the form of red laboratory results that are not abnormal for this patient.

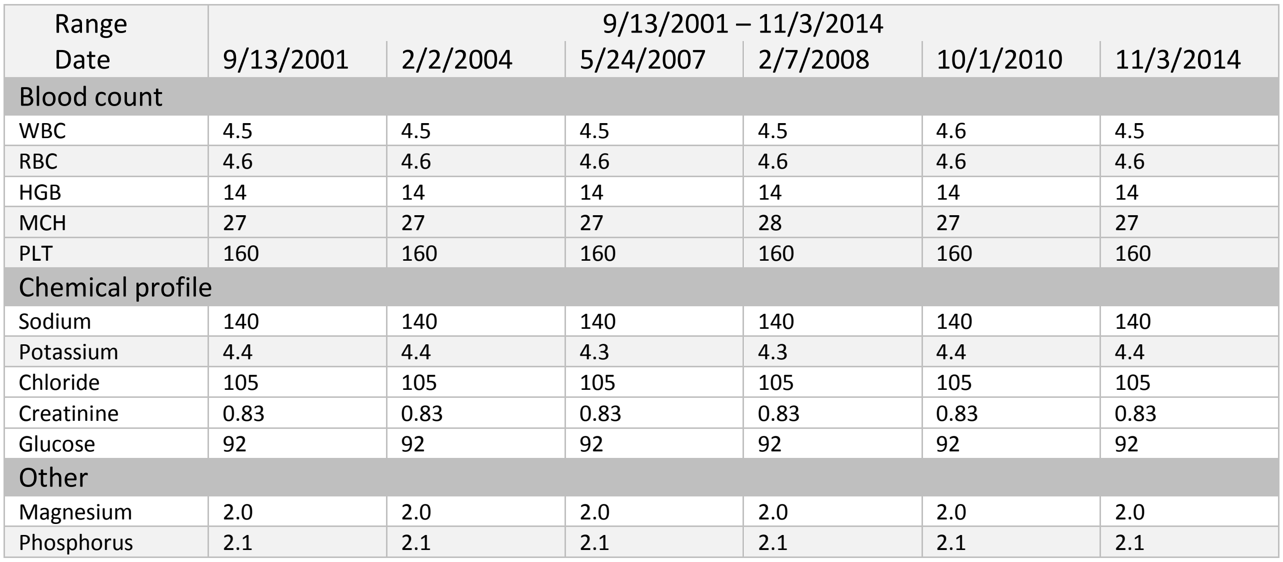
3. *Patient Timeline page*: lists the events that occurred since the patient presented to the ED. Each segment of the timeline is a text block. One issue is that these blocks are closely spaced and have inconsistent formats.

In this experiment, all pages were static and non-scrollable. All the information needed for the experiment was present on the respective page.



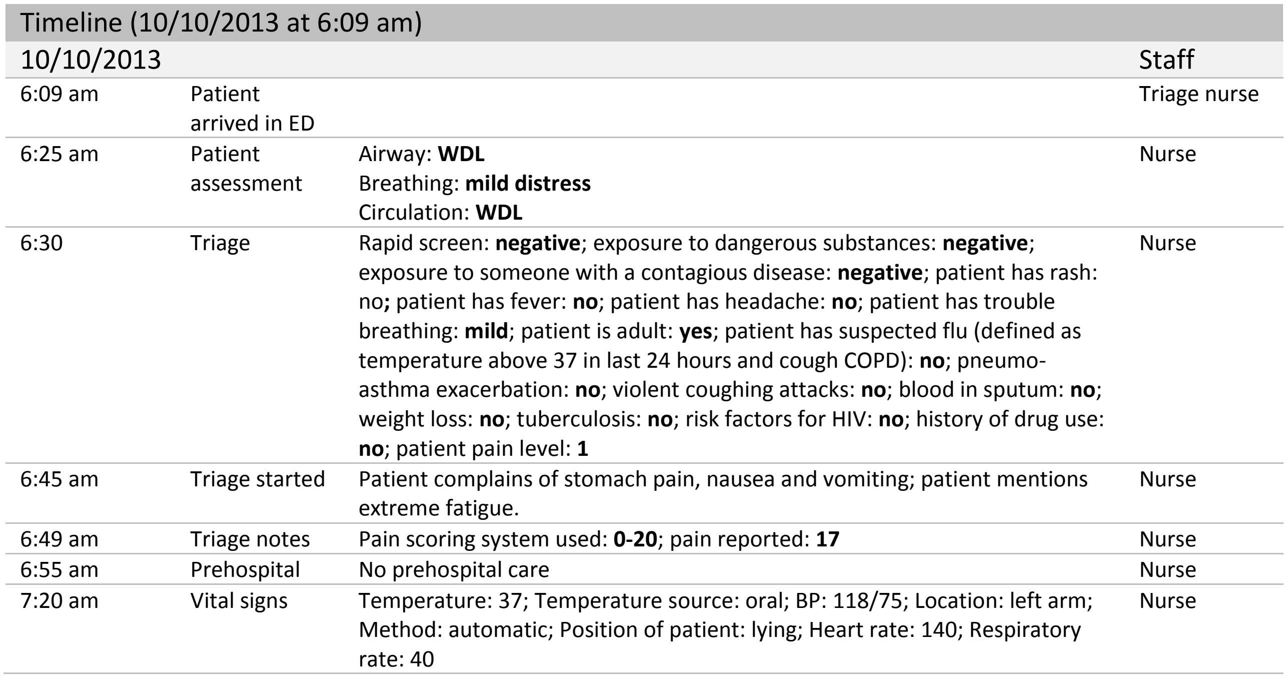
Medical History	Date	Comments
Chicken pox	2/5/1975	
Hiatal hernia	3/7/1998	
Appendicitis	2/25/1989	
Pneumonia	11/7/2008	
Surgical History	**Date**	
Colonoscopy	9/4/2011	
Bronchoscopy	12/2/2014	
Appendectomy	2/25/1989	
Nissen fundoplication	4/1/2000	
Social History	**History**	
Smoking	Never	
Alcohol	Occasional	
Family History	**Relation**	
Coronary artery disease	Father	
Pulmonary embolism	Mother	

(a)



Range	9/13/2001 – 11/3/2014					
Date	9/13/2001	2/2/2004	5/24/2007	2/7/2008	10/1/2010	11/3/2014
Blood count						
WBC	4.5	4.5	4.5	4.5	4.6	4.5
RBC	4.6	4.6	4.6	4.6	4.6	4.6
HGB	14	14	14	14	14	14
MCH	27	27	27	28	27	27
PLT	160	160	160	160	160	160
Chemical profile						
Sodium	140	140	140	140	140	140
Potassium	4.4	4.4	4.3	4.3	4.4	4.4
Chloride	105	105	105	105	105	105
Creatinine	0.83	0.83	0.83	0.83	0.83	0.83
Glucose	92	92	92	92	92	92
Other						
Magnesium	2.0	2.0	2.0	2.0	2.0	2.0
Phosphorus	2.1	2.1	2.1	2.1	2.1	2.1

(b)



Timeline (10/10/2013 at 6:09 am)			
10/10/2013			Staff
6:09 am	Patient arrived in ED		Triage nurse
6:25 am	Patient assessment	Airway: **WDL** Breathing: **mild distress** Circulation: **WDL**	Nurse
6:30	Triage	Rapid screen: **negative**; exposure to dangerous substances: **negative**; exposure to someone with a contagious disease: **negative**; patient has rash: no**;** patient has fever: **no**; patient has headache: **no**; patient has trouble breathing: **mild**; patient is adult: **yes**; patient has suspected flu (defined as temperature above 37 in last 24 hours and cough COPD): **no**; pneumo-asthma exacerbation: **no**; violent coughing attacks: **no**; blood in sputum: **no**; weight loss: **no**; tuberculosis: **no**; risk factors for HIV: **no**; history of drug use: **no**; patient pain level: **1**	Nurse
6:45 am	Triage started	Patient complains of stomach pain, nausea and vomiting; patient mentions extreme fatigue.	Nurse
6:49 am	Triage notes	Pain scoring system used: **0-20**; pain reported: **17**	Nurse
6:55 am	Prehospital	No prehospital care	Nurse
7:20 am	Vital signs	Temperature: 37; Temperature source: oral; BP: 118/75; Location: left arm; Method: automatic; Position of patient: lying; Heart rate: 140; Respiratory rate: 40	Nurse

(c)

Figure 3.2: Outlines of the three pages used in the experiment: (a) Medical History, (b) Laboratory Results, and (c) Patient Timeline. Note that these figures are approximate guidelines only and do not represent the exact text, font size, or dimensions of the displays used. Also, the additional patient and navigation information on these pages is not shown here.

Experiment setup. The EMR pages were displayed on a 19-inch monitor with a resolution of 1280x1024 pixels. The same ASL D-6 desktop-mounted eye tracker was used as in the previous experiments (sampling rate: 60 Hz; accuracy: less than 1 degree visual angle; precision: 0.5 degrees visual angle). The eye tracker was placed directly in front of the computer monitor such that the front of the eye tracker was around 5 cm from the screen. The participants were seated at a distance of around 50 cm from the screen, making for a visual angle of around 25 degrees in the horizontal direction and 20 degrees in the vertical directions. The text on the screen subtended 0.6 degrees visual angle, and the lighting was held constant for all participants. Calibration took place at the start of the experiment using a 9-point grid. The duration of the calibration procedure varied across participants but generally took around 5 minutes.

Experiment design. The three independent variables were clutter (low, high), stress (no stress, stress), and type of search task (simple, difficult). These were varied within participants, leading to a 2x2x2 full-factorial design. The combination of these three variables translated to 24 versions of the EMR pages (eight per page), each containing data for a different fictitious patient scenario. For each patient scenario, there was a corresponding and unique (1) summary of the patient's symptoms, similar to what a physician would receive in a triage note (2) search target; i.e., information that a physician might look for in a similar situation and that the participant was asked to locate, (3) noticing target (different from the search target); i.e., information that a physician needed to observe on their own to make the correct diagnosis, and (4) suggested diagnosis, which was presented to participants as the initial assessment by a medical student. Each patient scenario corresponded to one experiment trial, and participants were exposed to the same one trial per condition.

Participants were told to first answer the search question (i.e., find the search target) and then use the rest of their time to assess the likelihood of the diagnosis based on the available EMR data (noticing task). Note that they were not requested to make a *final* diagnosis about the patient, as such a task would entail examining more than just one or few EMR pages. Participants provided all answers verbally.

Collaborating physicians in the ED confirmed that the patient scenarios were realistic and mimicked some of the tasks they perform using their EMR. Table 3.2 shows 4 sample questions for the Medical History page, including the text that was presented to physicians and the search and noticing targets. The full scenario details for all three pages can be found in Appendix C.

The three experimental variables in this study were manipulated as follows:

- *Clutter*: clutter was operationally defined as the density of entries in the EMR displays. High clutter was created by adding information that was irrelevant to the current condition of the patient. For example, for a patient presenting with abdominal pain, irrelevant information could include entries about skin conditions or ear disorders. The low versus high clutter tasks were equivalent in that 1) there was the same amount of *relevant* information (i.e., the same number of relevant entries), 2) the eccentricity of the search and noticing targets were held to within less than a fifth of a page of difference (always in the same horizontal level and never more than 300 pixels apart vertically in a 1680x1050 pixel image), and 3) the two questions were comparable (e.g., they both required search for an entry in the medical history section). The font size for all pages was 12. For the medical history page, there was an average of 20.0 (SD = 4.4) entries or 67.2 (SD = 9.2) words in the low-clutter condition and 44.7 (SD = 3.0) entries in the high-clutter condition. For the laboratory results page, the low-clutter condition had an average of 31 (SD = 2.3) entries in the low-clutter condition and an average of 170 (SD = 11.6) in the high-clutter conditions. Finally, for the Patient Timeline page, there was an average of 11.7 (SD = 1.08) entries or 247 (SD = 38) words in the low-clutter condition and 16.5 (SD = 1.12) entries or 580 (SD = 33.1) words in the high-clutter condition.
- *Stress*: stress was induced using a combination of techniques. First, the high-stress scenarios involved life-threatening patient conditions, as opposed to less serious conditions for low-stress tasks. Second, a time limit of 20 seconds to perform the search and noticing tasks was imposed during the high-stress scenarios. This limit was determined empirically, following pilot tests with physicians. Third, participants were told that the person with the best performance in these high-stress tasks (in terms of

search time and accuracy) would be paid an extra $50, in addition to the regular compensation for participating in the experiment. Fourth, participants were instructed to complete the task as fast as possible yet still accurately, whereas in the low-stress condition, participants were asked to be both fast and accurate. Finally, the high-stress scenarios were accompanied by ventilator and heart rate monitor alerts, sounds that physicians would typically associate with critical conditions.

- *Task*: both simple and difficult search tasks were used. Simple search tasks involved finding only one given piece of information in the display (e.g., "Does the patient have a history of gastritis?"), whereas difficult tasks involved finding several instances of a given item and performing some comparison or judgment (e.g., "When was the patient's most recent aortic root repair?").

Table 3.2: Example of 4 scenarios (2 low/high sets) for the Medical History page. The first pair is a simple task, low-stress example, while the second is a difficult task, high-stress example.

Diagnosis	Clutter	Given text (high stress scenarios in *italics*; search target shown in **bold** here)	Noticing target
		Simple task	
Peptic ulcer	Low clutter	A 45-year-old female has severe abdominal pain, and nausea. Your medical student has evaluated the patient and believes she has cholecystitis. You open the patient's chart to check her medical history. Does the patient have a history of **gastric reflux**?	Cholecystectomy
	High clutter	An 80-year-old female has severe abdominal pain and nausea. Your medical student has evaluated the patient and believes she has appendicitis. You open the patient's chart to check her medical history. Does the patient have a history of **gastritis**?	Appendectomy
		Difficult task	
Aortic dissection	Low clutter	*A 60-year-old male presents with acute onset of severe chest pain radiating to the back. Your medical student believes he has aortic dissection. You open the patient's chart to check for a history of aortic issues. What year was his most recent aortic **bypass surgery**?*	Father has aortic aneurysm
	High clutter	*A 79-year-old male presents with acute onset of severe chest pain radiating to the back and shortness of breath. Your medical student believes he has aortic dissection. You open the patient's chart to check for a history of aortic issues. What year was his most recent aortic **root repair**?*	Mother had aortic dissection

The dependent measures consisted of performance data, the proposed eye movement metrics, and subjective data. The performance measures were *search time*, or the time to find the search target (excluding cases where the answer was incorrect, a response was not provided, or the trial timed out), *screen time* (the total amount of time spent scanning the display, including both search and noticing tasks), *search accuracy* (whether participants correctly identified the search target), and *noticing accuracy* (whether participants detected the important information in each EMR page). For both types of accuracy, an error was defined as any case where the participant either did not respond or provided the incorrect answer. Finally, subjective measures included clutter ratings and workload assessments using the NASA Task Load Index (TLX; Hart & Staveland, 1988).

Experiment procedure. First, participants were given a training session which included a sample task scenario. This was done in order to familiarize them with the different tasks and required responses. Participants were informed that this study was intended to evaluate the design of EMRs. Next, the eye tracker was calibrated, after which the experiment began. For each trial, the screen first displayed an *introduction page* that had a description of the patient's symptoms, the initial diagnosis made by a medical student, the specific question related to the proposed diagnosis, and whether participants had 20 seconds or unlimited time for that task. The presentation order of the variables was counterbalanced by dividing participants into two groups, each of which did the experiment in a fixed random order. The order of the second group was the inverse of the first. When participants were ready, they pressed a button to replace the question screen with the corresponding *EMR page*. Participants then searched through that page to find the answer to the given question. They were instructed that the search task was their priority, and they needed to provide the answer verbally first and also think aloud throughout.

After participants had provided their answer to the search task, they continued scanning the display to look for other case-relevant information for as long as they needed (though limited to 20 seconds in the high-stress conditions). The EMR page was then replaced by a *question page* that asked participants to explain whether and why they thought the initial diagnosis was likely or not. They completed all trials with no breaks in between, with the 24 pages taking around 30 minutes. After completing all trials, participants filled out a debriefing questionnaire in which they entered clutter ratings for sample images and NASA-TLX ratings for the low- and high-stress conditions. The full debriefing questionnaire can be seen in Appendix D.

Results

Unless otherwise specified, results were analyzed using a 2x2x2 repeated-measures analysis of variance (ANOVA) with listwise deletion. Bonferroni corrections were applied for multiple statistical tests. Significance was set at $p < 0.05$ and partial eta squared (η_p^2) was used as a measure of effect size. The ANOVA results are reported for statistically significant results only, with some descriptive values highlighting notable trends. The various experimental conditions will be referred to using L/H for low/high stress, and S/D for simple or difficult tasks. Error bars on graphs indicate the standard error of the mean (*SEM*).

Subjective results. Participants were asked to provide a clutter rating on a 10-point scale (with 10 being the highest possible degree of clutter), for both the low and high clutter versions of each page. These ratings were analyzed using a Wilcoxon Exact Sign Test for ordinal data.

Results showed a significant effect of clutter on the ratings for each page (see Table 3.3). For the Medical History page, the ratings increased from a median of 3 (interquartile range (*IQR*) = 2-4) in low clutter to 8 (*IQR* = 5-8). For the Laboratory Page, the ratings increased from 2 (*IQR* = 1-3) in low clutter to 5 (*IQR* = 4-7) in high clutter. Finally, for the Patient Timeline page, the ratings increased from 7 (*IQR* = 5-9) in low clutter to 9 (*IQR* = 8-9) in high clutter.

Table 3.3: Median clutter ratings for the low- and high-clutter conditions (values in parentheses indicate the interquartile range).

	Low clutter median (*IQR*)	**High clutter median (*IQR*)**	**Wilcoxon exact sign test**
Medical History	3 (2-4)	8 (5-8)	z = 3.31, p = 0.001
Laboratory Results	2 (1-3)	5 (4-7)	z = 3.19, p = 0.001
Patient Timeline	7 (5-9)	9 (8-9)	z = 2.69, p = 0.007

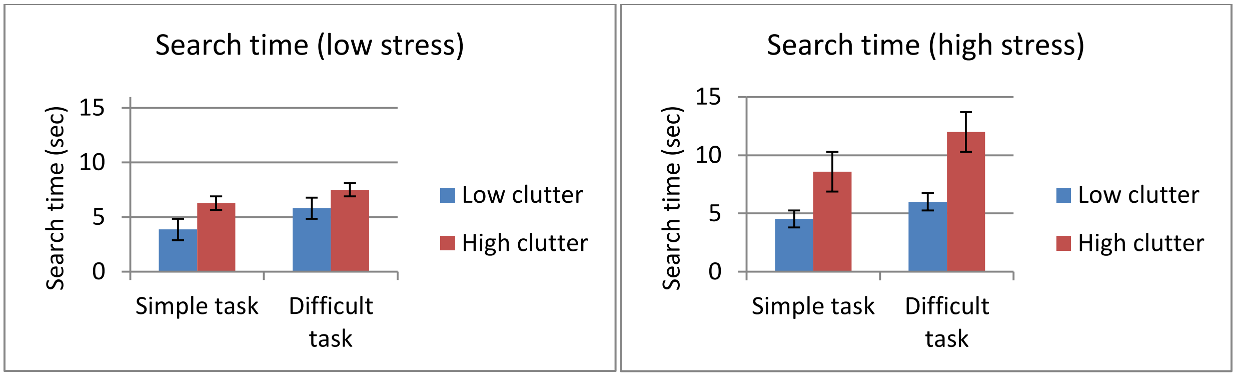
Participants also rated their perceived mental workload in the low-stress and high-stress conditions using NASA-TLX scales (scale: 0 to 20 (largest effect)). These rankings were also analyzed using a Wilcoxon test. As can be seen in Table 3.4, there were significant effects of stress on all six scales. The largest median increase for was temporal demand, which increased from 7 (*IQR* = 1.5-7.5) in low stress to 15 (*IQR* = 13.5-18) in high stress. The ratings for mental demand also increased from 8 (*IQR* = 5.5-11.5) in low stress to 12 (*IQR* = 9.5-14.5) in high stress, while performance ratings increased from 6 (*IQR* = 4-8) to 11 (*IQR* = 8-15) and effort ratings increased from 7 (*IQR* = 4-10) to 12 (*IQR* = 10.5-16).

Table 3.4: Median NASA-TLX ratings (values in parentheses indicate the IQR)

NASA-TLX scale (0-20)	**Low stress median (*IQR*)**	**High Stress median (*IQR*)**	**Wilcoxon exact sign test**
Mental demand	8 (5.5-11.5)	12 (9.5-14.5)	z = 2.677, p = 0.007
Temporal demand	4 (1.5-7.5)	15 (13.5-18)	z = 3.411, p = 0.001
Performance	6 (4-8)	11 (8-15)	z = 3.304, p = 0.001
Effort	7 (4-10)	12 (10.5-16)	z = 3.303, p = 0.001

Performance results. The following sections will describe the observed significant effects on search time, screen time, search accuracy, and screen accuracy.

***Search time*.** For the Medical History page, clutter caused a slight overall increase in search time from 5.8 (*SD* = 0.88) seconds in the low clutter condition to 7.7 (*SD* = 2.13) seconds in the high clutter condition, with a significant simple simple main effect of clutter in the high stress and difficult task condition only (*F*(1, 10) = 9.09, *p* = 0.013; Figure 3.3). There were no significant interaction effects between clutter and stress, or between clutter and task. There were also no significant main effects of clutter or any significant interaction effects between clutter and stress or clutter and task for the Laboratory Results and Patient Timeline pages.



(a) (b)

Figure 3.3: Response time in the different clutter, stress, and task conditions for the Medical History page.

Screen time. To calculate the screen time (the time that participants needed to extract relevant information from the EMR page), the high-stress scenarios that involved a 20-second time limit were excluded. For the low-stress conditions, clutter resulted in a significant increase in screen time (*F*(1, 14) = 11.6, *p* = 0.004, *F*(1, 14) = 17.9, *p* = 0.001, and *F*(1, 14) = 5.6, *p* = 0.032 for the Medical History, Laboratory Results, and Patient Timeline Pages, respectively; see

Figure 3.4). For the Medical History page, screen time increased from 22.2 (*SD* = 3.0) seconds in low clutter to 33.7 (*SD* = 3.4) seconds in high clutter. For the Laboratory Results page, screen time increased from 23.1 (*SD* = 0.53) seconds in low clutter to 33.0 (*SD* = 3.8) seconds in high clutter, and for the Patient Timeline page the values were 31.0 (*SD* = 1.2) seconds in low clutter and 38.2 (*SD* = 5.2) in high clutter.

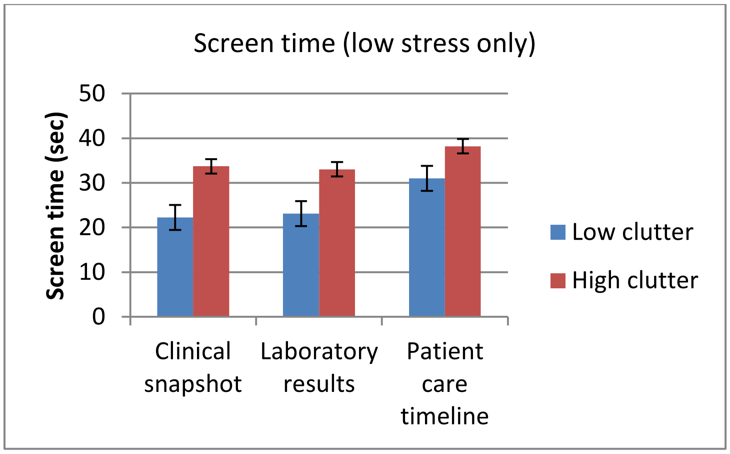
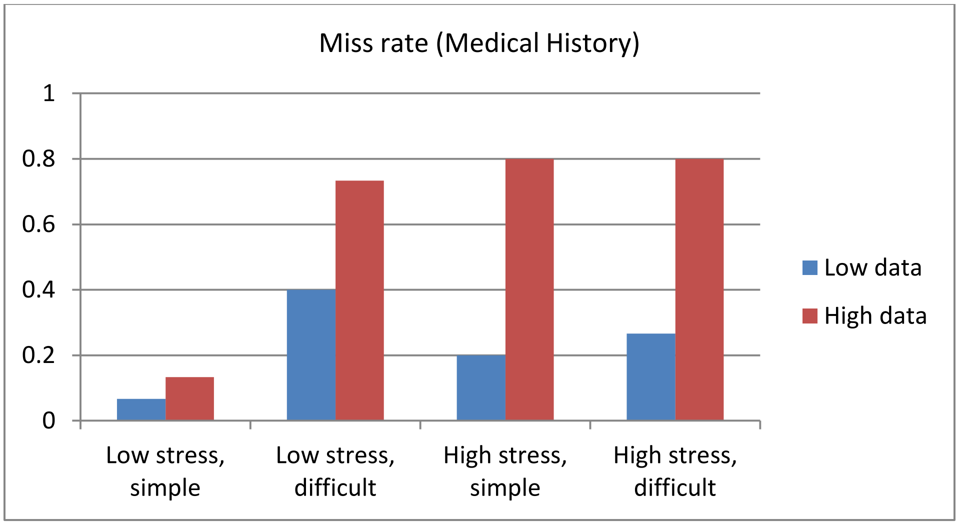


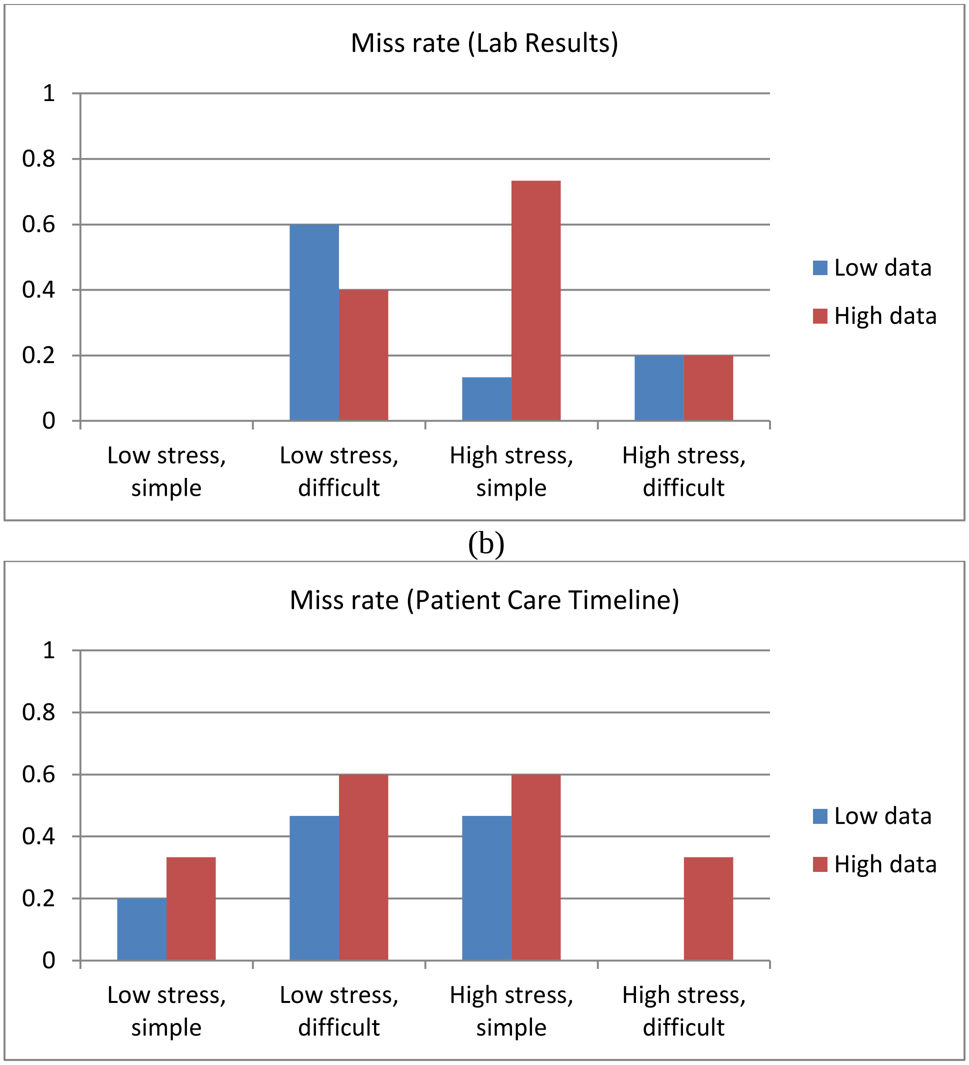
Figure 3.4: Screen time in seconds for the low stress conditions for all three EMR pages.

Search and noticing miss rate. For the Medical History page, out of a total of 60 search targets (8 pages x 15 participants), only 2 targets were missed in low clutter and 6 in high clutter across all participants. The average miss rate was 0.23 (*SD* = 0.12) in low clutter and this increased to 0.61 (*SD* = 0.28) in high clutter. The number of misses also increased slightly in the Patient Timeline page, from 5 misses in low clutter to 11 in high clutter (miss rate of 0.28 (*SD* = 0.19) in low clutter and 0.46 (*SD* = 0.13) in high clutter). On the other hand, for the laboratory results page, participants did not miss any targets in low or high clutter. To determine the effects of clutter, stress, and task on the likelihood of participants missing a target, a mixed-model binary logistic regression was performed for each page with clutter, stress, and task as fixed effects and participants as the random effects. None of the predictor variables were statistically significant for any of the pages.

The effect of clutter was more prominent for the noticing tasks. As described for the search miss rate, a mixed-model binary logistic regression was again used to analyze the effects of the different variables on the noticing miss rate. For the Medical History page, clutter (odds ratio (*OR*) = 5.2, 95% confidence interval (*CI*): 2.1 - 12.8, *p* < 0.001), stress (*OR* = 3.8, *CI*: 1.5 - 9.2, *p* = 0.003), and task (*OR* = 3.8, *CI*: 1.5 - 9.2, *p* = 0.003) were significant predictors of misses (see Figure 3.5). For the Medical History, high clutter resulted in a significant increase in the miss rate, from an overall mean of 0.067 (*SD* = 0.12) in low clutter to 0.61 (*SD* = 0.28; see Figure 3.5). This was not the case, however, for the Laboratory Results page, where none of the variables were significant predictors (the mean miss rate was 0.23 (*SD* = 0.22) in low clutter and 0.33 (*SD* = 0.27) in high clutter). As for the Patient Timeline page, only clutter was a significant predictor (*OR* = 0.36, *CI*: -1.8 - -0.21, *p* = 0.013). The mean miss rate for this page was 0.28 (*SD* = 0.19) in low clutter and 0.46 (*SD* = 0.13) in high clutter.



(a)



(b)

(c)

Figure 3.5: Number of noticing misses in the different conditions (LS: low stress, simple task; LD: low stress, difficult task; HS: high stress, simple task; HD: high stress, difficult task). Values here are summed across all participants.

Eye tracking results. Given that the Medical History page was the only one that showed consistent performance decrements in high clutter, the eye tracking data were analyzed for that page only. One participant's data had to be discarded due to poor eye tracking discrimination, resulting in a total of 14 participants for the eye tracking analysis. Results showed a main effect of clutter on the three spread metrics: convex hull area, spatial density, and NNI (see Table 3.5).

Convex hull area significantly increased from 2,817 (*SD* = 524) pixels^2 in low clutter to 3,834 (*SD* = 841) pixels^2 in high clutter. Similarly, spatial density significantly increased from 0.014 (*SD* = 0.0017) in low clutter to 0.017 (*SD* = 0.0023) in high clutter and NNI significantly increased from 1.8 (*SD* = 0.024) in low clutter to 2.5 (*SD* = 0.028) in high clutter. On the other hand, two directness metrics, scanpath length per second and mean saccade amplitude, significantly decreased in high clutter. Scanpath length per second decreased from 70.8 (*SD* = 6.0) pixels/sec in low clutter to 56.8 (*SD* = 5.3) in high clutter, whereas mean saccade amplitude decreased from 36.9 (*SD* = 3.7) pixels in low clutter to 30.6 (*SD* = 2.5) in high clutter. The duration metric of mean fixation duration was not significantly affected by clutter.

Table 3.5: Mean values of the search metrics in the different clutter conditions.

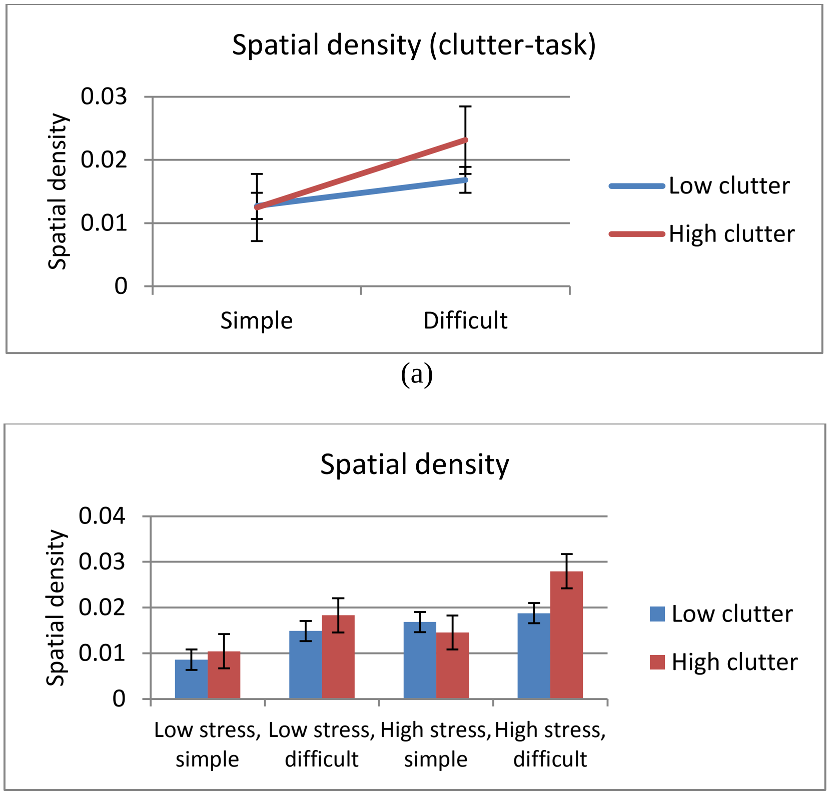
Eye tracking metric	**Low clutter mean (*SEM*)**	**High clutter mean (*SEM*)**	**Main effects of clutter**	**Interaction effects**
		Spread metrics		
Convex hull area (pixels^2)	2,817 (524)	3,834 (841)	*F*(1, 10)[1] = 4.958, *p* = 0.05, η_p^2 = 0.331	-
Spatial density	0.0148 (0.00170)	0.0178 (0.00230)	*F*(1, 13) = 6.822, *p* = 0.022, η_p^2 = 0.344	Clutter-task: *F*(1, 13) = 8.429, *p* = 0.012, η_p^2 = 0.393
Nearest neighbor index (NNI)	1.87 (0.0248)	2.58 (0.0287)	*F*(1, 13) = 12.13, *p* = 0.006, η_p^2 = 0.326	-
		Directness metrics		
Scanpath length per second (pixels/sec)	70.8 (6.00)	56.8 (5.3)	*F*(1, 13) = 10.452, *p* = 0.007, η_p^2 = 0.446	-
Mean saccade amplitude (pixels)	36.9 (3.7)	30.6 (2.5)	*F*(1, 13) = 5.058, *p* = 0.042, η_p^2 = 0.28	Clutter-stress: *F*(1, 13) = 12.902, *p* = 0.003, η_p^2 = 0.498

Backtrack rate (/sec)	0.90 (0.099)	0.81 (0.10)	Not significant	-
Rate of transitions (/sec)	1.7 (0.13)	1.6 (0.13)	Not significant	-
Duration metrics				
Mean fixation duration (sec)	0.42 (0.041)	0.46 (0.080)	Not significant	-

[1] Note that in three cases the convex hull area could not be calculated because there were not enough fixation points to create the triangulation matrix

In the case of spatial density, the main effects need to be interpreted with caution since there were also significant interaction effects (see Figure 3.6, and Table 3.5). In particular, there was a significant clutter-task interaction (Figure 3.6, left) and a simple main effect of clutter in the difficult conditions ($F(1, 13) = 8.8$, $p = 0.011$, $\eta_p^2 = 0.39$). As can be seen in Figure 3.6, the increase in spatial density in the difficult condition can be largely attributed to the increase in the high stress and difficult task condition.

There was also a significant clutter-stress interaction effect for mean saccade amplitude (see Figure 3.7 and Table 3.5). In this case, clutter resulted in a significantly shorter mean saccade amplitude in the low stress condition (simple main effect of clutter in low stress condition only; $F(1, 13) = 10.4$, $p = 0.007$, $\eta_p^2 = 0.44$).



(a)

Figure 3.6: (a) Significant interaction effects for spatial density; (b) note the most noticeable increase in the HD conditions.

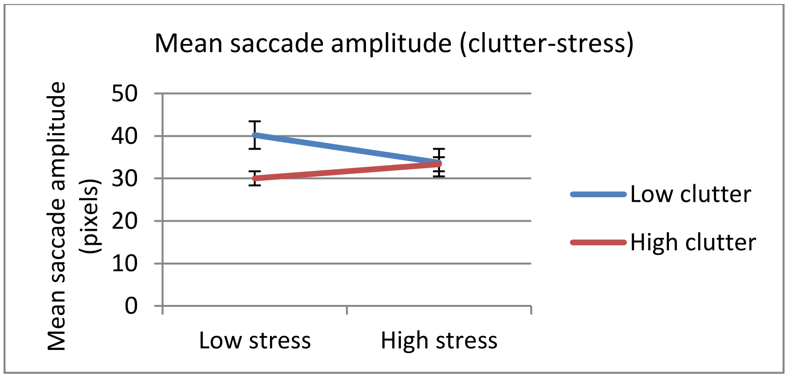


Figure 3.7: Significant clutter-stress interaction effects for the mean saccade amplitude

Experiment 2 Discussion

The goals of this study were twofold. First, the study served to assess and compare the effects of clutter on visual search and noticing tasks. The second goal was to examine whether stress interacted with, and possibly exacerbated the performance effects of, clutter.

The subjective data confirm that the manipulation of both clutter and stress was successful. Overall, the Medical History page was affected most significantly by clutter, resulting in increased search and screen time. In other words, it not only took participants longer to find their search target, but it also took more time to scan and extract all meaningful information from the display. This finding is consistent with results of most earlier studies on display clutter (e.g., Beck et al., 2012; Neider & Zelinsky, 2011). For the Laboratory Results and Patient Timeline pages, clutter significantly affected the screen times, but not the search times. The subjective ratings of clutter help interpret these results. In the case of the Laboratory Results page, the relatively low clutter ratings in both the low and high clutter condition suggest that excess information was not a major problem; participants may have easily filtered out irrelevant information. On the other hand, the timeline display received a relatively high median rating of seven even for the low clutter condition. One possible reason for this result is that the organization of this display was so poor that it rendered the effect of data density negligible.

Overall, the miss rate for the search task was very low throughout and increased only slightly in high clutter. This is not surprising given that they were searching through one static page and told to be sure to be accurate. However, clutter was detrimental to the noticing task. Despite the increased screen time for the high-clutter displays, participants missed more noticing targets in the high-clutter than in the low-clutter conditions. Except for the Laboratory Results page, clutter was a significant predictor of misses, and higher clutter always led to higher noticing miss rates in all stress and task conditions. This held true even in the low stress cases, where participants could spend as much time as they wanted on the display. Moreover, for the Medical History page, results were in line with the expectation that stress and task difficulty would exacerbate the effects of clutter. The simple main effect of clutter on response time was most prominent in the high stress and difficult task condition. The fact that stress led to increased search time in these cases, despite the fact that participants had a time limit and were instructed to search as fast as possible, confirms the suspected interaction between clutter and stress.

On the Medical History page, performance effects were reflected most clearly in spread metrics (specifically, convex hull area, spatial density, and NNI). This finding suggests that participants were distracted by the unimportant data, which drew their attention in the absence of any other cues or strong top-down model of where to locate information. The resulting large spread of fixations on irrelevant data likely resulted in the increased response time. High clutter also affected directness measures, particularly during search. Specifically, mean saccade amplitude and scanpath length per second decreased significantly. Thus, while fixations overall were spread widely across the screen, successive fixations were not very far apart. It seems that participants adopted a slower and more orderly search to compensate for high clutter. Finally, the duration metric was not significantly affected by clutter. This suggests that extracting and understanding information was not a problem for participants in this experiment.

The performance decrements resulting from high stress and task difficulty were also reflected in the spatial density and scanpath length. Thus, any algorithms looking to identify the effects of clutter for a judgment and multiple comparison task, especially under high stress, may want to consider using these metrics. On the other hand, mean saccade amplitude would appear to be less useful in the case of high stress as the effect of clutter was significant only in low stress and simple task conditions.

Discussion and Conclusion

This chapter detailed two experiments that were conducted mainly in order to 1) identify the eye movement metrics that are significantly affected by clutter during information acquisition and 2) use these eye movement metrics to explain performance effects of clutter by highlighting underlying changes in attention allocation. Table 3.6 shows a summary of the eye tracking results obtained from these three experiments.

With regards to the first goal, there were a number of metrics that were significantly affected by clutter across both experiments. Notably, the spread metrics (convex hull area, spatial density, and NNI) all significantly increased, whereas mean saccade amplitude – a directness metric – showed a significant decrease in both experiments. Other metrics, such as rate of transitions and mean fixation duration, were significantly affected only in Experiment 1, which suggests that particular aspects of these displays that were not present in Experiment 2 were responsible for these differences. For example, Experiment 1 contained a considerable amount of color variation, something that was not a feature of the Medical History page.

Results from Experiment 2 suggest that stress interacted with clutter to exacerbate its effects. In particular, the number of fixations, spatial density, and scanpath length were all more strongly affected by clutter in the presence of high stress. This suggests that these metrics would be preferred over others to detect the effects of clutter in stressful situations.

Regarding the second goal of the above studies, namely using eye movement metrics to explain any observed performance effects of clutter by analyzing underlying changes in attention allocation, clutter appears to cause eye fixations to be more spread out or dispersed across far-reaching areas of the display. This dispersion then contributed to the increased response time in the case of high clutter. At the same time, the scanpath length per second (i.e., speed of search) and distance between successive fixations were lower in high clutter, implying that users were searching the display more slowly and systematically. This tendency may have contributed further to the increased search time in high clutter.

Table 3.6: Summary of the eye tracking results for Experiments 1 and 2 in the case of an increase in the amount of data density in the display.

Eye tracking metric	**Visual search**	
	Experiment 1: visual search in a simple environment	**Experiment 2: visual search in the ED**
Spread metrics		
Convex hull area	*Increase*	*Increase*
Spatial density	*Increase*	*Increase*
Nearest neighbor index[1]		*Increase*
Directness metrics		
Scanpath length per second		Decrease
Mean saccade amplitude	Decrease	Decrease
Backtrack rate		
Rate of transitions	*Increase*	
Duration metrics		
Mean fixation duration	*Increase*	

[1]Note that this was not calculated in Experiment 1

In conclusion, the key points learned from the above two experiments were that

1) The effects of clutter (specifically, data density) on visual search can be detected using a number of eye movement metrics, namely convex hull area, spatial density, mean saccade amplitude, and scanpath length per second.
2) Spread metrics, and spatial density in particular, seem to be the most diagnostic and sensitive eye movement metrics available, followed by mean saccade amplitude among the directness metrics. These metrics indicate that the decrements in performance due to data density can be largely attributed to increased spread and slower search.

The discrepancies observed for some of the eye movement metrics pointed to the need to systematically vary data density and organization in order to isolate and better understand the contribution of each aspect. It was not clear, for example, why the rate of transitions and saccade amplitude rate were significantly affected by clutter in one experiment and not the other. Chapter 4 will detail the efforts to achieve this goal of delineating the effects of the two main aspects of clutter.

References

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision research*, *39*(17), 2947-2953.

Beck, M. R., Trenchard, M., van Lamsweerde, A., Goldstein, R. R., & Lohrenz, M. (2012). Searching in clutter: Visual attention strategies of expert pilots. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1411-1415). Boston, MA: Sage Publications.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 443-446.

Dornheim, M. A. (1995, January 30). Dramatic incidents highlight mode problems in cockpits. Aviation Week and Space Technology, 6–8.

Handford, M. (1997a). *Where's Waldo Now?* Somerville, MA: Candlewick Press.

Handford, M. (1997b). *Where's Waldo? The Wonder Book.* Somerville, MA: Candlewick Press.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology,52*, 139-183.

Joint Planning and Development Office (JPDO). (2010*). Concept of Operations for the Next Generation Air Transportation System*. Version 3.1, 10 April 2010. Washington, DC: Federal Aviation Administration.

Klein, R. M., & MacInnes, W. J. (1999). Inhibition of return is a foraging facilitator in visual search. *Psychological science*, *10*(4), 346-352.

Neider, M. B., & Zelinsky, G. J. (2011). Cutting through the clutter: Searching for targets in evolving complex scenes. *Journal of vision*, *11*(14), 7.

NextGen Integration and Implementation Office. (2010). *NextGen Implementation Plan*. Washington, DC: Federal Aviation Administration.

Nikolic, M. I., & Sarter, N. B. (2007). Flight deck disturbance management: A simulator study of diagnosis and recovery from breakdowns in pilot-automation coordination. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(4), 553-563.

Otero-Millan, J., Troncoso, X. G., Macknik, S. L., Serrano-Pedraza, I., & Martinez-Conde, S. (2008). Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator. *Journal of Vision*, *8*(14), 21.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision, 10*(4), 437-442.

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(3), 347-357.

Schmidt, R. W. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. W. Schmidt (Ed.), Attention and awareness in foreign language learning (pp. 1–63). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

CHAPTER 4

The Contribution of Data Density and Display Organization to Clutter and its Effects on Attention and Performance

In the previous experiments, clutter was varied by manipulating data density. However, data density is only one display-based aspect of clutter referenced in the proposed definition of the phenomenon. The second aspect, display organization, has also been proposed as one contributor to clutter in earlier work (e.g., Bravo & Farid, 2008; Doyon-Poulin et al., 2012; Rosenholtz, Li, Mansfield, & Jin, 2005). It is therefore not clear to what extent the observed performance and attentional effects observed in the previous experiments were the result of data density, poor organization, or a combination of both factors. In addition, Experiment 2 showed that stress exacerbated the effects of high data density; however, it is not clear to what extent stress would interact with display organization also.

These gaps were addressed in Experiment 3. The specific aims were to 1) isolate the effects of data density and poor organization, in combination with stress, on performance and attention during visual search, and 2) identify the eye movement metrics that can best trace and differentiate between these effects. In this study, participants were asked to perform search tasks in a fairly simple simulated graphics program. This type of display was selected as it afforded the possibility of carefully and systematically varying both data density and poor organization.

Methods

Participants

The participants in this experiment were 20 engineering students from the University of Michigan (12 male and eight female). Their average age was 22.5 years (standard deviation (*SD*) = 5.4), and they all had used Adobe Photoshop ® before. All participants gave informed consent and were paid $15 for completing the experiment. Participants had self-reported normal or corrected to normal vision; contact lenses were allowed but glasses were not, due to the limitations of the eye tracker. This study was approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board (ID: #HUM00078246).

Simulated Graphics Program

A mock graphics program was created that consisted of icons used in familiar graphics packages (for a similar approach, see Goldberg and Kotval (1999)). Icons from different Adobe ® graphics suites were presented around the perimeter of the display, as well as some closer to the center (see Figure 4.1). The icons were all assigned to a unique group based on their function (for example, Write, Align, Color, and Grids).

Experiment Setup

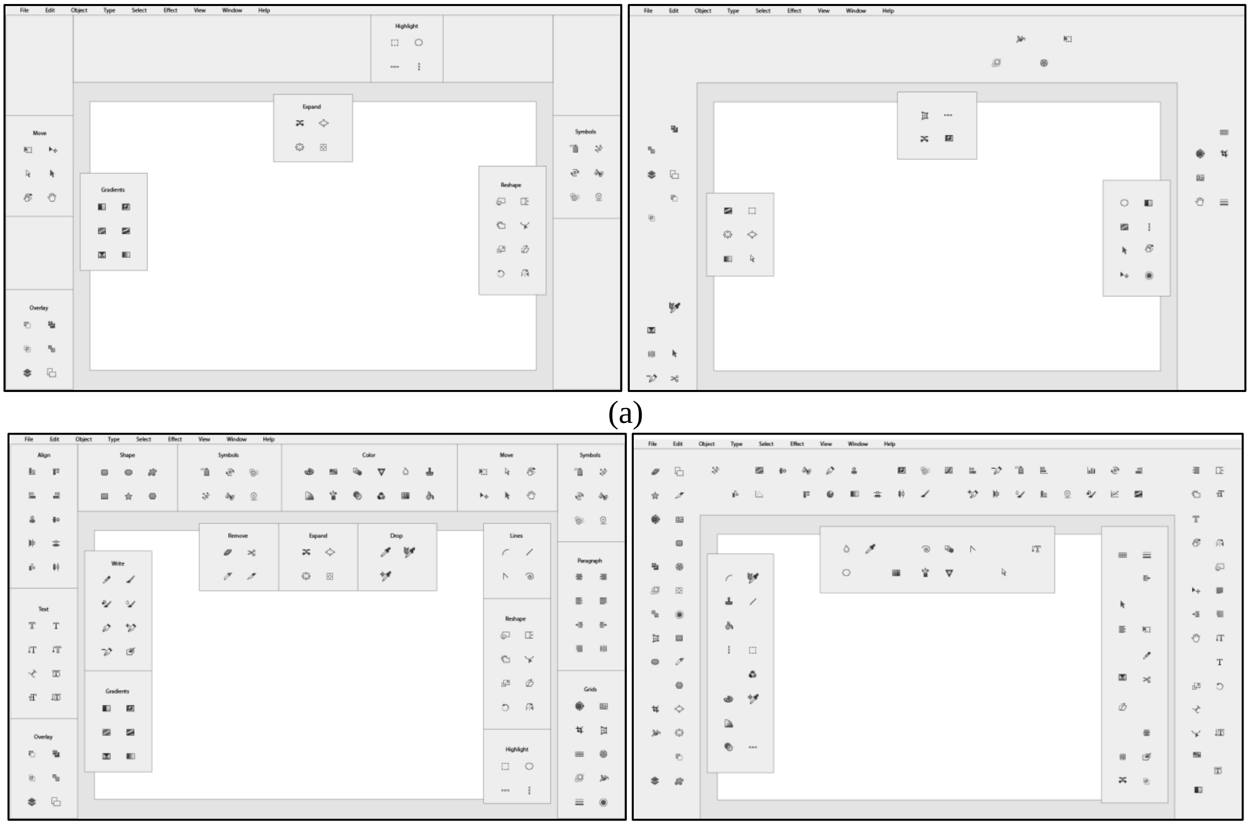
The simulated graphics program was displayed on a 19-inch monitor with a resolution of 1280x1024 pixels. An Applied Science Laboratories D-6 desktop-mounted eye tracker was used (sampling rate: 60 Hz; accuracy: less than 1 degree visual angle; precision: 0.5 degrees visual angle). This eye tracker was placed in front of the computer monitor such that the aperture was approximately 5 cm from the screen. The participants were seated at a distance of around 50 cm inches from the monitor, making for a visual angle of around 25 degrees in the horizontal direction and 20 degrees in the vertical directions. The lighting conditions were the same for all participants. The calibration procedure, which took around 5 minutes at the start of the experiment, was accomplished using a 9-point grid.

Experiment Design

The three independent variables in this study were the density of data (low, high), display organization (good, poor), and stress (no stress, stress). These were varied within participants, leading to a 2x2x2 full-factorial design. In each experiment trial, participants had to search for a target icon in a given display. The variables in each trial were manipulated as follows:

- *Density of data (low, high):* the density of data was operationally defined as the number of graphics/paint icons within the display (see Figure 4.1). The high-data displays all had the complete set of 127 icons, whereas all low-data displays had between 38 and 45 icons. This slight difference resulted from the fact that not all icon groups consisted of the same number of elements. They contained four, six, or eight icons, depending on the icons available in the Adobe ® suites.

- *Organization (good, poor):* in the well-organized displays, the icons were placed in their respective groups, surrounded by a thin outline box. The number and sequence of the icons within these groups was always the same, and the group name was provided above each group. In contrast, in the poorly organized display, the icons were randomly distributed across the display.
- *Stress (no stress, stress):* for the no-stress trials, participants were told only that they needed to be as accurate as possible in their search for the target (i.e., to keep searching until they find the correct target). However, in the stress trials, participants had to find the target as quickly as possible to collect additional points. Participants received 20 points if they responded within 1 second, 19 points for responding within 2 seconds, and so forth, up until 20 seconds. If finding the target took them longer than 20 seconds, they would lose one point per second. In addition, participants were cautioned against compromising accuracy; they were informed that incorrect answers would result in losing 20 points. To increase participants' motivation to do well, the three participants with the highest total number of points at the end of the experiment received an additional $75, $50, and $25 dollars, respectively. During the stress trials, there was also an auditory count from 1 to 20 to let participants know how many seconds had elapsed.



(a)

(b)

Figure 4.1: (a) low-data displays with good organization (left) and poor organization (right) and (b) high-data displays with good organization (left) and poor organization (right).

Participants were presented with three instances of each experimental trial, making for a total of twenty-four trials and twenty-four corresponding (unique) displays. All of the targets were located in one of three eccentricities (around 10 º, 25 º, and 30 º visual angles) in the perimeter of the display. Eccentricity of the target was counterbalanced between the trials such that each experiment condition was performed in each of the three eccentricities. There were two different search targets for each full set of eight conditions, and these two were always drawn from the same group of icons. The targets were deliberately selected to be easily recognizable (as opposed to obscure or less-used) icons. In addition, participants underwent training before the experiment to make sure they were familiar with all the target icons and the groups.

The results of Experiment 2 had shown that search for multiple instances of a target was more significantly affected by clutter than search for just one occurrence (see Chapter 3). Participants in this study were therefore instructed that the target could appear more than once and that they should search the display until they detected all instances of the target. In the 24 experiment trials, there was always exactly one instance of the target. In addition, there were 16 "dummy" trials where the target was either absent, shown in the middle section of the display, or repeated (i.e., there were two or three instances of the target). These trials served to ensure that participants always searched the entire display for multiple targets. However, the performance data from these trials was not included in the analysis.

The presentation order of the variables was counterbalanced by dividing participants into two groups, each of which completed the experiment in a fixed random order. The order of the second group was the inverse of the first, and the stress condition was varied in blocks of ten. Participants did sets of ten low-stress or high-stress trials (see Figure 4.2).

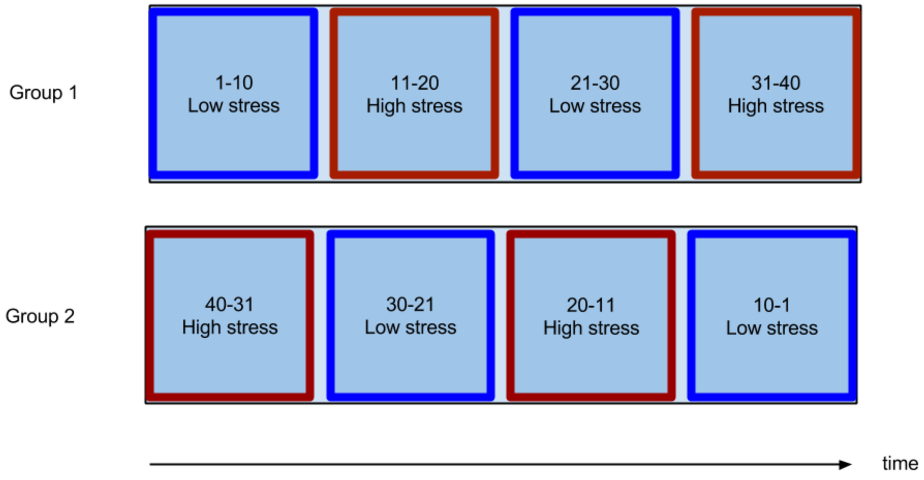


Figure 4.2: Illustration of the order of trials for Group 1 and Group 2. Each of the numbered scenarios within each block were randomly either low or high clutter

Experiment Procedure

Participants were first asked to read and sign the consent form. Next, they completed a 20-minute training session. After being informed of the goals and rules of the experiment, participants were shown the various icon groups. All of the target icons (i.e., the subset that they would have to find later) were highlighted and named. After reviewing all the groups with the experimenter, participants had to complete three tests: 1) they were shown images of the target icons and asked to provide the name of each icon, 2) they were shown all of the icon groups without their names and asked to provide the name of each group, and 3) they were shown the name and image of the target icons and had to identify the group that each icon belonged to. Participants could repeat each test until they achieved the 100% accuracy required to proceed.

Next, participants had to complete a set of four practice trials. The instructions and controls for these trials were identical to those of the actual experiment. For each trial, participants were first presented with the target icon (name and image) that they needed to look for. The screen also indicated whether points could be collected (stress condition) or not (no-stress condition). They could look at the target icon as long as necessary, and then pressed the right arrow key when they were ready to start the search. The experiment display appeared, and the response time counter started. If participants thought there was no target or multiple instances of the target, they pressed the “down” arrow key. They pressed the “up” arrow key in case of one and only one instance of the target.

Following the training session, the eye tracker was calibrated, and participants proceeded with the forty experiment trials. After completing all trials, participants filled out a debriefing questionnaire, where they rated the amount of clutter in 8 sample displays from the experiment on a scale from 0 to 10 (with ‘10’ representing the highest possible amount of clutter). They also provided NASA Task Load Index (TLX; Hart & Staveland, 1988) ratings for the stress and no stress conditions as a means of evaluating our manipulation of stress. The full debriefing document can be seen in Appendix E. The experiment took around twenty minutes to complete.

Results

Unless otherwise specified, the data was analyzed using a three-way repeated measures analysis of variance (ANOVA) with listwise deletion. Bonferroni corrections were applied for multiple comparisons. Only significant results ($p < 0.05$) are reported here. In the description of results, L/H, G/P, and N/S refer to the various (L)ow/(H)igh data density, (G)ood/(P)oor organization, and (N)o stress/(S)tress conditions. On two occasions, eye tracking calibration was unsuccessful, and participants’ eye tracking data was discarded. However, these participants still completed the experiment, and their performance data are included in the analysis.

Subjective Results

The median ratings of clutter were 3.0 (*IQR* = 2-4), 4.7 (*IQR* = 3.75-6.25), 6.2 (*IQR* = 4.75-8), and 9.0 (*IQR* = 7.7-10) in the LG, LP, HG, and HP conditions, respectively. There was a significant effect of data ($F(1, 19) = 239$, $p < 0.001$, $\eta_p^2 = 0.85$) and organization ($F(1, 19) = 53.8$, $p < 0.001$, $\eta_p^2 = 0.739$) on the clutter ratings (with ‘10’ being the highest level of clutter), but no significant interaction effects were observed. The ratings increased from 3.8 (*IQR* = 3.43-4.31) in low data to 7.6 (*IQR* = 6.93-8.31) in high data. In the case of poor organization, the ratings also increased from 4.6 (*IQR* = 2.0-7.62) in good organization to 6.8 (*IQR* = 4.00-10.00) in poor organization.

The NASA-TLX ratings were analyzed using a Wilcoxon exact sign test for ordinal data for low and high stress. Results show that there were significant effects of stress on all of the relevant dimensions (see Table 4.1). For the mental demand scale, the ratings increased from 10 (*IQR* = 5.5-13.0) in low stress to 14.0 (*IQR* = 13-16.2) in high stress. For temporal demand, they increased from 8.0 (*IQR* = 4.7-12.5) to 17.5 (*IQR* = 15.7-18.2), while for performance, the ratings increased from 6.0 (*IQR* = 4.0-7.2) to 8.5 (*IQR* = 7.0-10.5). Finally, for the effort scale, the ratings increased from 11.0 (*IQR* = 7.5-14.0) to 15.0 (*IQR* = 12.7-16.2) in the high stress case.

Table 4.1: NASA-TLX results (scale: 0 to 20, with 20 representing the highest possible level of the respective dimension). Values in parentheses indicate the interquartile range (IQR)

NASA-TLX scale (0-20)	**Low stress median (*IQR*)**	**High stress median (*IQR*)**	**Wilcoxon exact sign test**
Mental demand	10.0 (5.5-13.0)	14.0 (13-16.2)	z = 3.42, p = 0.001
Temporal demand	8.0 (4.7-12.5)	17.5 (15.7-18.2)	z = 3.82, p < 0.001
Performance	6.0 (4.0-7.2)	8.5 (7.0-10.5)	z = 2.55, p = 0.011
Effort	11.0 (7.5-14.0)	15.0 (12.7-16.2)	z = 3.71, p < 0.001

Performance Results

Response time. Only correct answers were considered in the calculation of response time. The mean response time was found to be 4.1 (*SD* = 0.82) seconds for LGN, 3.8 (*SD* = 0.93) for LGS, 5.4 (*SD* = 1.51) for LPN, 5.4 (*SD* = 1.5) for LPS, 8.9 (*SD* = 2.8) for HGN, 7.6 (*SD* = 1.8) for HGS, 12.9 (*SD* = 6.2) for HPN, and 10.1 (*SD* = 3.7) for HPS. As expected, high data density ($F(1, 19) = 88.2$, $p < 0.001$, $\eta_p^2 = 0.823$) and poor organization ($F(1, 19) = 23235$, $p <$

0.001, η_p^2 = 0.55) resulted in a significant increase in response time. The dependent measure increased from 4.7 (*SD* = 0.72) seconds in low data to 9.9 (*SD* = 1.9) seconds in high data (see Figure 4.3), and from 6.1 (*SD* = 2.1) seconds in good organization to 8.4 (*SD* = 3.2) seconds in high organization. Stress, on the other hand, resulted in a significant overall decrease in response time, from 7.8 (*SD* = 3.4) seconds in no stress to 6.7 (*SD* = 2.3) seconds in stress (*F*(1, 19) = 10.6, *p* = 0.004, η_p^2 = 0.358).

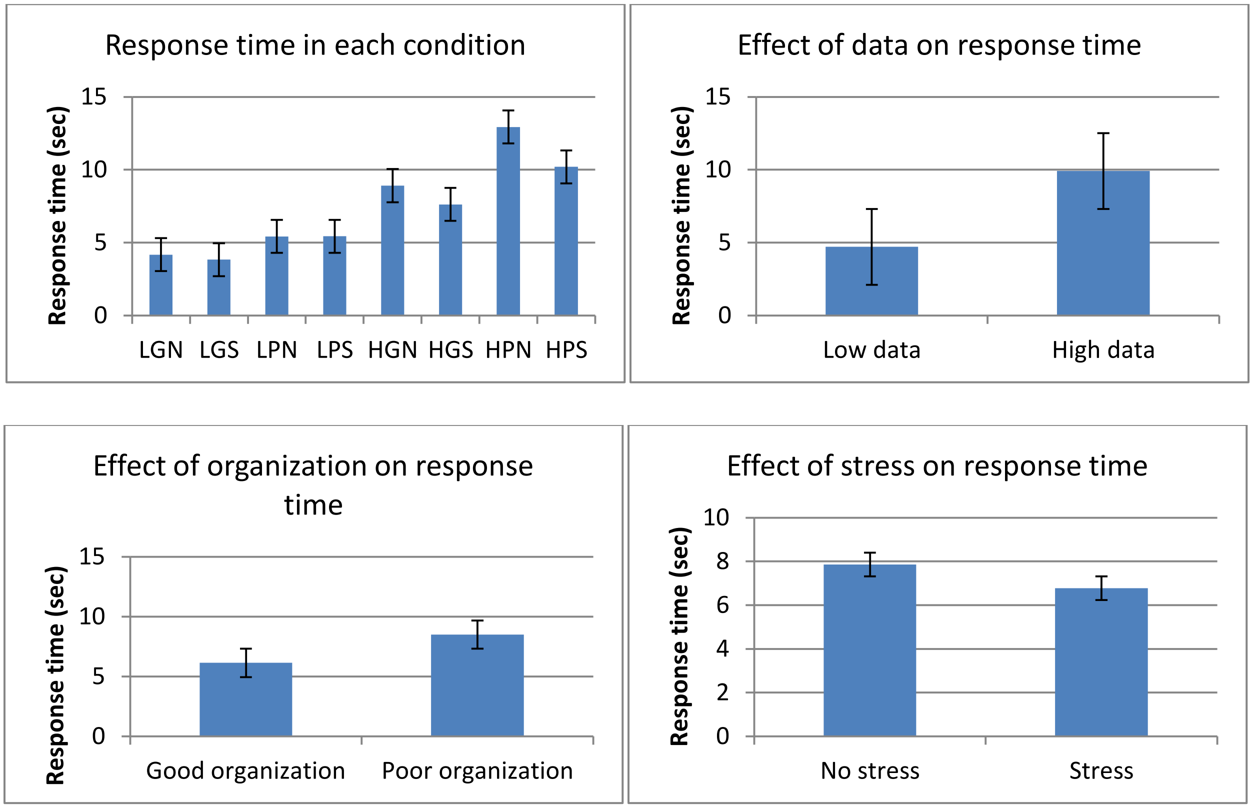
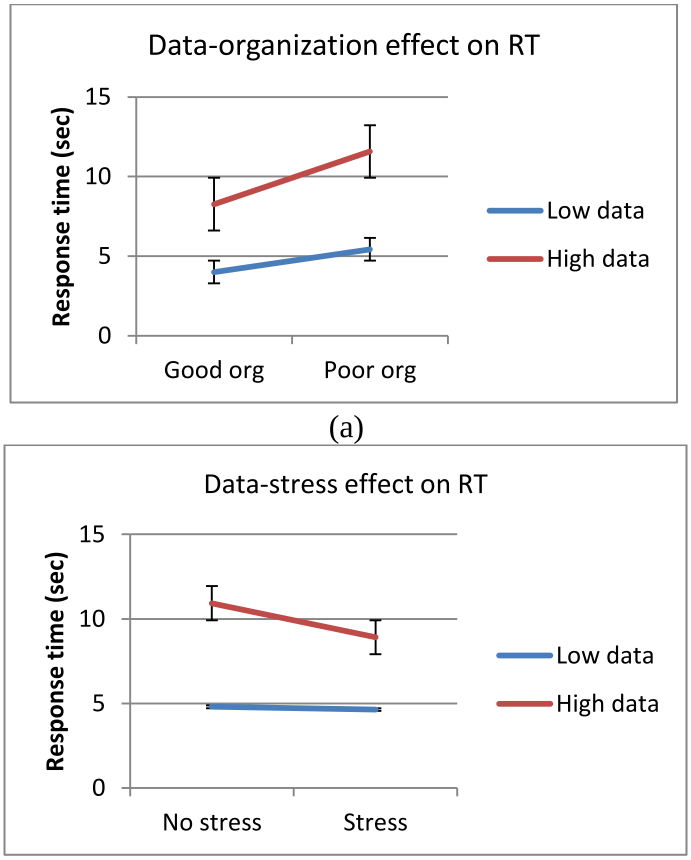


Figure 4.3: Effects of data, organization, and stress on the overall response time (the different conditions are defined by L: low data, H: high data; G: good organization, P: poor organization; N: no stress, S: stress)

These overall effects need to be interpreted with caution, however, as there were significant interaction effects between data and organization (*F*(1, 19) = 6.9, *p* = 0.017, η_p^2 =

0.26), as well as between data and stress ($F(1, 19) = 11.2$, $p = 0.003$, $\eta_p^2 = 0.371$). In both the good ($F(1, 19) = 109.1$, $p < 0.001$, $\eta_p^2 = 0.852$) and poor organization ($F(1, 19) = 53.8$, $p < 0.001$, $\eta_p^2 = 0.739$) conditions, the response time was longer for high data density (see Figure 4.4a). However, this effect was more pronounced in the poor organization condition. Also, while the response time for high data density was significantly higher in both the low ($F(1, 19) = 62.3$, $p < 0.001$, $\eta_p^2 = 0.766$) and high ($F(1, 19) = 111.3$, $p < 0.001$, $\eta_p^2 = 0.854$) stress conditions, the difference between high and low data was less pronounced in high stress (see Figure 4.4b).



(a)

(b)

Figure 4.4: Interaction between (a) data and organization and (b) data and stress

Miss rate. A miss was defined as a participant pressing the “down” arrow key to (incorrectly) indicate either the absence of a target or the presence of multiple targets. Figure 4.5

shows the total number of misses summed across all participants and trials. Results were analyzed using a mixed model binary logistic regression with data, organization, and stress as the fixed effects and participants as the random effect. Results show that data (*OR* = 5.0, *p* = 0.008, confidence interval (*CI*) = [0.42, 2.8]) and organization (*OR* = 6.4, *p* = 0.004, *CI* = [0.58, 3.1]) were significant predictors of misses, with an increased chance of a miss in the high data and poor organization conditions. The mean miss rate was 0.10 (*SD* = 0.30) in all of the four low clutter conditions. The miss rate was 0 in the HGN condition, 0.05 (*SD* = 0.21) in HGS, 0.35 (*SD* = 0.47) in HPN, and 0.65 (*SD* = 0.65) in the HPS condition. The miss rate increased from an average of 0.033 (*SD* = 0) in low data to 0.087 (*SD* = 0.086) in high data, and from an average of 0.002 (*SD* = 0.013) in good organization to 0.10 (*SD* = 0.075) in poor organization.

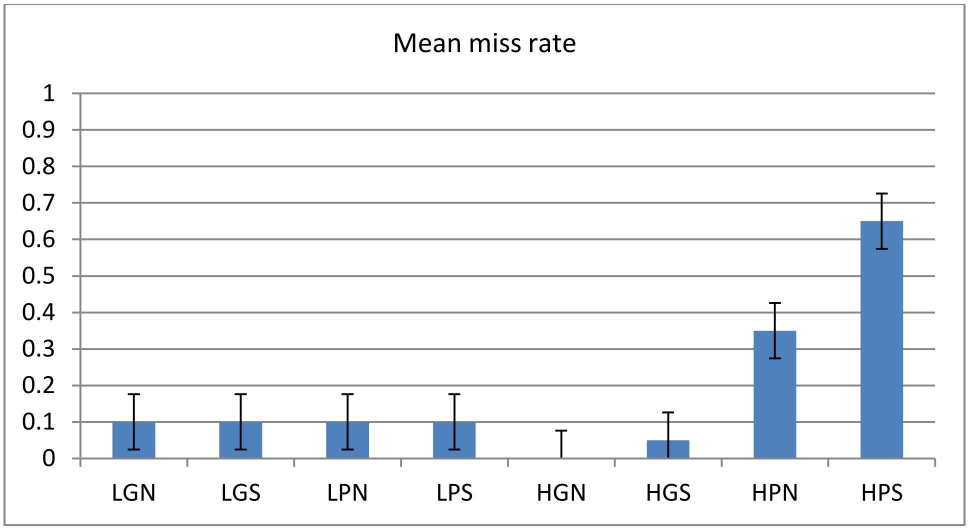


Figure 4.5: Number of misses summed across all participants (the different conditions are defined by L: low data, H: high data; G: good organization, P: poor organization; N: no stress, S: stress)

In addition, there was a significant interaction between data and organization (*OR* = 0.18, *p* = 0.036, *CI* = -3.3, -0.11]). High data led to more misses only when combined with poor organization (see Figure 4.6). In case of good display organization, the number of misses did not differ significantly between the high and low data trials.

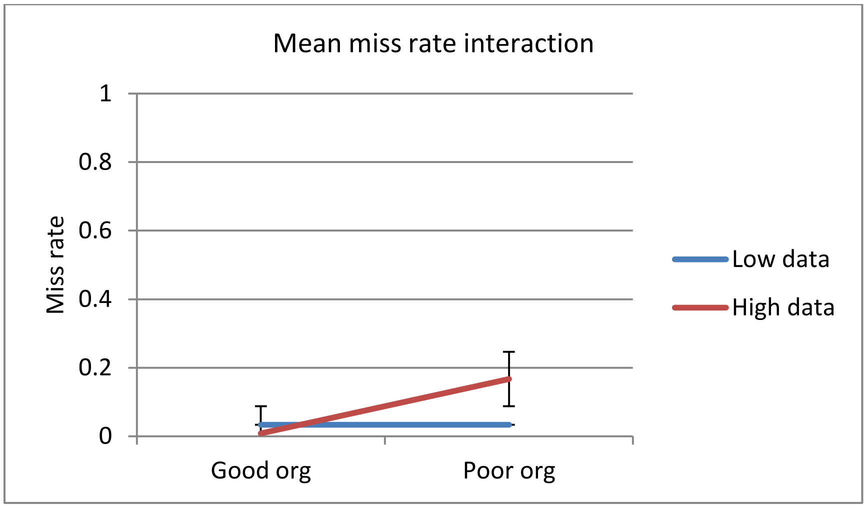


Figure 4.6: Interaction effect on the number of misses between data and organization

Eye Tracking Results

Overall, there was a significant main effect of data density on all metrics, except the rate of transitions (see Table 4.2). The spread metrics all increased significantly in high data. Convex hull area increased from 10,400 (*SD* = 2,400) pixels^2 in low data to 19,200 (SD = 4,500) in high data, spatial density increased from 0.063 (*SD* = 0.01) to 0.13 (*SD* = 0.03), and NNI increased from 0.44 (*SD* = 0.09) to 0.62 (*SD* = 0.07). On the other hand, the directness metrics and mean fixation duration decreased significantly. Scanpath length per second decreased from 143 (*SD* = 30) pixels/sec in low data to 123 (*SD* = 28) pixels in high data, mean saccade amplitude decreased from 55 (*SD* = 9) to 42 (*SD* = 8) pixels, backtrack rate decreased from 0.12 (*SD* =

0.070) to 0.053 (*SD* = 0.030) backtracks/sec, and mean fixation duration decreased from 0.26 (*SD* = 0.040) seconds to 0.24 (*SD* = 0.030) seconds.

Organization significantly affected all metrics, except scanpath length per second and backtrack rate. For two spread metrics – convex hull area and spatial density – as well as for mean fixation duration, poor organization caused a significant increase. Convex hull area increased from 14,300 (*SD* = 5,300) pixels^2 in good organization to 15,300 (*SD* = 5,900) pixels^2 in poor organization, spatial density increased from 0.093 (*SD* = 0.03) to 0.12 (*SD* = 0.05), and mean fixation duration increased from 0.25 (*SD* = 0.03) seconds to 0.263 (*SD* = 0.043) seconds. Conversely, in the case of NNI, mean saccade amplitude, and rate of transitions, poor organization resulted in a significant decrease. NNI decreased from 0.56 (*SD* =0.1) in good organization to 0.50 (*SD* = 0.1) in poor organization, mean saccade amplitude decreased from 50 (*SD* = 11) pixels to 46 (*SD* = 11) pixels, and rate of transitions decreased from 1.85 (*SD* = 0.3) to 1.73 (*SD* = 0.3).

There was a significant data-organization interaction for spatial density, with poor organization resulting in a larger increase in the high data condition as opposed to the low condition. There were also data-stress interactions effects for spatial density, NNI, and rate of transitions. For both NNI and rate of transitions, stress resulted in a more prominent increase in the high data as opposed to low data condition, whereas for spatial density, high stress resulted in a larger decrease. The results of the statistical analysis of interaction effects can be seen in Table 4.3.

Table 4.2: Significant main and interaction effects for the eye movement metrics calculated over the whole response time period. All mean values and standard deviations (in parentheses) are reported, but ANOVA results are given for significant effects/interactions only

	Main effects		**Interaction effects**	
Eye movement metrics	**Data** Low data High data	**Organization** Good organization Poor organization	Data-organization interaction	Data-stress interaction
Location metrics				
Convex hull area (pixels^2)	10,400 (2,400) 19,200 (4,500) $F(1, 16) = 335, p < 0.001, \eta_p^2 = 0.9$	14,300 (5,300) 15,300 (5,900) $F(1, 16) = 8.1, p = 0.012, \eta_p^2 = 0.3$		
Spatial density	0.063 (0.01) 0.14 (0.03) $F(1, 16) = 288, p < 0.001, \eta_p^2 = 0.9)$	0.093 (0.03) 0.12 (0.05) $F(1, 16) = 43, p < 0.001, \eta_p^2 = 0.7)$	$F(1, 16) = 22, p < 0.001, \eta_p^2 = 0.5$	$F(1, 16) = 12, p = 0.003, \eta_p^2 = 0.4$
Nearest neighbor index (NNI)	0.44 (0.09) 0.62 (0.07) $F(1, 16) = 304, \mathrm{p} < 0.001, \eta_p^2 = 0.9$	0.56 (0.1) 0.50 (0.1) $F(1, 16) = 16, \mathrm{p} = 0.001, \eta_p^2 = 0.5$		$F(1, 16) = 7.8, p = 0.013, \eta_p^2 = 0.3$
Directness metrics				
Scanpath length per second (pixels/sec)	143 (30) 123 (28) $F(1, 16) = 81, p < 0.001, \eta_p^2 = 0.8$	139 (28) 127 (32) $F(1, 16) = 14, p = 0.001, \eta_p^2 = 0.4$		
Mean saccade amplitude (pixels)	55 (9) 42 (8) $F(1, 16) = 246, p < 0.001, \eta_p^2 = 0.9$	50 (11) 46 (11) $F(1, 16) = 22, p < 0.001, \eta_p^2 = 0.5$		
Rate of transitions (/sec)	1.80 (0.3) 1.80 (0.3)	1.85 (0.3) 1.73 (0.3)		$F(1, 16) = 5.46, p = 0.033, \eta_p^2 = 0.254$

		$F(1, 16) = 5.1$, $p = 0.037$, $\eta_p^2 = 0.2$	
Backtrack rate (/sec)	0.12 (0.07) 0.053 (0.03) $F(1, 16) = 37$, $p < 0.001$, $\eta_p^2 = 0.6$	0.095 (0.07) 0.082 (0.06)	
Duration metrics			
Mean fixation duration (sec)	0.26 (0.04) 0.24 (0.03) $F(1, 16) = 17$, $p = 0.001$, $\eta_p^2 = 0.5$	0.25 (0.03) 0.263 (0.043) $F(1, 16) = 5$, $p = 0.029$, $\eta_p^2 = 0.2$	

Table 4.3: Details of interaction effects for the eye movement metrics calculated over the whole response time period. Values are only provided in the case there was evidence of significant data-organization interaction or data-stress interaction, as indicated in Table 4.2. ANOVA results given for statistically significant results only.

	Simple main effects of data **Low data** **High data**			
Eye movement metrics	Good organization	Poor organization	No stress	Stress
Spread metrics				
Spatial density	0.059 (0.01) 0.126 (0.0272) $F(1, 17) = 205$, $p < 0.001$, $\eta_p^2 = 0.9$	0.067 (0.01) 0.15 (0.03) $F(1, 17) = 266$, $p < 0.001$, $\eta_p^2 = 0.9$	0.063 (0.01) 0.14 (0.03) $F(1, 17) = 233$, $p < 0.001$, $\eta_p^2 = 0.9$	0.063 (0.01) 0.13 (0.02) $F(1, 17) = 231$, $p < 0.001$, $\eta_p^2 = 0.9$
Nearest neighbor index (NNI)			0.45 (0.08) 0.60 (0.06) $F(1, 17) = 194$, $p < 0.001$, $\eta_p^2 = 0.9$	0.44 (0.1) 0.64 (0.08) $F(1, 17) = 170$, $p < 0.001$, $\eta_p^2 = 0.9$
Directness metrics				
Rate of transitions			1.79 (0.3) 1.75 (0.3)	1.77 (0.3) 1.84 (0.3)

Discussion and Conclusion

The goal of this study was to assess and distinguish between the performance and underlying attentional effects of two aspects of clutter – high data density and poor organization – in combination with stress. To that end, participants' performance on a series of visual search tasks was recorded and eye tracking was used to trace their eye movements while performing these tasks. The previously-proposed eye movement metrics were then calculated to explain any performance effects of clutter.

Performance and Subjective Results

The significant differences between the ratings for the two levels of data and organization suggest that these factors did contribute to appreciable differences in perceived clutter. More importantly, both aspects of clutter significantly degraded performance, resulting in longer response times and more misses. Poor organization exacerbated the effects of high data density for both response time and the number of misses. In particular, the number of misses was greater in high data only when there was also poor organization. This highlights the often neglected but important role of poor organization in display clutter and, conversely, suggests that good organization can be used to mitigate the effects of excess data.

The NASA-TLX ratings suggest that the manipulation of stress was successful. However, contrary to Experiment 2, where stress interacted with clutter to increase response time, here there was no clear evidence of performance decrements due to stress. Rather, stress resulted in a significant decrease in response time. In addition, stress led to a more pronounced decrease in response time in the high data condition. Stress also resulted in only a slight and non-significant increase in the number of misses in the high data condition. This suggests that the manipulation of stress in Experiment 3 led to increased motivation only but not a level of stress that could negatively influence performance. This could be attributed to the simple nature of the domain and tasks in this study as compared to Experiment 2, which involved a more realistic scenario, stress led to significant performance decrements.

Eye Tracking Results

Having established that there were significant performance decrements due to clutter, the eye tracking data was then analyzed to capture changes in the underlying attentional processes. We calculated eye movement metrics that belong in the three categories of spread, directness, and duration. Table 4.4 summarizes the main eye tracking results of Experiment 3, displaying the findings together with those of Experiments 1 and 2 for comparison.

Spread metrics. In the case of high data, a larger spread of fixations was observed. In the case of poor organization, the spread metrics increased in similar fashion, with the exception of NNI which showed the opposite effect. Since NNI reflects the distance between the closest fixation points, this result suggests that the grouping of icons led to fixations being farther apart. This suggests that NNI may be a good metric to distinguish between the effects of data and organization. Moreover, among the spread metrics, spatial density was the only one that reflected the same interaction effects of data-organization that were seen for response time. This confirms that it a very useful measure of clutter, which has consistently been shown across all the experiments and in various domains (Moacdieh & Sarter, 2012; Moacdieh, Prinet, & Sarter, 2013; Moacdieh, & Sarter, 2015). Convex hull area was also diagnostic of clutter in all conditions.

Directness metrics. For data density and poor organization, the directness metrics suggested a rather systematic and deliberate search in the high-data conditions, with fixations closely spaced. Mean saccade amplitude in particular appears to be significantly affected by both aspects of clutter and across all three experiments, highlighting it as a very useful metric. There appears to more discrepancy in terms of backtrack rate and rate of transitions, which may need further experiments to better understand their link to clutter. Both metrics decrease with increasing clutter, however, confirming the move to more systematic search in the case of high clutter. It seems that backtrack rate is most strongly reflects data density, whereas rate of transitions is more indicative of poor organization.

Duration metrics. In the case of high data density, the duration metrics indicated that the speed at which participants searched the display was higher in high data density. In other words, participants were quickly moving from one location to the next closest, without spending much time fixating any one location. However, in the case of poor organization, there was higher mean fixation duration , which shows that it took participants more time to extract information from each area, making search even slower. In the case of good organization, they could tell “at a glance” if a target was located within one group whereas, in the poor organization condition, they had to devote more time to processing each icon. Fixation duration is then another metric that can be used to differentiate between the two aspects of clutter.

This study thus showed that eye movement metrics can be used to differentiate between the two main contributors to clutter – data density and poor display organization – and it established which eye movement metrics are best suited for this purpose. Ultimately, in order for these metrics to be used as the basis for an adaptive system, they need to provide the same diagnostic value in real time. The next step towards this goal is described in Chapter 5.

Table 4.4: Summary of the main eye tracking results for the first three experiments. The entries indicate the effects of high data density and poor organization on each of the metrics.

Eye tracking metric	**Visual search**			
	Experiment 1: simple images	**Experiment 2: ED EMRs**	**Experiment 3: simulated graphics**	
	High density	**High density**	**High density**	**Poor organization**
Spread metrics				
Convex hull area	*Increase*	*Increase*	*Increase*	*Increase*
Spatial density	*Increase*	*Increase*	*Increase*	*Increase*
Nearest neighbor index[1]		*Increase*	*Increase*	Decrease
Directness metrics				
Scanpath length per second		Decrease	Decrease	Decrease
Mean saccade amplitude	Decrease	Decrease	Decrease	Decrease
Backtrack rate			Decrease	
Rate of transitions	*Increase*			Decrease
Duration metrics				
Mean fixation duration	*Increase*		Decrease	*Increase*

References

Bravo, M. J., & Farid, H. (2008). A scale invariant measure of clutter. *Journal of Vision, 8*(1):23, 1-9

Doyon-Poulin, P., Robert, J., & Ouellette, B. (2012). Review of visual clutter and its effects on pilot performance: A new look at past research. In *Proceedings of the 2012 IEEE/AIAA 31st Digital Avionics Systems Conference (DASC)* (pp. 2D1-1-11). Williamsburg, VA. IEEE.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631-645.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology,52*, 139-183.

Moacdieh, N. M., Prinet, J. C., & Sarter, N. B. (2013, September). Effects of Modern Primary Flight Display Clutter Evidence from Performance and Eye Tracking Data. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 11-15). SAGE Publications.

Moacdieh, N. M., & Sarter, N. B. (2012, September). Eye movement metrics: A toolbox for assessing the effects of clutter on attention allocation. In*Proceedings of the Human Factors and Ergonomics Society Annual Meeting*(Vol. 56, No. 1, pp. 1366-1370). SAGE Publications.

Moacdieh, N., & Sarter, N. (2015). Clutter in Electronic Medical Records Examining Its Performance and Attentional Costs Using Eye Tracking. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 0018720814564594.

Rosenholtz, R., Li, Y., Mansfield, J., & Jin, Z. (2005). *Feature congestion: a measure of display clutter*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 761-770). Portland, OR: ACM Press.

CHAPTER 5

Development of an Eye Tracking-Based Model to Predict the Effects of Clutter in Real Time

In Chapter 4, eye movement metrics were identified that can discern the effects of high data density and poor display organization on attention and performance. To be able to use these metrics as the basis for an eye tracking-based, adaptive display, it was necessary to next determine whether they can achieve the same goal in real time, i.e., when calculated over a short window of time. If so, these metrics could then be used to build a model that predicts the effects of clutter and, ideally, does so early on during visual search. Display adjustments could then be triggered before performance breakdowns occur. Also, given that some of these eye tracking metrics have been shown to differentiate between the effects of density and organization, the adjustments can potentially be targeted to address the particularly prominent aspect of clutter.

To this end, Experiment 4 again asked participants to search for various painting-related icons in the same displays used in Experiment 3. Data density and organization were varied as before. There were several notable differences between the two experiments, though. In Experiment 4, the eye tracking data was collected and analyzed in real time, and the metrics were calculated over a short, moving window of time. These real-time eye movement metrics then formed the basis of a model to predict the performance effects of clutter. Several binary modeling approaches can be used for this purpose, most notably machine learning techniques such as support vector machines (SVM; e.g., Liang et al., 2007). However, here logistic regression was chosen because of its robustness, relative simplicity, and ability to provide insight into the contribution of each of the metrics (see Chapter 1 for more details). Moreover, in Experiment 4, participants went through the sets of displays twice: the first round was used to collect the training data set and build the model while the second round was used to test the prediction model. Another key difference between Experiments 3 and 4 is the fact that stress was not manipulated in the latter because this factor did not lead to significant performance decrements in Experiment 3. Stress, which is a prominent concern in the emergency department (ED) environment, will be revisited in the last ED experiment (Chapter 6).

In summary, the overall goal of Experiment 4 was to build on the findings from the earlier experiments and predict performance decrements in real time, as opposed to tracing and explaining them after the fact. More specifically, the objectives were to determine whether 1) eye movement metrics calculated over a short window of time could reflect the effects of data density and display organization on attention, 2) these eye movement metrics could be used to predict overall response time, and 3) this prediction could be made early enough in the search process to prompt display adjustments prior to significant performance decrements.

Methods

Participants

The participants in this study were 10 engineering students from the University of Michigan (5 male and 5 female; average age: 23.0 years (*SD* = 3.9)). They all confirmed that they had used Adobe Photoshop ® before. All participants gave informed consent and were compensated $25 for approximately 90 minutes of their time. Participants had self-reported normal or corrected to normal vision; contact lenses were allowed but glasses were not, due to the limitations of the eye tracker. This study was approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board (ID: #HUM00078246).

Experiment Setup

The same images from a simulated graphics program that were used in Experiment 3 were presented here. Also, the same 19-inch monitor (resolution: 1280x1024 pixels) and ASL D-6 desktop-mounted eye tracker (sampling rate: 60 Hz; accuracy: less than 1 degree visual angle; precision: 0.5 degrees visual angle) were used. The participants were seated around 50 cm from the monitor, making for a visual angle of around 25 degrees in the horizontal direction and 20 degrees in the vertical directions. The same lighting conditions were used for all participants. Eye tracking data packets were sent to the experiment computer in real time via TCP/IP over a local network. All of the code to connect to the eye tracker, collect the eye tracking data, and create/test the models was created in Matlab R2011b.

Experiment Design

The independent variables in this study were data density (low, high) and display organization (good, poor). These were defined and manipulated as in Experiment 3. The same displays used in Experiment 3 (including those with multiple targets and absent targets) were used for the first round of search tasks (i.e., during the creation of the training set). For the second round of search tasks (i.e., the testing phase), the same display configurations were used but the locations of the targets were changed. All participants performed the experiment trials in the same fixed, random order and they were told to prioritize accuracy over speed in all cases.

The dependent variables were once again the performance measures of response time and miss rate, as well as the eye tracking data, in this case calculated in real time over a three-second period. This time window was selected as it was less than the average response time in the low data and good organization condition (as seen in Experiment 3) but, at the same time, long enough to allow for the calculation of the eye movement metrics. There was an overlap of two seconds between successive time windows such that new eye tracking measures were generated every second following the first three seconds.

The raw eye tracking data obtained in real time was first used to calculate fixations based on the dispersion algorithm described by Goldberg and Kotval (1999). The minimum time threshold for fixations was set at 100 ms and the minimum distance threshold at 1 degree of visual angle. These parameters were the same as the ones in the ASL software that was used to obtain the metrics in earlier experiments to ensure that the metrics are comparable across studies.

The fixations were then used to calculate three eye movement metrics: scanpath length, mean saccade amplitude, and mean fixation duration. These metrics were found to be good predictors of high data and poor organization in Experiment 3 and were therefore deemed sufficient for this first attempt at real-time clutter detection. Also, limiting the number of metrics was necessary because any additional metrics added a non-negligible lag in the real-time data processing.

Model Creation

Several choices had to be made in the creation of the model. The first consideration was the choice of a target variable that would be used to predict likely performance breakdowns and significant struggles with information search due to clutter. In keeping with the goal of developing a procedure that is display-independent, response time (RT) was selected as the target variable. The idea was not to predict RT with a high degree of accuracy; rather, the goal was to predict whether RT would likely be greater than a pre-set threshold. In this study, this threshold was selected based on the results of Experiment 3. The median RT in that experiment was approximately 7 sec which was chosen here as the cutoff between acceptable ($RT < 7$) or problematic, long ($RT \geq 7$) responses. In the remainder of this document, “long RT” will refer to values above this threshold.

Four models were created and tested. The first model, which will be referred to as the RT model, was meant to predict long RTs as described above. The second model, labeled DataRT, was intended to predict long RTs in the high data density conditions. In other words, this model was going to predict performance decrements in the presence of, and likely because of, high data density. In the same vein, the third model (OrgRT) was expected to predict long RTs in the poor organization condition. Finally, the objective of the fourth model (DataOrgRT) was to predict long RTs in the combined high data and poor organization condition. This last model would thus analyze the RT decrements and underlying attentional effects that may result from the interaction between density and organization.

The second choice was the type of model for making the above predictions. As mentioned earlier, logistic regression was selected for this purpose. Logistic regression models take as input a set of training instances – each of which is a set of predictors – and then generate the likelihood of the target variable belonging to one of two categories, 1 or 0 (here, a target variable of 1 always corresponded to the cases with long RTs). A set of labeled instances called the training set is typically used to generate the model, which is then tested on a different set of instances, the testing set. In this study, each instance corresponded to the set of three eye movement metrics (the predictors) calculated over a given three-second time window. For the RT model, all instances associated with long RTs were labeled as '1' while all others were labeled as 0. For the DataRT model, the high data trials with long RT were labeled as '1'. Similarly, for the OrgRT model, the poor organization instances with long RT were labeled as '1', whereas in the DataOrgRT model, a label of '1' was assigned to instances with high data, poor organization, and long RT. The eye movement metrics were all standardized prior to running the logistic regression models in order to be able to compare all coefficients.

The last decision to make was whether to build one model that incorporates data from all participants or to create one dedicated model for each participant. The former approach implies that the model is more generalizable and avoids overfitting. However, researchers have pointed out how eye movements tend to vary considerably between people (Andrews & Coppola, 2013; Castelhano & Henderson, 2009; Rayner et al., 2007); therefore, one model for each participant would likely result in higher predictive accuracy, an approach adopted with eye tracking before (Jin et al., 2013; Liang et al., 2007). This option was adopted in this study.

Experiment Procedure

After reading and signing the consent form, participants went through the same training procedure described in Experiment 3. The eye tracker was set up and calibrated, and then participants completed the first set of search tasks. Next, they were given a 5-minute break. During that time, the performance and eye tracking data from the first set was used to build a model specific to each participant's eye movements. Participants then completed a second different set of search tasks. This time, the model was used to predict RT in real time. Participants were not notified when a long RT was predicted; rather, this information was stored for later analysis only. This was necessary to be able to record what the actual RT was and then determine if the model accurately predicted this result.

Results

Significance was set at $p < 0.05$. In the description of results, LG, LP, HG, and HP refer to the various (L)ow/(H)igh data density and (G)ood/(P)oor organization conditions. All error bars in the figures show the standard error of the mean. Note that one participants' data had to be discarded because of a misunderstanding of the instructions, resulting in data for a total of 9 participants.

Performance Results

RT. RT and miss rates were calculated for all trials. A repeated-measures analysis of variance (ANOVA) with listwise deletion was used. Bonferroni corrections were applied for multiple comparisons. The values largely mirror the ones obtained in Experiment 3 (see Figure

5.1), with response times of 5.13 (*SD* = 1.74), 4.97 (*SD* = 1.50), 7.20 (*SD* = 2.00), and 13.06 (*SD* = 4.60) seconds in the LG, LP, HG, and HP conditions, respectively. There was a significant increase in response time from 5.05 (*SD* = 1.71) in low data to 10.13 (*SD* = 4.75) seconds in high data ($F(1, 8) = 29$, $p = 0.001$, $\eta_p^2 = 0.7$). Similarly, there was a significant increase from 6.17 (*SD* = 2.23) seconds in good organization to 9.02 (*SD* = 5.40) seconds in poor organization ($F(1, 8) = 9.8$, $p = 0.014$, $\eta_p^2 = 0.5$). There was also a significant interaction between data density and organization ($F(1, 8) = 21.14$, $p = 0.002$, $\eta_p^2 = 0.725$), with longer RT in the high data density and poor organization conditions. Moreover, there was no significant effect of data density on RT in the good organization condition, whereas there was a significant increase in the poor organization condition ($F(1, 8) = 39.28$, $p < 0.001$, $\eta_p^2 = 0.831$).

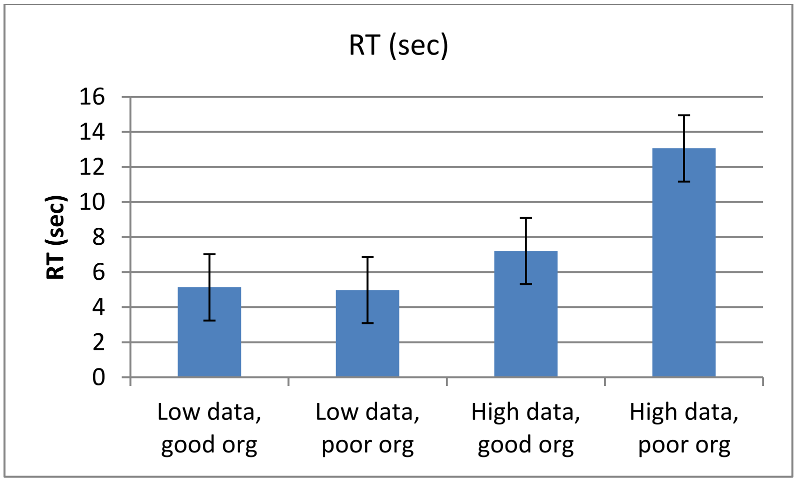
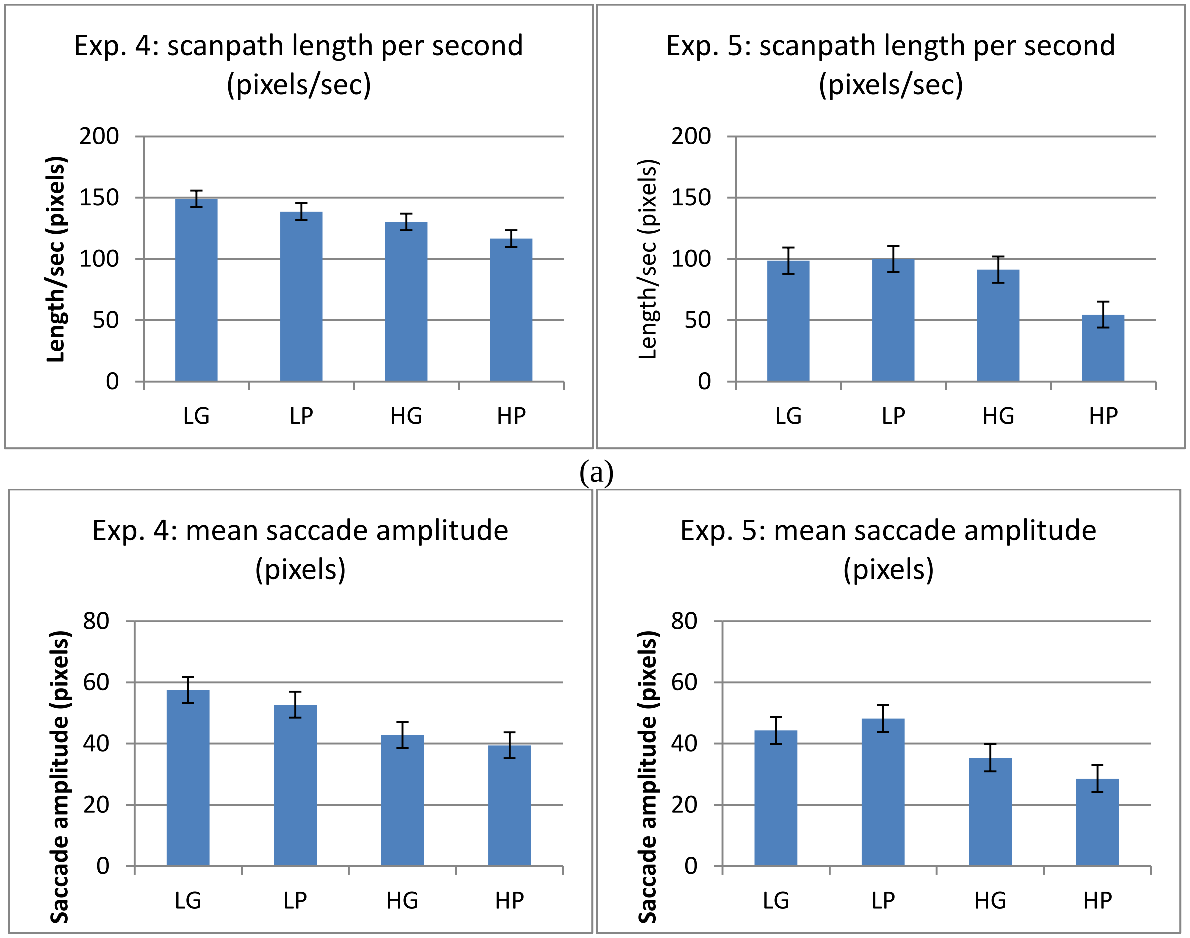


Figure 5.1: Values of RT in the four different conditions

Miss rate. There were very few misses overall in this experiment. Across all participants, there was a total of 7 misses, with 2 in the LG condition, 1 in LP, 1 in HG, and 3 in HP. Two participants were responsible for 6 of these misses (one participant had 4 misses and another had 2). No statistical analysis was performed for these results.

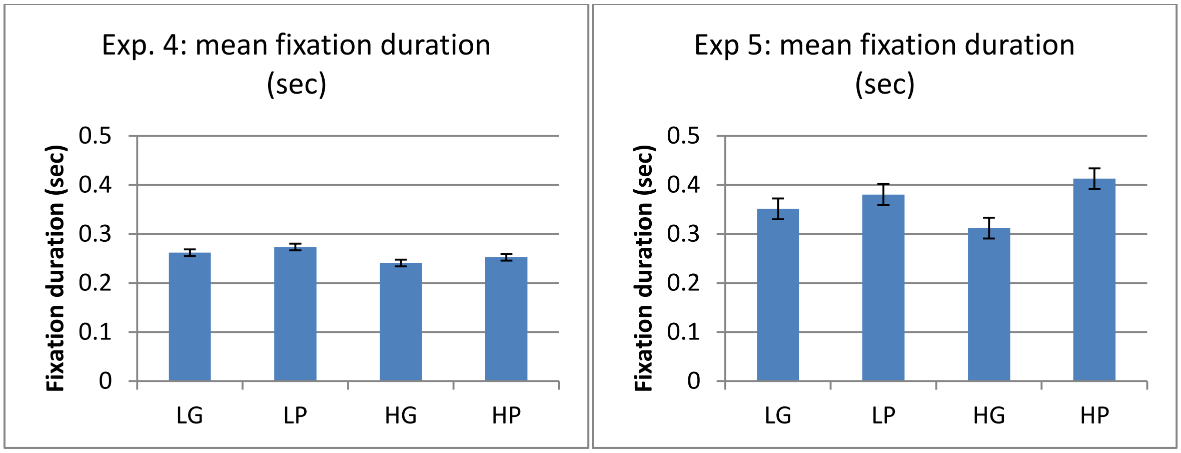
Eye tracking results

Descriptive results. The descriptive eye tracking results help obtain an initial understanding of the real-time effects of clutter on attention. The mean scanpath length per second, mean saccade amplitude, and mean fixation duration, all calculated over a three-second window, can be seen in Figure 5.2. For reference, these are plotted together with the results from Experiment 3 where the same metrics were calculated for the duration of each entire search task.



(a)

(b)



(c)

Figure 5.2: Mean values of (a) scanpath length per second, (b) mean saccade amplitude, and (c) mean fixation duration for both Experiments 3 and 4 (LG: low data, good organization; LP: low data, poor organization; HG: high data, good organization; HP: high data, poor organization)

The changes in the three metrics for the various display conditions mirror the patterns observed in Experiment 3a. Both scanpath length per second and mean saccade amplitude tended to be lowest in the high data density and poor organization condition, whereas mean fixation duration tended to be higher in the case of poor organization. The unequal number of instances in each of the conditions meant that formal analysis of variance was not possible here. Nevertheless, the descriptive measures provided a preview of the expected model results.

Model results. Table 5.1 shows the parameters that were used to evaluate the four models. In addition to overall accuracy, other values of interest included the true positive rate (TPR; the sensitivity or hit rate), TNR (true negative or correct rejection rate, also known as specificity), PPV (positive predictive value; the proportion of all positively-labeled cases that are true positives or hits), and NPV (negative predictive value; proportion of negatively-labeled cases that are true negatives or correct rejections). Also, *d'* and *β* were calculated where d' (sensitivity) represents the ability of the model to detect longer RTs (*d'* is 0 if the probability of detecting RT is 50%), while the response bias β indicates whether the model is liberal (more likely to overestimate RT; $\beta < 1$), or conservative (more likely to underestimate RT; $\beta > 1$).

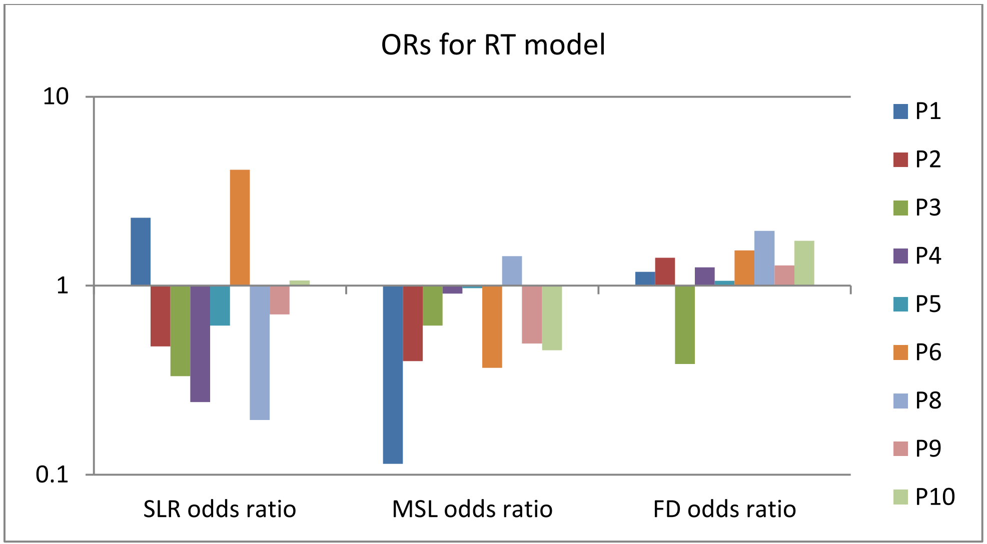
Table 5.1: Summary of logistic regression model results

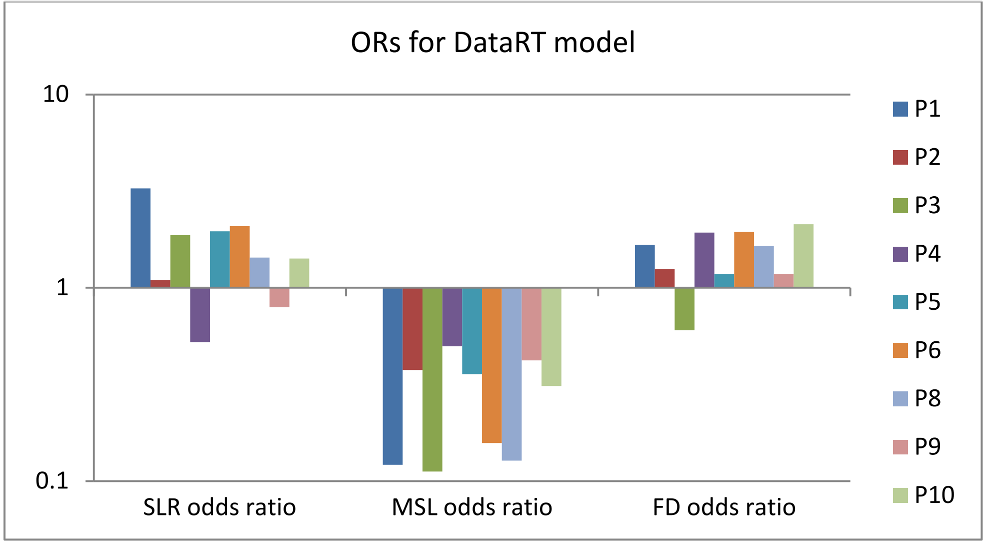
Average (SD)	**Accuracy**	**TPR**	**TNR**	**PPV**	**NPV**	**d'**	**β**
RT	0.77 *(5.74)*	0.90 *(0.14)*	0.67 *(0.16)*	0.59 *(0.12)*	0.95 *(0.04)*	1.60 *(0.44)*	0.95 *(0.72)*
DataRT	0.71 *(10.3)*	0.96 *(0.07)*	0.59 *(0.18)*	0.5 *(0.1)*	0.98 *(0.03)*	1.38 *(0.54)*	0.77 *(0.43)*
OrgRT	0.79 *(7.9)*	0.97 *(0.04)*	0.71 *(0.11)*	0.56 *(0.13)*	0.98 *(0.02)*	1.72 *(0.57)*	0.86 *(0.31)*
DataOrgRT	0.75 *(9.4)*	0.95 *(0.07)*	0.68 *(0.12)*	0.47 *(0.12)*	0.98 *(0.02)*	1.64 *(0.57)*	0.85 *(0.46)*

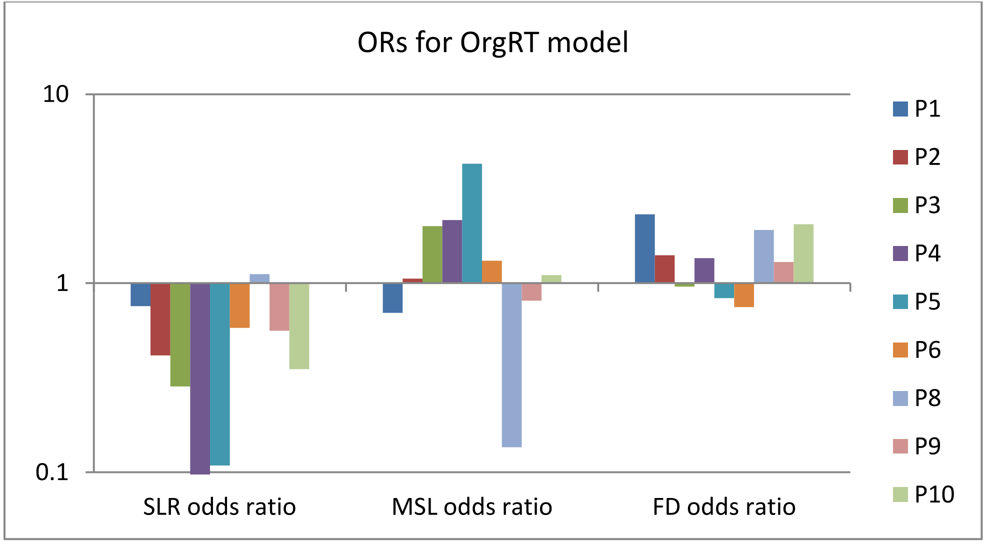
The average accuracy for all of the models was greater than 0.7, with average values of 0.77, 0.71, 0.79, and 0.75 for the RT, DataRT, OrgRT, and DataOrgRT models, respectively. A one-sample t-test confirmed that all of these average values are significantly different from chance, which is defined as accuracy of 0.5 (RT model: $t(8) = 13.2$, $p < 0.001$; DataRT: $t(8) = 5.7$, $p < 0.001$; OrgRT: $t(8) = 10.2$, $p < 0.001$; DataOrgRT: $t(8) = 7.3$, $p < 0.001$). The TPR for all four models was relatively high for all models, at 0.9 (*SD* = 0.15), 0.96 (*SD* = 0.074), 0.97 *SD* = 0.048), and 0.95 (*SD* = 0.075) for RT, DataRT, OrgRT, and DataOrgRT, respectively. The TNR, on the other hand, was found to be 0.67 (*SD* = 0.16), 0.59 (*SD* = 0.18), 0.71 (*SD* = 0.11), and 0.68 (*SD* = 0.12) for RT, DataRT, OrgRT, and DataOrgRT, respectively. It followed that the PPV values were 0.59 (*SD* = 0.12), 0.5 (*SD* = 0.10), 0.56 (*SD* = 0.13), and 0.47 (*SD* = 0.12). The NPV were 0.95 (*SD* = 0.049), 0.98 (*SD* = 0.038), 0.98 (*SD* = 0.02), and 0.98 (*SD* = 0.029) for

RT, DataRT, OrgRT, and DataOrgRT, respectively. The d’ values were 1.6 (*SD* = 0.44), 1.38 (*SD* = 0.55), 1.72 (*SD* = 0.57), and 1.64 (*SD* = 0.57) for RT, DataRT, OrgRT, and DataOrgRT, respectively. These values were all found to be significantly greater than chance, which is defined as d’ = 0 (RT model: $t(8) = 10.1$, $p < 0.001$; DataRT: $t(8) = 7.1$, $p < 0.001$; OrgRT: $t(8) = 8.4$, $p < 0.001$; DataOrgRT: $t(8) = 8$, $p < 0.001$).

For each model, the odds ratio (OR; i.e., the exponential of the log ratio or model coefficients) was calculated as well. Odds ratios greater than 1 indicate that an increase in that factor will increase the likelihood of the target variable being equal to 1; the larger the OR, the larger the odds of having the target variable equal to 1. Conversely, odds ratios smaller than 1 signify a decreased likelihood of having the target variable equal to 1. Figure 5.3 shows the ORs for the different models created and Table 5.2 shows the means of the values. For the scanpath length per second (SLR), the mean ORs were less than 1 for OrgRT and DataOrgRT, but greater than 1 for RT and DataRT. Looking at the graphs, there is a strong tendency for the ORs to be less than 1, except for DataRT. For mean saccade amplitude (MSL), the ORs were for the large part less than 1 for all models, except for OrgRT. As for mean fixation duration, the mean ORs were all greater than 1, although for DataOrgRT in particular there was some variation in the ORs.


ORs for RT model
10
1
0.1
SLR odds ratio
MSL odds ratio
FD odds ratio
P1
P2
P3
P4
P5
P6
P8
P9
P10


ORs for DataRT model
10
1
0.1
SLR odds ratio
MSL odds ratio
FD odds ratio
P1
P2
P3
P4
P5
P6
P8
P9
P10


ORs for OrgRT model
10
1
0.1
SLR odds ratio
MSL odds ratio
FD odds ratio
P1
P2
P3
P4
P5
P6
P8
P9
P10

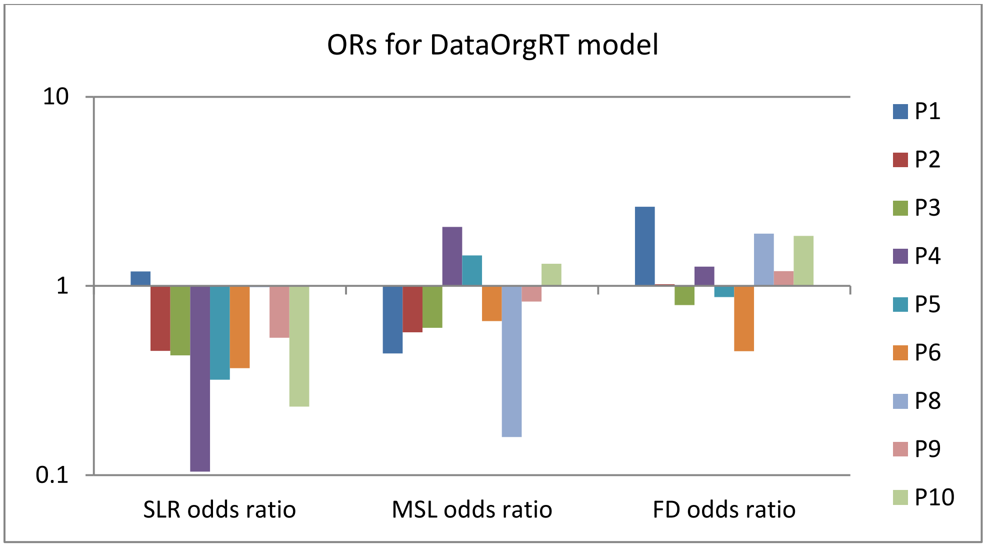


Figure 5.3: Odds ratios (ORs) for the different models

Table 5.2: Summary of model odds ratios

Odds ratios Means *(SD)*	**Scanpath length per second**	**Mean saccade amplitude**	**Mean fixation duration**
RT	1.11 *(1.22)*	0.64 *(0.37)*	1.31 *(0.41)*
DataRT	1.60 *(0.76)*	0.28 *(0.27)*	1.5 *(0.45)*
OrgRT	0.47 *(0.30)*	1.51 *(1.14)*	1.43 *(0.52)*
DataOrgRT	0.51 *(0.33)*	0.9 *(0.55)*	1.33 *(0.63)*

Finally, the average time at which the model correctly predicted a long RT was calculated. These average times were found to be 3.84, 3.65, 4.09, and 4.01 seconds for the RT, DataRT, OrgRT, and DataOrgRT models, respectively. These times were all found to be significantly less than the average response time of 7 seconds (RT model: $t(8) = -24.1$, $p < 0.001$; DataRT: $t(8) = -20.6$, $p < 0.001$; OrgRT: $t(8) = -12.6$, $p < 0.001$; DataOrgRT: $t(8) = -11.7$, $p < 0.001$).

Discussion and Conclusion

This experiment had three objectives. The first goal was to determine whether three eye movement metrics would capture the previously observed effects of data density and display organization on attention when these metrics were calculated over a short three-second time window. The average values for the metrics in this study indicate that, for the most part, this is the case (see Table 5.3). A few discrepancies were observed which are likely due to one of two factors. First, in Experiment 3, fixations were calculated using the dedicated ASL software; however, since this software does not allow for real time analysis of eye tracking data, the fixations in Experiment 4 were analyzed using Matlab code that was developed specifically for this purpose. Although the main ASL fixation algorithm was replicated in Matlab, the ASL software provided additional data pre-processing, such as the removal of blinks. These steps were not performed in the Matlab code, which may have contributed to some discrepancies in the values of the metrics. In addition, recall that the values obtained in Experiment 3 included the effects of stress. One of these effects was a slight decrease in fixation duration which may also have contributed to the observed differences.

In addition to the descriptive results, the model odds ratios also provide some insight into the effects of data density and display organization on performance and attention. In all cases, the mean fixation duration odds ratio was greater than 1, suggesting a positive association between longer fixation duration and larger RT in all conditions. The DataOrgRT model reflected the noticeably low scanpath length per second and mean saccade amplitude in the high data density and poor organization condition, as observed in the descriptive results. Both the mean OR coefficients for scanpath length per second (0.51) and mean saccade amplitude (0.9) were less than 1 for this model, indicating an inverse relationship with RT in the case of high data and poor organization. The association was particularly strong for scanpath length per second. In other words, in the high data density and poor organization conditions, participants who took long to respond were more likely to have had a slower rate of search and higher fixation duration. Moreover, based on the DataRT and OrgRT models, it seemed that the low saccade amplitude rate (i.e., slow search) was associated with long RT in the poor organization condition, whereas low mean saccade amplitude (closely-spaced fixations) was associated with long RT in high data density. However, more research and more modeling results will be needed to fully understand these effects.

The second goal of this study was to determine whether the three eye movement metrics could be used to predict long RT. Of the four models, OrgRT produced the highest accuracy at 79%, followed by RT (77%) and DataOrgRT (75%). The DataRT model produced the worst accuracy at 71%. Note that organization, as well as organization combined with data density, could be predicted more accurately than data density alone. While overall, accuracy of the models will need to be improved further, which may be achieved by adding more eye movement metrics, it is important to point out that the TPR was greater than 90% for all models, reaching 97% for OrgRT. This means that very few cases with long RT were missed (i.e., very few false negatives). Instead, the challenges were false alarms in which RT was acceptable but were classified as long RT. This is confirmed by the values of β, which were all less than 1, suggesting liberal models that favor false alarms over misses. Still, making sure that no long RT conditions are missed is arguably more critical in real-life applications. The d’ value for all models was greater than 1, with the value for OrgRT largest at 1.72, confirming that it was the best model.

The third and final goal of this study was to determine whether the predictions from these models could be obtained early enough to trigger useful adaptations. This was confirmed as the average prediction time was significantly lower than the average RT in all four conditions.

Table 5.3: Summary of the effect of high data density and poor organization on the metrics.

Eye tracking metric	**Visual search**					
	Exp. 1: simple images	**Exp. 2: ED EMRs**	**Exp. 3: simulated graphics**		**Exp. 4: real-time metrics**	
	High density	**High density**	**High density**	**Poor organization**	**High density**	**Poor org.**
Spread metrics						
Convex hull area (pixels^2)	*Increase*	*Increase*	*Increase*	*Increase*		
Spatial density	*Increase*	*Increase*	*Increase*	*Increase*		
Nearest neighbor index[1]		*Increase*	*Increase*	Decrease		
Directness metrics						
Scanpath length per second (pixels/sec)		Decrease	Decrease		Decrease*	Decrease*
Mean saccade amplitude (pixels)	Decrease	Decrease	Decrease	Decrease	Decrease*	Decrease*
Backtrack rate (/sec)			Decrease			
Rate of transitions (/sec)	*Increase*			Decrease		
Duration metrics						
Mean fixation duration (sec)	*Increase*		Decrease	*Increase*	*Increase**	*Increase**

*These are based on descriptive and not statistical results

In summary, the results of Experiment 5 show that a real-time model that can predict the effects of clutter is feasible. This initial experiment relied on three metrics only: scanpath length per second, mean saccade amplitude, and mean fixation duration. These metrics, calculated over a moving 3-second time window, largely reflected the effects observed earlier over the whole RT period. The four tested models, which were trained to predict RT in different conditions, were able to achieve accuracy rates of 70 – 80%. These rates are significantly greater than chance but need to be improved further. To this end, future research should explore the addition of more eye movement metrics to the models. The high (greater than 90%) TPR and early detection ability (around 4 seconds) were both very promising, however. The final step performed in this line of research was to test these models, together with real-time display adjustments, in a very different type of domain, the emergency department (ED). This step was taken in Experiment 5.

References

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision research*, *39*(17), 2947-2953.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 6.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631-645.

Jin, L., Niu, Q., Jiang, Y., Xian, H., Qin, Y., & Xu, M. (2013). Driver sleepiness detection system based on eye movements variables. *Advances in Mechanical Engineering*, *2013*.

Liang, Y., Reyes, M. L., & Lee, J. D. (2007). Real-time detection of driver cognitive distraction using support vector machines. *Intelligent Transportation Systems, IEEE Transactions on*, *8*(2), 340-350.

Rayner, K., Pollatsek, A., Drieghe, D., Slattery, T.J., & Reichle, E.D. (2007). Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, and Engbert (2006). Journal of Experimental Psychology: General, 134

CHAPTER 6

Exploring the Feasibility and Effectiveness of Real-Time, Eye Tracking-Based Models and Display Adjustments In EMRs

The findings from the previous experiments (described in Chapters 4 and 5) suggest that the use of eye tracking-based models for predicting the effects of display clutter on attention and performance is feasible and promising. However, those earlier studies were conducted in a highly-controlled environment involving pure target search tasks, and no display changes were triggered to prevent or overcome performance decrements due to clutter. The next and final step in this line of research was thus to 1) test the accuracy and predictive value of the developed models in a more complex data-rich environment with more demanding tasks, and 2) examine whether display adjustments that are triggered in real time, in response to observed clutter effects, area acceptable to users and significantly improve performance.

These two questions were addressed in Experiment 5, which was conducted in the context of electronic medical records (EMRs) in a hospital emergency department (ED). Similar to the earlier studies that led up to this final experiment, clutter and stress were manipulated while participants performed search and noticing tasks. A more complex and open-ended search task was used here as a test of whether the model would work in a task other than pure target search. Based on Experiment 2, it was argued that a longer, more difficult task would also be more affected by clutter. Moreover, a critical additional feature of Experiment 5 was the introduction of display adjustments aimed at counteracting the effects of clutter. Note that the focus here was not on developing optimal decluttering techniques for ED EMRs. Rather, fairly simple display adjustments were chosen and implemented in an effort to explore the feasibility and acceptability of the overall approach of real-time, eye tracking-based display adaptations.

Past studies that have investigated adaptive decluttering techniques typically provided the display adjustments based on some display algorithm (e.g., St. John, Smallman, Manes, Feher, & Morrison, 2005). In contrast, in the current research, display adaptations were triggered in real time, if and when eye tracking data showed early signs of clutter effects. This approach was adopted to avoid distracting participants with unnecessary changes to the display if they were not needed. Moreover, this lag before the display adjustments appear is necessary in order to capture the interaction between the display-based aspects of clutter and other top-down factors that could influence the effects of clutter on performance. The experience of the user, the level of stress, and the current task are just some examples of top-down factors that can interact with the quantity and organization of data. This approach is in keeping with the definition of clutter described in Chapter 1, where the emphasis is on the performance decrements associated with clutter. In other words, it would be impossible to establish one display that works for all users and situations; rather, by focusing on the detecting early signs of performance decrements, the display can be tailored to suit the needs of the current user. This may prove to be particularly valuable in the case of displays used by multiple people or that can be modified.

In summary, the specific research questions for this final study were:

1. Will the appearance of display adjustments a few seconds into a search task help participants or cause distraction, as reflected by performance and subjective measures?

2. Can the modeling approach presented in Chapter 5 accurately predict the effects of clutter in a more complex, stressful, and data-rich environment?

Like Experiment 4, Experiment 5 also consisted of two parts. The first part was a training phase where physicians' eye movements were recorded while they performed their search tasks. This data was used to create the clutter prediction models. In the second part, these models were used to predict, in real time, whether physicians would likely need a long time to complete their search tasks. If so, display adjustments were triggered to assess their benefits and potential costs to attention management and task performance.

Methods

Participants

The participants in this experiment were 12 medical practitioners (five female and seven male; six residents and six physician assistants) who were employed in the University of Michigan Department of Emergency Medicine. Their average age was 32.6 years (standard deviation (*SD)* = 4.6), and their average years of experience in this ED was 4 years (*SD* = 2.2). All participants had been trained to use the EMR currently employed in the ED, and their average years of experience with this EMR was 2 years (*SD* = 0.90). Participants had self-reported normal or corrected to normal vision; contact lenses were allowed but glasses were not, due to the limitations of the eye tracker. For reasons that will be discussed later, participants were divided into two groups, Group 1 and Group 2. Each group contained three residents and three physician assistants. There was no significant difference between the two groups in terms of age, level of experience, or proficiency with the EMR. All participants gave informed consent and received compensation for their time. This study was approved by the University of Michigan Health Sciences and Behavioral Sciences Institutional Review Board (ID: #HUM00078246).

Simulated EMR

The EMR used in this study extended the one that was used for Experiment 2 (see Chapter 3). Previously, three pages of the currently-used EMR were replicated: the Medical History page, Laboratory Results page, and Patient Timeline page. Results indicated that the Medical History page, which displays an unordered list of patients' medical, surgical, social, and family history, showed the most significant effects of clutter. As a result, the Medical History page was chosen to be the focus of this final experiment. It was the page on which participants had to perform their main search task, and based on which the eye tracking-based model would predict their response time. This was also the only page on which clutter (in this case, data density) was systematically manipulated and on which display adjustments were tested.

In order to create more realistic scenarios and also to investigate whether participants would notice important patient data across fragmented EMR information, two additional EMR pages were replicated using XAML and C#: Chart Review and Patient Summary. These are the two main pages that physicians use to extract information about a patient (as opposed to those related to documentation and order entry). Chart Review is the area that stores previous physician and nurse notes which are related to earlier ED encounters, office visits, or surgical procedures. The Patient Summary page stores all of a patient's medical information, including previous medical/surgical history, current medications, allergies, ordered medications, labs, images, and current vital signs taken at triage. Both Chart Review and Patient Summary contain a number of tabs that separated information into different areas; the full details of what each tab contained can be seen in Appendix F. Finally, the Lab Results page that had been implemented for Experiment 2 was used here also. This page displays a table of current and previous laboratory tests. As before, copyright restrictions mean that no screenshots can be displayed here.

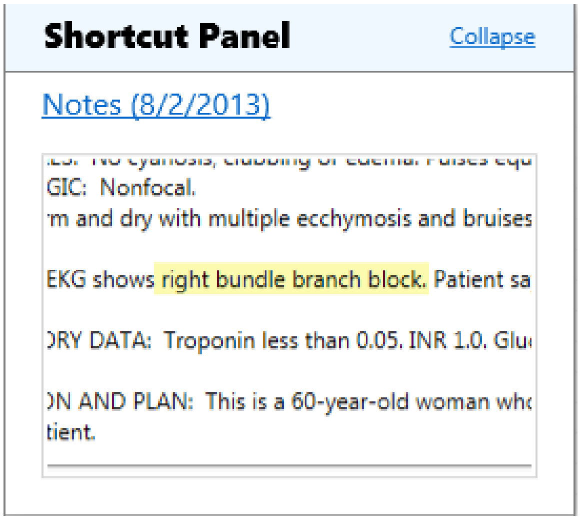
All EMR pages were filled with fictitious but realistic patient information to create 16 medical scenarios. We made sure that all of the information in each patient's EMR was consistent and medically sound. For example, the listed medical history matched the patient's previous physician notes, which were also consistent with the patient's medications. Depending on the particular patient scenario, laboratory or imaging results could also be available. Some of the scenario details were based on material taken from the Harvard Medical School Simulation Casebook (Howard, Siegelman, Guterman, Hayden, & Gordon, 2011). All of the scenarios were reviewed and approved by the collaborating physicians at the University of Michigan Hospital.

Display Adjustments

Two types of display adjustments were implemented and triggered in response to predicted long RTs:

- *Increasing the salience of the target*: the first adjustment consisted of highlighting the search target on the display (the background color of the particular entry in the Medical History page turned yellow; for a similar approach in a different domain, see Yeh and Wickens (2001)). The inherent structure of the display was not changed.

- *Improving the organization of the display*: the second, adjustment was the appearance of a shortcut panel on the right side of the Medical History page (see Figure 6.1 for an example of one of the panels). These panels showed small (250x175 pixels) screenshots of pages/areas of the EMR that were considered useful for the particular scenario. The most relevant information on that page was highlighted, and a link was provided that, when clicked, would take the user to the respective EMR page/area. The panel was designed such that it appeared in a blank space on the screen and thus did not cover any EMR data.



ES: No cyanosis, clubbing or edema. Pulses equ
GIC: Nonfocal.
m and dry with multiple ecchymosis and bruises

EKG shows right bundle branch block. Patient sa

ORY DATA: Troponin less than 0.05. INR 1.0. Glu

ON AND PLAN: This is a 60-year-old woman wh
tient.

Figure 6.1: Example of one of the panels used in the experiment, which links to a physician note from 8/2/2013. The section of the note that seems most relevant to this particular case is highlighted.

For both types of adjustment, it was assumed that a natural language processing (NLP) algorithm could be developed to identify, based on available EMR data, what information might be useful and should be brought to the attention of the physician in order to perform a medical diagnosis. The development of such algorithms has been the focus of considerable research in the medical and computer science domains (e.g., Biese et al., 2013; Jonnalagadda et al., 2013) but was beyond the scope of this dissertation. To simulate the presence of such an NLP algorithm, the correct information (based on discussions with physicians) for all the highlighting and panel adjustments was hardcoded into the scenarios.

Eye tracking-based Model

The same approach to modeling response time (RT) and data-response time (DataRT) that was described for Experiment 4 was used here (see Chapter 5). Recall that the RT model serves to predict long RT, whereas the DataRT model aims to predict long RT in the high data conditions. OrgRT and DataOrgRT were not used in this experiment since organization was not systematically varied. The reason it was not varied was that such a change would involve asking physicians to work with an unfamiliar display, creating a confound. Based, in part, on the results from Experiment 2, the threshold used to separate long from acceptable RTs was set at 25 seconds. This was the median value of noticing times. The three-second time window for calculating eye movement metrics was used once again, and the same Matlab code was used for this experiment.

In the training phase, participants completed a search task on a total of 10 screenshots of the EMR Medical History page. Six of these images were low in clutter, and 4 were high in clutter (slightly more low-clutter images were used since these provide fewer instances). For each image, participants were asked to search for risk factors for a particular medical condition. Thus, participants had to perform this more involved search task that required searching through the entire page and performing judgments. As mentioned earlier, this was a deliberate choice to test the models in a different type of task, largely based on the previous finding that search performance for multiple targets is more strongly affected by clutter. As in Experiment 4, stress was not manipulated in the training phase. As in Experiment 5, the search times from the training phase were divided into three-second windows, which formed the instances of the RT and DataRT logistic regression models.

The results of the training phase were then used in the testing phase, where the two models were employed to predict RTs in the 16 patient scenarios. The predictions were stored but not used to trigger display adjustments. Instead, the display adjustments were always triggered by the experimenter three seconds into the scenario. This served to address the question of the acceptability and the performance effects of adjustments that are presented early on, independent of model predictions. In other words, it was necessary to be able to test the costs and benefits of early real-time display adjustments regardless of whether the model predicted clutter early on or not.

Experiment Setup

The EMR pages were displayed on a 19-inch monitor with a resolution of 1280x1024 pixels. The same ASL D-6 desktop-mounted eye tracker was used as in the previous experiments (sampling rate: 60 Hz; accuracy: less than 1 degree visual angle; precision: 0.5 degrees visual angle). The eye tracker was placed directly in front of the computer monitor such that the aperture was around 5 cm from the screen. The participants were seated at a distance of around 50 cm from the monitor, making for a visual angle of around 25 degrees in the horizontal direction and 20 degrees in the vertical directions. The lighting conditions were held constant for all participants. Calibration took place at the start of the experiment using a nine-point grid. The duration of the calibration procedure varied across participants but generally took around five minutes. To the left of the participant's monitor, another screen was installed to enable the experimenter to observe the output of the model, trigger the display adaptations, and keep track of the eye tracking data in real time.

Experiment Design

The independent variables were clutter (low, high), stress level (no stress, stress), and type of display adaptation (highlighting, panel). Clutter and stress were manipulated as in Experiment 2 (please refer to Chapter 3). The only difference was that participants in this experiment had a 45-second time limit in the stress condition, as compared to a 20-second time limit in Experiment 2. This change appeared necessary since, as will be detailed later, participants' tasks in this study involved searching through several pages of the EMR (as opposed to a single, static page in Experiment 3). Clutter and stress were varied within participants, whereas the type of adaptation was varied between two groups of participants, Group 1 (highlighting) and Group 2 (panel).

The result was a 2x2x2 fractional factorial design, where each participant went through eight low-clutter and eight high-clutter scenarios, for a total of 16 scenarios. Half of the scenarios were no-stress, while the other half were stress scenarios (the manipulation of clutter and stress was fully factorial). Of the eight high-clutter displays, four were subject to adaptation. The scenarios that involved adaptation for one group did not include any adaptation for the second group. Figure 6.2 illustrates the design of the groups and highlights the between-subjects comparisons.

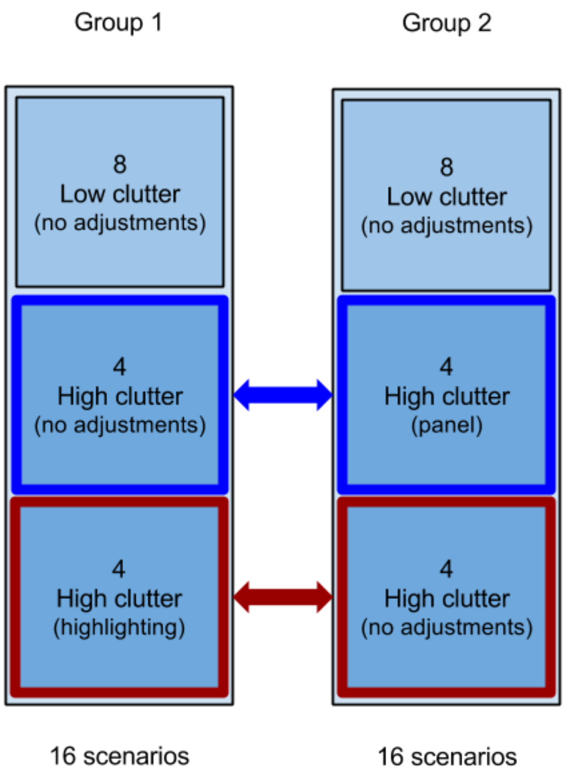


Figure 6.2: Breakdown of clutter for the 16 scenarios and the adjustments in each of the two groups. Half of each set of scenarios are no-stress and the other half are stress scenarios.

Similar to the procedure described in detail in Chapter 3, each variable combination or experiment trial translated to a different patient scenario. For each scenario, participants were given (1) a summary of the patient's symptoms, similar to what a physician would receive in a triage note, (2) a search task (for information that participants had to locate in the medical history page to help them make their diagnosis), (3) two noticing targets (information that participants needed to detect on their own to make the correct diagnosis), and (4) a suggested diagnosis. This diagnosis was presented as an initial assessment by a medical student; participants then needed to look for evidence to support or disprove this suggestion.

To guide their decision process, participants were told to first answer the given search question using the Medical History page. All search questions required participants to review the display and identify multiple targets that matched the search criterion. Participants then continued to review the EMR pages to look for any other relevant information. In each case, they had to notice two pieces of information, one on the Medical History page (Noticing 1) and a second that could be shown anywhere in the other pages of the EMR (Noticing 2). Both of the noticing tasks were designed to be highly relevant to the diagnosis. The details of two sample scenarios can be seen in Table 6.1; the full scenario details can be found in Appendix H. Each pair of scenarios (e.g., 1a and 1b) was equivalent in terms of the locations of the two noticing targets. For Noticing 1 targets, they were both in the same section of the Medical History page, whereas for the Noticing 2 targets, they were both in the same area (e.g., notes or Laboratory Results). Note that the highlighting was thus meant to help with the Noticing 1 tasks, whereas the shortcut panel targeted the Noticing 2 tasks.

Finally, the dependent measures were the search time (i.e., time to complete the search task), Noticing 1 time (to complete the Noticing 1 task), Search miss rate, Noticing 1 miss rate, and Noticing 2 miss rate. Note that the response time for the Noticing 2 tasks was not considered because the amount of data in the EMR outside of the Medical History page was not held constant across scenarios (it depended on the particular patient case). All times were obtained by listening to the recording of participants, who were instructed to give all answers verbally and think aloud throughout each trial.

Table 6.1: Two sample scenarios from the ones used in Experiment 4. The full scenario details can be seen in Appendix G.

Trial	Pre-trial information presented to participants			Correct answers [locations in the EMR]		
	Patient triage note	**Suggested diagnosis**	**Search question**	**Noticing task 1**	**Noticing task 2**	**Correct diagnosis**
Low clutter	Mark Lawrence is a 46-year old male. He presents to the ED with dizziness and loss of balance.	Acute anemia	Does the patient have a history of blood disorders?	Recent pharyngitis [medical history]	History of pseudo-anemia [notes]	Labyrinthitis
High clutter *(Panel)*	Katherine Willis is a 60-year old female. She presents to the ED with chest pain, cough, and difficulty breathing. The EKG is notable for a right bundle branch block.	Myocardial infarction	Does the patient have a history of heart disease?	Chronic obstructive pulmonary disease [medical history]	Patient has always had a right bundle branch block [notes]	Pneumonia

Experiment Procedure

After reading and signing the consent form, participants were given instructions for the first part of the experiment, the training phase. The eye tracker was calibrated, and participants completed search tasks for the 10 displays of the training phase. They were instructed to prioritize accuracy over speed and to think aloud throughout each trial. Their answers were recorded by the experimenter. Once they completed the training phase, they were given a short break while the model was set up. Next, they were given instructions for the main part of the experiment, which involved 16 scenarios representing various ED cases. They were told that they would need to complete 16 scenarios that represented various ED cases.

For each scenario, participants were provided with an introductory page that would show 1) a triage note describing the symptoms of the patient, 2) a diagnosis that had been suggested by a medical student, and 3) a specific question to answer using only the medical history page. The introductory page would also indicate whether participants had unlimited time or 45 seconds to complete that scenario. The presentation order of the scenarios was counterbalanced by dividing both Groups 1 and 2 into two groups, each of which performed the experiment in a fixed random order. The order of the second group was the inverse of the first. Participants were told that their goal was to determine whether the medical student's diagnosis made sense and, if not, what might be a more likely diagnosis. Participants were also informed about the two types of display adjustment: highlighting (Group 1) and panel (Group 2). They were told that these adjustments would occasionally appear to point out potentially relevant information and that it was up to them to decide whether to make use of the information.

When participants were ready, they pressed a button to replace the question screen with the corresponding EMR of that patient. They performed their search and noticing tasks and then pressed the spacebar to display the final question page. The same question was always presented here: "Based on what you have seen about the patient, do you think that [the medical student's diagnosis] is likely? Please explain why or why not". If participants mentioned any information related to the noticing task here but had failed to do so while performing the task, this was recorded and not considered a miss; however, no response time was stored. Participants completed all 16 trials with no breaks in between, for a total time of around 30 minutes. After completing all trials, participants filled out a debriefing questionnaire in which they entered NASA-Task Load Index (TLX; Hart & Staveland, 1988) ratings for the low- and high-stress conditions and also provided their feedback on the display adjustments (highlighting for Group 1 and the panel for Group 2). The full debriefing questionnaire can be seen in Appendix D.

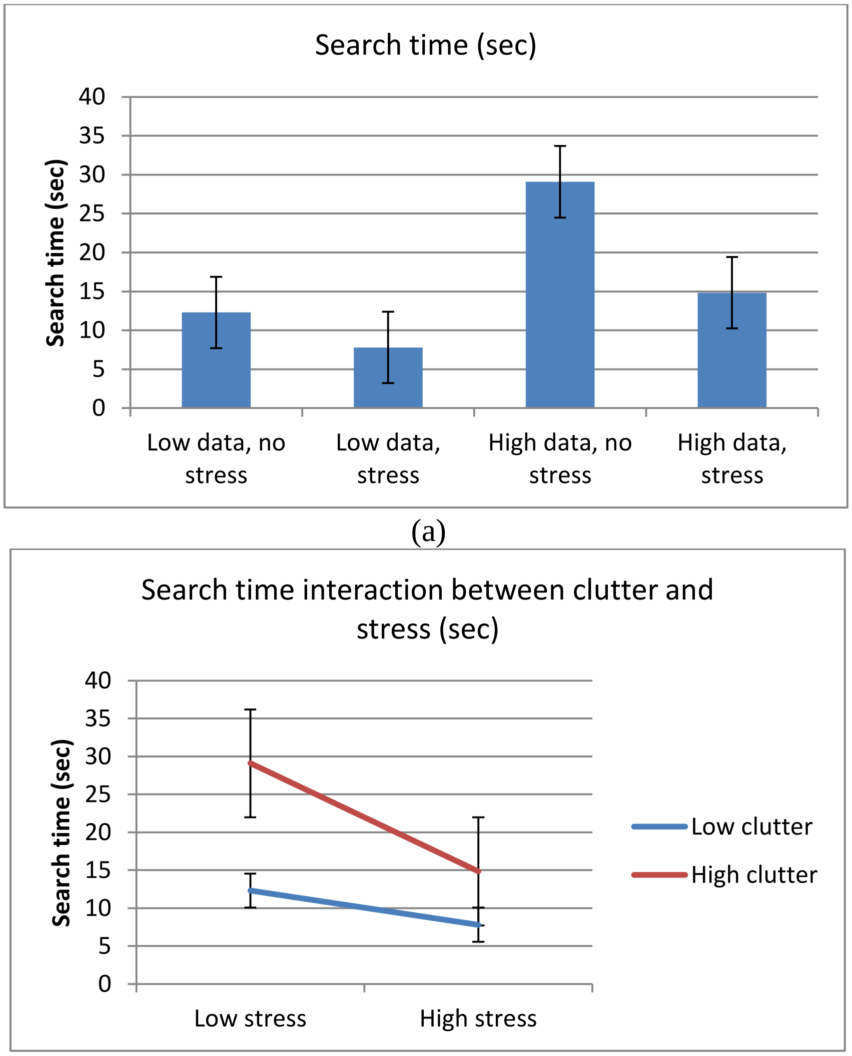
Results

For the analysis of the performance and eye tracking data, the significance level was set at p < 0.05. Non-significant statistical results are not reported. The eye tracking data of three participants had to be discarded because, for various reasons, they could not be properly seated in front of the eye tracker. Error bars on graphs indicate the standard error of the mean (*SEM*).

Search Tasks: Effects of Clutter and Stress

Only search and noticing times for correct responses were included in the analysis. Except for one miss that was excluded, all other search tasks were performed successfully. The mean search time was 12.3 (*SD* = 3.83), 7.80 (*SD* = 2.37), 29.08 (*SD* = 8.53), and 14.83 (*SD* = 4.15) in the LN, LS, HN, and HS conditions, respectively. A repeated-measures ANOVA (with listwise deletion and Bonferroni corrections) showed a significant effect of clutter on search time (F(1, 11) = 79.45, p < 0.001, η_p^2 = 0.878), which increased from an average of 10.1 (*SD* = 3.90) seconds in the low clutter condition to 21.9 (*SD* = 9.79) seconds in the high clutter condition (see Figure 6.3a). There was also a significant effect of stress on search time (F(1, 11) = 56.25, p < 0.001, η_p^2 = 0.836), which decreased from 20.6 (*SD* = 10.6) seconds on average in the no-stress condition to 11. (*SD* = 4.88) 31seconds in the stress conditions. Finally, a significant interaction effect between clutter and stress (F(1, 11) = 19.81, p = 0.001, η_p^2 = 0.643) was observed, such that the decrease in search time due to high stress was more pronounced in the high clutter condition (see Figure 6.3b). The effect of clutter was significant in both the low-stress (F(1, 11)

= 58.31, $p < 0.001$, $\eta_p^{2} = 0.841$) and high-stress ($F(1, 11) = 43.337$, $p < 0.001$, $\eta_p^{2} = 0.788$) conditions.



(a)

(b)

Figure 6.3: Effects of clutter and stress on search time: a) Search times for the different clutter and stress conditions and b) clutter-stress interaction effects

An independent samples t-test was used to examine possible negative effects of display adjustments on search time for each high-clutter scenario. Specifically, the search times for scenarios 2, 6, 10 and 14 were compared between Group 1 and Group 2, where information was highlighted for Group 1 while Group 2 did not experience any adaptations. Similarly, the search times were compared for scenarios 4, 8, 12 and 16 where Group 1 was not provided with adaptations but Group 2 was presented with the panel. Results show that there were no significant differences in the search times in any of the cases (see Table 6.2).

Table 6.2: Search time averages (and standard deviation) in the high-clutter conditions

Search time in secs *(SD)*	**High-clutter scenarios**							
	2	6	10	14	4	8	12	16
No adjustment	32.2 (16.53)	12.0 (3.57)	28.2 (4.10)	8.8 (392)	22.6 (5.31)	19.8 (6.59)	32.5 (12.5)	14.2 (4.7)
Highlighting	31.0 (15.75)	12.3 (5.31)	36.7 (12.3)	10.8 (8.3)	-	-	-	-
Panel	-				23.2 (9.42)	16.4 (2.06)	21.6 (9.04)	20.1 (7.77)

Noticing 1 Tasks: Effects of Clutter and Stress

As with search time, response times for the first noticing task were considered only for correct responses. Figure 6.4 shows the average values in each of the clutter and stress conditions, which were 11.75 (*SD* = 3.59), 9.5 (*SD* = 5.63), 21.56 (*SD* = 12.55) and 11.65 (*SD* = 5.86) seconds in the LN, LS, HN, and HS conditions, respectively. There were no significant effects of clutter or stress on Noticing 1 times, nor was there a significant interaction effect between clutter and stress. However, there was a trend toward longer Noticing 1 times in high clutter, with an average of 10.8 (*SD* = 5.16) seconds in low clutter and 17.6 (*SD* = 10.9) seconds in high clutter. Similarly, Noticing 1 response times decreased somewhat due to stress, from 18.1 (*SD* = 11.14) seconds on average in the no-stress condition to 11.2 (*SD* = 5.84) seconds in the stress conditions.

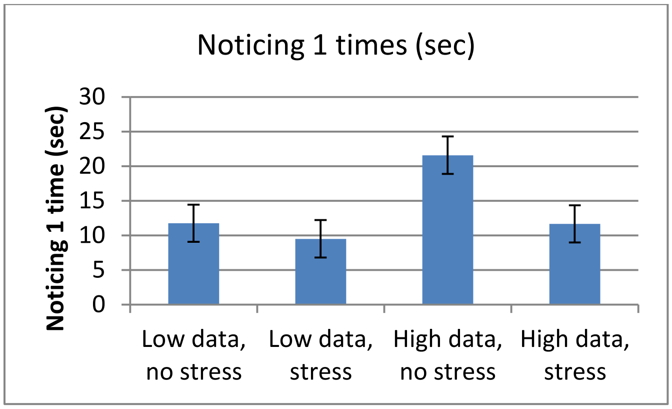


Figure 6.4: Noticing 1 times in the different clutter and stress conditions

The response times for the Noticing 1 tasks as a function of the different adaptation conditions can be seen in Table 6.3. Statistical analysis of these response times was not possible due to the large number of Noticing 1 misses or late detections (an average of 8 out of 12 Noticing 1 misses per trial). The entries marked N/A in the following tables indicate that all responses were either misses or provided at the end of the session, in which case no response time was obtained. Note that, in all cases, Noticing 1 time was equal to or shorter in the case of highlighting, as compared to no highlighting.

Table 6.3: Noticing 1 times (SD) in the different display-adjustment conditions

Noticing 1 time in secs *(SD)*	**Scenarios**							
	2	6	10	14	4	8	12	16
No adjustment	43.0 (0.00)	16.0 (0.00)	25.0 (21.2)	4.67 (0.47)	N/A	6.00 (0.00)	12.0 (0.00)	18.3 (4.49)
Highlighting	26.0 (11.8)	10.67 (4.03)	12.0 (3.67)	4.50 (1.12)	-	-	-	-
Panel	-	-	-	-	19.0 (0.00)	7.00 (0.00)	N/A	18.7 (3.96)

Comparing the number of misses across scenarios may give a better idea of the effects of display adaptations (see Table 6.4). A Fisher's exact test was used to determine whether there were any significant differences between groups. Only Scenario 2 showed a significant difference (p = 0.015 and with the strength of association being Φ = 0.845, p = 0.003), with significantly more misses when there was no highlighting. In all other cases, the number of misses was smaller in the case of highlighting, as compared to no highlighting; however, these differences did not reach significance. The panel, which was implemented to help improve Noticing 2 performance, not Noticing 1, did not lead to any significant differences in Noticing 1 times.

Table 6.4: Search miss rate in the different display-adjustment conditions (calculated as the sum of misses divided by 6, which is the total number of trials)

Miss rate (out of 6)	**Scenarios**							
	2	6	10	14	4	8	12	16
No adjustment	0.83	0.83	0.33	0.5	1	0.83	0.66	0.33
Highlighting	0	0.5	0.16	0.16	-	-	-	-
Panel	-	-	-	-	0.83	0.83	1	0.33

Noticing 2 Tasks: Effects of Display Adjustments

As mentioned earlier, the response time across all different scenarios for Noticing 2 tasks was not analyzed given that the amount of data outside the medical history page was not held constant between scenarios. However, the goal was to compare the same scenarios in the case where there was a panel or not. As can be seen in Table 6.5, it was difficult to discern differences due to the large number of misses. Those misses can be seen in Table 6.6. A Fisher's exact test was again used to compare the number of misses between groups. Here, only Scenario 8 showed a significant effect, with 6 misses in case of no adaptation, compared to no misses when the panel was presented (p = 0.002; strength of association: Φ = -1, p = 0.001).

Table 6.5: Average Noticing 2 response times (and SD) in the different display-adjustment conditions

Noticing 2 time in secs *(SD)*	**Scenarios**							
	2	6	10	14	4	8	12	16
No adjustment	63.5 (24.5)	31.2 (3.89)	62.5 (4.5)	25.3 (8.99)	57.0 (17.9)	N/A	62.0 (17.5)	N/A
Highlighting	78.0 (16.3)	23.6 (4.19)	66.3 (23.7)	33.3 (10.2)	-	-	-	-
Panel	-	-	-	-	109 (51)	29.5 (16.5)	49.6 (23.2)	30.6 (4.98)

Table 6.6: Noticing 2 miss rate in the high-clutter conditions (calculated by dividing the total number of misses across all participants by 6, the total number of trials)

Miss rate (out of 6)	**Scenarios**							
	2	6	10	14	4	8	12	16
No adjustment	0.66	0.33	0.33	0.5	0.5	1	0	1
Highlighting	0.5	0.16	0.16	0.16	-	-	-	-
Panel	-	-	-	-	0.5	0	0.16	0.5

Eye Tracking-Based Model Results

The modeling approach which was successful in Experiment 5 did not yield comparable results in this case, with an average three-fold cross validation accuracy for the training phase of 0.56 (RT model) and 0.53 (DataRT model). The train and test accuracy was also lower at 0.175 and 0.15 for the RT and DataRT models, respectively. Additional parameters of d' and β were thus not investigated further.

This poor predictive value could be inferred from the descriptive eye tracking results, which were not consistent with the patterns observed in previous experiments. The average scanpath length per second increased slightly, from 36.9 (SD = 7.37) pixels/sec in low-clutter to 41.3 (SD = 11.1) pixels/sec in high-clutter but there was no consistent pattern overall across participants. Similarly, mean fixation duration increased slightly on average, from 0.73 (SD = 0.19) seconds in low clutter to 0.77 (SD = 0.14) seconds in high clutter, as did mean saccade amplitude (24.8 (SD = 4.28) pixels in low clutter and 22.9 (SD = 7.74) pixels in high clutter). In both cases, though, these changes were not consistent across participants.

Subjective Results

Participants rated their perceived mental workload in the low-stress and high-stress conditions using NASA-TLX scales (scale: 0 to 20 (largest effect)). These rankings were analyzed using a Wilcoxon test for ordinal data. As can be seen in Table 6.7, there was a significant effect of stress only on the temporal demand scale, which increased from 6.5 (IQR = 4-14.5) in low stress to 17 (IQR = 13.7-17.7) in high stress.

Table 6.7: Results of the NASA-TLX ratings along the different dimensions (values in parentheses indicate the interquartile range (OWE)

NASA-TLX scale (0-20)	**Low stress median (IQR)**	**High stress median (IQR)**	**Wilcoxon exact sign test**
Mental demand	10 (9.2-11.5)	14.5 (11.5-16.7)	Not significant
Temporal demand	6.5 (4-14.5)	17 (13.7-17.7)	z = 2.807, p = 0.005
Performance	6.5 (4-9.7)	8.5 (5.2-13.0)	Not significant
Effort	12 (7-12.7)	13.5 (10.7-16.7)	Not significant

As part of the debriefing questionnaire, participants also provided their feedback and insight on the problem of EMR clutter and the possible benefits of the display adaptations. Overall, all participants acknowledged that clutter is a problem in their EMR. The main issue that was mentioned was the presence of a large amount of irrelevant data, in both the medical history page and in the physician notes. For example, every minor surgical procedure is listed under surgical history, even if it is a routing intervention. This sometimes makes for excessively long and meaningless lists of surgical history. Other participants pointed out how the organization of data in list form was not helpful and made it difficult to extract important information. The number of clicks required to navigate to desired places/information and the problem of fragmented data were also mentioned, as was the fact that there is potential for confusion with similar-looking terms (e.g., mycobacterial and myocardial).

Each group of participants was asked for feedback on the display adjustment employed in this experiment. Participants in group 1 (highlighting) generally felt that the adaptation was helpful, but two participants mentioned that it could be distracting when the highlighted information does not align with what they are looking for. Also, one physician pointed out that highlighting could anchor physicians on a particular diagnosis and make them miss contradicting information that supports a different diagnosis. In general, the impression was that highlighting would be most helpful if reserved for highly-relevant and critical information, such as heart disease and chronic illnesses. Only one person was completely opposed to highlighting, claiming that there was no need for any guidance.

With one exception, the participants in group 2 found the panel to be generally useful. Some pointed out that it was most useful for important laboratory results, like the elevated troponin during a myocardial infarction. The one participant who did not find the panel useful claimed that he did not even notice its appearance. None of the participants mentioned being distracted by the panel.

Discussion and Conclusion

Experiment 5, the final experiment in this line of research, was intended to implement and study the costs, benefits, and acceptance of eye tracking and performance-based real-time display adaptations in a realistic, stressful environment – the ED EMR. The goal was to shed light on a number of questions that will be discussed in turn in the following sections.

Question 1: Will the appearance of display adjustments a few seconds into a task help participants notice critical information or cause distraction, as reflected by performance and subjective measures?

This question addresses the possible tradeoff between costs and benefits of real-time display adjustments. The costs of display adjustments can be separated into those that may occur even with correct adjustments (i.e., pointing out information relevant to the diagnosis) as well as costs associated with incorrect adjustments. The main costs that would be of concern with the presence of correct display adjustments would be user distraction, frustration, and degraded performance. As for the costs of an incorrect adjustment, this could range from mere distraction to making an incorrect diagnosis. However, this second type of costs was not studied here as it is more strongly related to the NLP algorithm that is employed. Rather, the focus was on the effects of correct adjustments and their implications for performance.

With regards to highlighting, results showed that it helped improve Noticing 1 performance, both in terms of reduced response times as well as a decrease in the number of Noticing 1 misses. In all cases, Noticing 1 times were the same or lower, and the number of Noticing 1 misses were fewer in the presence of highlighting. As expected, highlighting seemed to be most beneficial when it pointed out an item in the surgical history list, an area of the medical history display that was consistently singled out by physicians as very difficult to search through. At the same time, there were no significant differences in *search* times for scenarios that contained highlighting versus those that did not. This suggests that the highlighting did not cause any notable distraction or cause participants to deviate from their main task despite the fact that, in the debrief, some participants mentioned that they felt the highlighting was distracting, and that it would be most beneficial if limited to critical issues.

The results for the shortcut panel suggest that it, too, was beneficial to performance, although this was not as consistently the case as for the highlighting. The largest benefit of the shortcut panel was seen when it pointed to laboratory results. Participants pointed out this benefit as well, as did the physicians during the earlier interviews (see Chapter 2). In particular, highly critical and time-sensitive information such as the troponin value that was displayed here is one type of data that physicians indicated they would like to be alerted to as soon as possible since delays could have significant bearings on the treatment outcome. When the panel pointed to current medications, there was no benefit, with results showing the same number of misses and even longer response time in the case of the panel. And when the shortcut was to a physician note, there were mixed results. In one case, there was considerable benefit as the information was always missed unless the panel was provided. However, in another case, there was no improvement. It is important to note that this variability between physicians is a typical feature of the ED environment and the medical domain as a whole (e.g., Mercuri & Gafni, 2011). Nevertheless, as with highlighting, there were no significant differences between search times when the panel appeared versus when it did not. There was also no discernable negative impact of the panel on the Noticing 1 tasks. This suggests that the panel was not particularly distracting, a conclusion that is supported by participants' subjective feedback. Adding display adaptations to the side of the display, as opposed to modifying some part of the display that is being used (as in the case of highlighting), may be preferable as it is less disruptive.

Question 2: Can the same modeling approach that was shown to be successful in a simple task environment accurately predict the effects of clutter in a more complex, stressful, and data-rich setting?

Overall, clutter affected response time as expected, with results showing significantly longer response times in the case of high clutter. However, the models for predicting clutter effects that were shown to be successful in the context of pure search tasks in a simulated graphics display environment (see Chapter 5) were not as effective in the ED environment.

The main reason for this discrepancy between Experiment 4 and Experiment 5 was the nature of the search questions employed here. In Experiment 5, a different, more complex and open-ended search task was used (e.g., “Find risk factors for stroke”). This was done to further test the models in a different context and also to create longer, more difficult scenarios (following from the results of Experiment 2).These more complex tasks, however, necessarily involved a degree of analysis and judgment on the part of physicians. Rather than immediately search for whatever target they were given, as was done before, participants now had to draw on their medical knowledge base to decide what some possible targets might be. Following this phase, they then had to determine, for each medical entry, whether that matched or was related to the possible options they had thought of. Physicians’ experiences and personal beliefs also played a large role in this task, where some physicians deemed a certain condition to be a risk factor while others did not. Thus all of these factors helped introduce significant variability across scenarios and physicians, which was not present earlier. This variability made it very difficult to trace the effects of clutter.

Another factor that may have played a role was the manipulation of clutter. Recall that in Experiment 4, the OrgRT model (i.e., model of RT in poor organization) showed the best predictive accuracy, highlighting the crucial role that organization plays in the creation and perception of clutter. However, in this experiment, only the amount of data was varied systematically since it represents the main challenge in the ED.

Thus, the current version of the proposed model appears to be best suited to predicting the effects of poor organization and its interactions with data density in the context of targeted search tasks. Further adjustments to the model are needed to increase its accuracy and robustness. For example, incorporating spread metrics may help with better detecting the effects of data density as they were shown to be highly diagnostic of the effects of clutter. In addition, further investigation into why stress was not induced in this experiment as in Experiment 5, as evidenced by the NASA-TLX metrics, will also be needed.

In conclusion, the results of this experiment indicate that real-time display adaptations are beneficial and accepted by users. Future studies could carefully design, test, and refine various forms of display adjustments to identify which work best across tasks and domains. The results of this experiment also highlighted the need to continue exploring means of improving the eye tracking-based models developed in Chapter 5. Including additional metrics or an adopting an alternative modeling approach may lead to more robust models. Together, the study of modeling approaches and suitable display adaptations can help ensure the reliable and timely detection of critical information.

References

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*,*52*, 139-183.

Howard Z, Siegelman J, Guterman E, Hayden EM, Gordon JA (2011). *Simulation Casebook.* The Gilbert Program in Medical Simulation, Harvard Medical School, Retrieved from http://mycourses.med.harvard.edu/ResUps/GILBERT/pdfs/HMS_7607.pdf.

St. John, M. , Smallman, H. S., Manes, D. I., Feher, B. A., & Morrison, J. G. (2005). Heuristic automation for decluttering tactical displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(3), 509-525.

Mercuri, M., & Gafni, A. (2011). Medical practice variations: what the literature tells us (or does not) about what are warranted and unwarranted variations.*Journal of evaluation in clinical practice*, *17*(4), 671-677.

CHAPTER 7

Conclusion

Display clutter is a widely-acknowledged but ill-defined problem that affects users in a range of situations – from website users (Grahame et al., 2004) to pilots (e.g., Kaber et al., 2008) – all of whom need to search for or detect information quickly and accurately. Clutter leads to delays and failures in information search and noticing (e.g., Alexander et al., 2008; Beck, Lohrenz, & Trafton, 2010; Bravo & Farid, 2006; Tullis, 1988) which, particularly in complex, safety-critical domains, can compromise system efficiency and safety. Accidents such as the Three Mile Island nuclear disaster (Lusted, 2012) have highlighted the detrimental effects of clutter, underlining the need to develop improved display designs that support operators in coping with large amounts of data. Overcoming clutter is particularly important in conditions of high stress, a noted feature of data-rich and safety-critical domains (e.g., Arora et al., 2010; Samel, Vejvoda, & Maas, 2004; Sexton, Thomas, & Helmreich, 2000). In such domains such as medicine, aviation, or process control, investing in means to overcome the effects of clutter is important and can help prevent major performance breakdowns.

One major challenge for clutter research is the lack of agreement on a definition of the phenomenon and the factors that constitute it. Chapter 1 detailed the efforts to unify, structure, and critique proposed aspects of clutter. A literature review of clutter definitions and measurement approaches emphasized the need to go beyond display-based definitions of clutter to include user-based factors, such as stress and experience, which can modulate the effects on attention and performance (Moacdieh & Sarter, 2014). This approach implies that the development of countermeasures to clutter requires a means of detecting and tracing its effects on performance and attention in real time – an approach that is in marked contrast to existing techniques that consider clutter a display feature and measure it using image processing algorithms (e.g., St. John et al., 2005). In this line of research, eye tracking was used to try and trace the effects of clutter on attention and thus explain the underlying reasons for observed performance decrements, both in the absence and presence of stress. During later stages of this work, eye tracking also formed the basis for triggering display adaptations in real time and in advance of performance costs of clutter. In addition to examining the effectiveness and feasibility of this approach, we also collected data on its acceptability to users.

In summary, the goals of this dissertation research were to

1. *Goal 1*: identify and empirically validate a comprehensive set of eye movement metrics that can capture and explain the effects of clutter (both high data density and poor organization) and stress on information acquisition
2. *Goal 2*: determine which of these metrics are best suited to tracing the effects of clutter in real time and use these metrics to build a model to predict the effects of clutter prior to significant performance decrements
3. *Goal 3*: implement, test, and evaluate the acceptability of real-time display adjustments to prevent attentional breakdowns in attention allocation and information acquisition.

This research was carried out in different contexts. The application domain that represented a complex, data-rich environment was emergency medicine. More specifically, the focus was on the electronic medical record (EMR) currently used by physicians at the University of Michigan Emergency Department (ED). Chapter 2 provided an overview of the domain, the EMR, and of research activities that served to familiarize myself with this field of work. Observations, interviews, and surveys with medical practitioners in the ED confirmed that clutter represents a significant problem for them in this high-workload, time-critical, and safety-critical environment (Kachalia et al., 2007).

In order to achieve Goal 1, the next step in this line of research was to establish the link between clutter, performance, and eye movements. Several metrics were examined that capture the spread, directness, and duration of fixations. Chapter 3 detailed two experiments that were conducted in order to analyze the effects of clutter on performance and use eye tracking to trace and explain these effects. This proposed eye movement metrics were thus tested and validated to identify the ones that best reflect the effects of clutter. Experiment 1 explored the effects of clutter on *visual search* in a *highly-controlled* environment. Then, Experiment 2, conducted in the context of the ED domain, built on the first experiment and incorporated *both search and noticing* tasks. Results showed that clutter affected both performance and a number of eye movement metrics. The eye movement metrics helped explain how the performance effects of clutter were mediated by its impact on attention management. In particular, the observed increased response time appeared to result from distraction by irrelevant data and a slower, more methodical search strategy in case of high clutter. Another finding was that stress exacerbated the effects of clutter, as reflected in both performance and eye tracking data.

Chapter 4 described efforts to distinguish between the two main aspects of clutter – data density and display organization – and their effects on performance and attention. In Experiment 3, the two clutter components were varied systematically in the context of a simulated graphics program. Results once again confirmed the detrimental effects of clutter on visual search performance. Notably, results revealed an interaction effect between data density and organization, where poor organization exacerbated the effects of high density. The eye movement metrics reflected these findings. They indicated that the increased search time was a function of increased spread, more deliberate search, and increased fixation duration. Poor display organization, in particular, contributed to slower search, which explains why fixation duration was not affected in the previous experiments where organization was not manipulated. In addition to mean fixation duration, the Nearest Neighbor Index (NNI) proved to be another metric that can differentiate between the effects of data density and poor organization.

The effects of stress that were observed in Experiment 2 were not replicated in Experiment 3. This may be explained by the fact that, with student participants performing simple search tasks, the motivation and risks involved may not have been sufficient to induce stress.

Experiment 4 was conducted to achieve Goal 2 of this dissertation, which was to determine which eye movement metrics are best suited to tracing and predicting the effects of clutter in real time. As described in Chapter 5, this experiment employed the same manipulations that were used as in Experiment 3; however, this time, data from a training phase were used to build and then evaluate logistic regression models in subsequent experimental trials. Each one of four models used three eye movement metrics to predict long response times in one of four different conditions: overall, high data density, poor organization, or high density *and* poor organization. Results showed that predicting long response times due to poor display organization could be modeled most accurately, once again highlighting the critical role of organization in slowing down search and contributing to clutter. The findings also indicate that eye tracking is well suited for detecting the effects of poor organization - a welcome finding considering that organization is very difficult to capture using other approaches, such as image processing which is better suited for calculating factors like density and color variation.

Finally, to achieve Goal 3 – the implementation, testing, and evaluation of real-time display adjustments, Experiment 5 tested the above four models in the more complex and data-rich environment of ED EMRs. Two types of display adjustments – highlighting and a shortcut panel – were triggered in real time by the experimenter to test their benefits to performance and their acceptability to physicians. The adaptations proved to be beneficial in terms of helping physicians notice important information. Based on their feedback, participants considered the adjustments to be useful, especially for highly critical data. In addition, based on both performance and subjective data, the adjustments did not appear to distract participants from their main tasks. These findings confirm that real-time display adjustments are a promising approach to counteracting the effects of clutter. In this experiment, there was again little effect of stress on performance or on the subjective ratings of stress. The time limit that was imposed on the search task was estimated based on Experiment 2 but it was likely too long. Finally, in contrast to Experiment 4, the eye tracking-based models of the effects of clutter exhibited poor accuracy in this case. This resulted could be due to the fact that the search tasks employed in Experiment 5 required physicians to make judgments and not merely search for a given target. This induced variability and may have significantly affected participants' eye movements, resulting in less consistent patterns of attention allocation.

Intellectual Merit and Broad Impact

This research makes significant contributions, both at a theoretical and an applied level. It adds to the knowledge base on visual attention and clutter by examining, systematically, the effects and interactions of data density, display organization, and stress on visual search and noticing. The work also contributes to the development of eye tracking-based metrics and models as a means to trace the effects of clutter on attention allocation in real time and trigger display adaptations in response to expected performance breakdowns. In particular, the departure from display-based metrics was one advantage of the proposed eye movement metrics, as was the introduction of directness metrics, which had not been greatly explored. Also, the appropriateness and effectiveness of various eye movement metrics for capturing different aspects of clutter was highlighted. Overall, spatial density was shown to be the most sensitive spread metric overall. NNI and mean fixation duration emerged as the most effective means of differentiating between the density and organization aspects of clutter. The ability to differentiate between clutter components is critical for identifying shortcomings and informing the targeted redesign of displays.

The use of eye tracking for detecting and tracing the presence/effects of clutter in real time is another novel contribution of this work. By developing a method for capturing clutter effects early on, over short time windows, this work lays the foundation for the creation of adaptive displays that can prevent, rather than merely respond to, the performance effects of clutter. While more work is needed to identify the most effective forms of adaptation, this research highlights the promise of the general approach, as evidenced both by performance improvements and subjective data showing that physicians felt favorable towards real-time display adaptations.

Ultimately, the findings from this research will help prevent performance breakdowns in complex and increasingly data-rich domains such as medicine, aviation, or process control. Ensuring that operators can locate information in a timely manner and do not miss critical information constitutes a major step toward promoting efficiency and safety in these domains.

Future Work

As with all research and given the necessarily limited scope of a dissertation, this work leaves a number of questions unanswered and suggests several directions for future studies. These will constitute part of the research plan I hope to implement as an Assistant Professor at the American University of Beirut starting Fall 2015.

First, this research investigated two of the main aspects of clutter, namely data density and display organization, but other factors remain to be explored. Issues such as color variation, target-background similarity, crowding, familiarity, and salience can influence attention allocation, eye movements and ultimately performance. Manipulating these variables systematically, and identifying their contribution to clutter, will likely lead to a better understanding of which eye movement metrics best trace clutter and possibly lead to improved and more robust models to predict clutter. Using more than two levels of clutter may also help better trace the effects of different aspects of clutter on performance. In addition, with the help of more powerful computing capabilities and techniques such as multi-threading, I hope to be able to increase the number of metrics that are employed in real time which may also improve the accuracy and robustness of the models further. It may then be possible to develop generic models of the effects of clutter that can be applied for any person or domain where a digital display is used.

Another research direction I am interested in is to further explore the effectiveness and acceptability of real-time display adaptations. In this dissertation research, the adjustments were deliberately few and simple, representing an initial test of whether users would accept such changes or not. Several other techniques could be more effective for guiding attention. For example, graded feedback where a signal is repeated with increased salience if it is not detected could be a promising approach that would minimize unnecessary disruptions by making use of preattentive processing (Sorkin, Kantowitz, & Kantowitz, 1988). The use of other modalities, such as audition or touch, for supporting visual search and noticing could also prove beneficial and avoid the risk of adding to an already saturated visual display (Sarter, 2006). In addition, in these studies it was assumed that the task the user was always performing the same task with the display (e.g., medical diagnosis for the current patient). However, other display adjustment techniques could involve displaying a query panel to allow users to indicate what it is they are looking for.

Finally, having had a longstanding interest in the medical domain, I plan to conduct further studies with medical personnel and patients in order to better understand their needs and requirements for EMRs. Guidance for the design or certification of EMRs is still missing, and usability issues continue to plague users and compromise patient safety (Terry, 2013). Developing and testing various implementations of EMRs, using eye tracking and performance data, could help in the development of guidelines for the design and implementation of better EMRs. This could be especially useful as current regulations do not address or take into account patient safety but rather focus more on issues such as billing and meeting meaningful use regulations (Hoffman & Podgurski, 2011).

References

Alexander, A. L., Stelzer, E. M., Kim, S. H., & Kaber, D. B. (2008). Bottom-up and top-down contributors to pilot perceptions of display clutter in advanced flight deck technologies. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1180-1184). New York City, NY: Sage Publications.

Arora, S., Ashrafian, H., Davis, R., Athanasiou, T., Darzi, A., & Sevdalis, N. (2010). Emotional intelligence in medicine: a systematic review through the context of the ACGME competencies. *Medical education*, *44*(8), 749-764.

Beck, M. R., Lohrenz, M. C., & Trafton, J. G. (2010). Measuring search efficiency in complex visual search tasks: global and local clutter. *Journal of Experimental Psychology: Applied; Journal of Experimental Psychology: Applied*, *16*(3), 238-250.

Bravo, M. J., & Farid, H. (2006). Object recognition in dense clutter. *Attention, Perception, & Psychophysics*, *68*(6), 911-918.

Hoffman, S., & Podgurski, A. (2011). Meaningful use and certification of health information technology: what about safety?. *The Journal of Law, Medicine & Ethics*, *39*(s1), 77-80.

Kaber, D. B., Alexander, A. L., Stelzer, E. M., Kim, S. H., Kaufmann, K., & Hsiang, S. (2008). Perceived clutter in advanced cockpit displays: measurement and modeling with experienced pilots. *Aviation, space, and environmental medicine*, *79*(11), 1007-1018.

Kachalia, A., Gandhi, T. K., Puopolo, A. L., Yoon, C., Thomas, E. J., Griffey, R., ... & Studdert, D. M. (2007). Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. *Annals of emergency medicine*, *49*(2), 196-205.

Moacdieh, N., & Sarter, N. (2014). Display clutter: A review of definitions and measurement techniques. *Human Factors.* doi: 0018720814541145.

Lusted, M. (2012). *The Three Mile Island Nuclear Disaster (Essential Events).* Essential Library

Samel, A., Vejvoda, M., & Maass, H. (2004). Sleep deficit and stress hormones in helicopter pilots on 7-day duty for emergency medical services.*Aviation, space, and environmental medicine*, *75*(11), 935-940

Sarter, N. B. (2006). Multimodal information presentation: Design guidance and research challenges. International Journal of Industrial Ergonomics, 36, 439–445.

Sexton, J. B., Thomas, E. J., & Helmreich, R. L. (2000). Error, stress, and teamwork in medicine and aviation: cross sectional surveys. *Bmj*, *320*(7237), 745-749.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *30*(4), 445-459.

St. John, M. S., Smallman, H. S., Manes, D. I., Feher, B. A., & Morrison, J. G. (2005). Heuristic automation for decluttering tactical displays. Human Factors: The Journal of the Human Factors and Ergonomics Society, 47(3), 509-525.

Terry, N. P. (2013). Meaningful adoption: what we know or think we know about the financing, effectiveness, quality, and safety of electronic medical records.*Journal of Legal Medicine*, *34*(1), 7-42.

Tullis, T. S. (1988). Screen design. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 377–411). Amsterdam: North-Holland.

APPENDICES

Appendix A: Resident Survey

Table A.1: List of all questions in the survey deployed among residents.

Questions	Options
What is your age?	10-year intervals from 20 to 80 (dropdown menu)
What is your current level of training?	Resident, physician, of physician assistant
What is your specialty?	Dropdown list of all medical specialties
Which EMR do you currently use?	18 options
Which EMRs have you used?	18 options
Please rate your comfort level with your current EMR	Likert scale 1-5
Please rate your satisfaction level with your current EMR	Likert scale 1-5
Please rate your comfort level with technology in general	Likert scale 1-5
Please rate the importance of good visual information display, in general	Likert scale 1-5
Please rate the importance of visual information display for clinical safety	Likert scale 1-5
In my experience, visual information display has contributed to the AVOIDANCE of a clinical error.	Likert scale 1-5
In my experience, visual information display has contributed to the CAUSATION of a clinical error	Likert scale 1-5
Please rate the importance of visual information display for clinical EFFICIENCY	Likert scale 1-5
How should visual information display be improved?	Free text
Which aspects of electronic health records should be addressed to improve clinical safety?	Free text
Which aspects of electronic health records should be addressed to improve clinical efficiency?	Free text
In the following section, please rate the amount of information you PERCEIVE in each image [25 images]	Scale 1-100

Appendix B: Physician Interviews

For each patient scenario, physicians had to answer the following two questions:

1. What information do you need to know at this point to be able to proceed with your diagnosis? Please try to list these in (relative) order of importance.
2. How would you proceed with the EMR (if applicable) in order to obtain the information that you need? Please walk me through the steps you are likely to take.

Table B.1: List of all the patient case scenarios presented to physicians during the interviews.

Order	Patient condition*	Presented scenario	Low/high stress	Stage 2/5**
1	Septic arthritis	A 45-year-old female presents to the ED with fever, nausea, and severe pain in both knees, which are swollen. She is having difficulty moving. She was in a mountain biking accident last week that resulted in numerous lesions and required stitches.	High stress	Stage 2
2	Ischemic stroke	A 75-year-old male presents to the ED feeling dizzy and with double vision. He is not speaking coherently and is being supported by his wife and son. His wife says he seems particularly weak on his right side.	High stress	Stage 2
3	DVT	A 38-year-old female presents to the ED with pain and swelling in her left calf. It started three days ago when she returned from a trip to Malaysia. At first she thought it was only a strained muscle and iced it but she is concerned about the swelling	Low stress	Stage 2
4	Aortic dissection	A 60-year-old male presents to the ED with acute onset of severe chest pain radiating to the back, nausea, and	High stress	Stage 2

		numbness.		
5	Hypotension	A 24-year-old female is brought to the ED by her friends who have all been taking part in a soccer game. The game started at noon and had been going on for 80 minutes when the patient mentioned that she was feeling dizzy and was going to faint. Her friends saw that she was losing her balance and insisted she goes to the ED.	Low stress	Stage 2
6	Unstable angina	A 65-year-old male presents to the ED with chest and neck pain, breathlessness, sweating, and nausea.	High stress	Stage 2
7	Peptic ulcer	A 45-year-old female presents to the ED with severe abdominal pain, vomiting, and nausea. She says she has been losing weight for some time. She has had pain when eating as well.	Low stress	Stage 2
8	Liver disease	A 72-year-old male presents to the ED with diarrhea, fatigue, slight abdominal pain, vomiting, and yellowish skin	Low stress	Stage 2
9	Gastric reflux	A 32-year-old female presents with chest pain, bloating, and regurgitation. She says it all started following lunch she had today and that she has never experienced pain like this before.	Low stress	Stage 2
10	Asthma	A 25-year-old female went on her daily run but decided to take a new route this morning. As she was out running past a construction site, she started having difficulty breathing. Patient slowed down the pace but still became increasingly short of breath and anxious.	Low stress	Stage 5
11	Migraine	A 29-year-old male presents with a throbbing left-sided headache that began gradually in the morning. This is similar to past headaches the patient has had, though somewhat more severe. Pain is localized behind the left eye and started “out of the blue” this morning as patient woke up, about 5 hours prior to arrival at emergency department. Patient tried Tylenol with no relief and is now starting to feel as though his/her vision may be getting a little blurry on the left side, where it really hurts.	Low stress	Stage 5
12	Anterior myocardial infarction	A 60-year-old male complains of crushing chest pain radiating to his neck and jaw on the left side. Symptoms started one hour ago during a business meeting. Patient had to excuse himself from the meeting as he	High stress	Stage 5

		became obviously diaphoretic and pale. Patient reports nausea and lightheadedness after the onset of the “crushing” chest pain.		
13	Compartment syndrome	A 19-year-old male spun out of control on the interstate and was brought by ambulance to the ED. The medics found patient lying on the ground, combative and with an obvious deformity of his right proximal lower extremity, which was splinted. At the scene, patient was not answering questions appropriately.	High stress	Stage 5
14	Alcoholic ketoacidosis	A 56-year-old male presents with uncontrollable nausea, vomiting, and acute abdominal pain. Patient claims to not have been able to keep food down for two days.	High stress	Stage 5
15	Narcotics overdose	A friend called EMS after he found patient on apartment couch not acting like himself. Friend reports patient has no known drug allergies or significant past medical history. Patient not following commands, occasionally yells “Get me outta here” or “leave me alone.”	High stress	Stage 5
16	Kidney stones	A 44-year-old female presents to the ED with extreme pain in her back and fever.	Low stress	Stage 5
17	Cellulitis	A 46-year-old male has been bitten by a spider on his left leg two days ago. The area is red, hot, and very painful.	Low stress	Stage 5
18	Acute cholecystitis	A 42-year-old female presents with persistent, sharp, severe RUQ pain for the past five hours. Patient states that she started experiencing “indigestion” about an hour after going out for breakfast with friends (ate fried potatoes, an omelet, and buttered toast); symptoms have not abated since that time. Patient tried taking some antacids, but this didn’t help.	Highs stress	Stage 5

*this was not given to physicians
**based on the distinction made in Chapter 2

Following these scenarios, physicians were also asked to expand on the following questions:

1. What do you think about the level of information in the EMR – is it too little, too much, or just right?

2. Do you find that information overload in EMRs is a problem? If so, how does it affect how you extract information from the EMR (your strategies, pages you avoid, etc.)?
3. Have you ever experienced a situation where you wanted to find something and were delayed because of the amount of information or display of information in the EMR? Please explain
4. In general, what information usually takes a lot of time to find and would benefit from a redesign?
5. Have you ever experienced a situation where you missed something because of the amount of information or display of information in the EMR and then had to reevaluate?
6. In general, what information is easily missed and would benefit from a redesign?

Appendix C: Experiment 2 Emergency Department Scenarios

Table C.1: Medical History page: text of the experiment trials shown to participants and the targets that they had to search for or notice

Diagnosis	Clutter	Given text (high stress scenarios in *italics*; search target shown in **bold** here)	Noticing target
		Simple tasks	
Peptic ulcer	Low clutter	A 45-year-old female has severe abdominal pain, vomiting, and nausea. Your medical student has evaluated the patient and believes she is suffering from cholecystitis. You open the patient's chart to check her medical history. Does the patient have a history of **gastric reflux**?	Cholecystectomy
	High clutter	An 80-year-old female has severe abdominal pain and nausea. Your medical student has evaluated the patient and believes she is suffering from appendicitis. You open the patient's chart to check her medical history. Does the patient have a history of **gastritis**?	Appendectomy
Ischemic stroke	Low clutter	*A 69-year-old male presents to the ED with dizziness. Your medical student has evaluated the patient and thinks he could be suffering from an inner ear disorder. You open the patient's chart to review his medical history. What year was the patient diagnosed with **diabetes**?*	Hypertension
	High clutter	*A 71-year-old male presents to the ED with a severe sudden headache. Your medical student has evaluated the patient and thinks he could be suffering from migraines. You open the patient's chart to review his medical history. What year was the patient diagnosed with **hypertension**?*	Atrial fibrillation
		Difficult tasks	
DVT	Low clutter	A 50-year-old female has pain and swelling in her left calf. Your medical student has seen the patient and believes she may have a deep vein thrombosis (DVT). You open the patient's chart to review her medical history. You also want to see if anyone in her family has had a DVT. Who is the **closest relative** to have had a DVT?	Tibia surgery

	High clutter	A 73-year-old male has pain, redness, and swelling in his right calf. Your medical student has seen the patient and believes he has a DVT. You open the patient's chart to review his medical history and that of his family. Who is the **closest relative** to have had a pulmonary embolism?	Knee surgery
Aortic dissection	Low clutter	*A 60-year-old male presents to the ED with acute onset of severe chest pain radiating to the back, nausea, and numbness. Your medical student has seen the patient and believes he has aortic dissection. You open the patient's chart to check for a history of aortic issues. What year was his most recent aortic **bypass surgery**?*	Father has aortic aneurysm
	High clutter	*A 79-year-old male presents to the ED with acute onset of severe chest pain radiating to the back and shortness of breath. Your medical student has seen the patient and believes he has aortic dissection. You open the patient's chart to check for a history of aortic issues. What year was his most recent aortic **root repair**?*	Mother had aortic dissection

Table C.2: Laboratory Results page: text of the experiment trials shown to participants and the targets that they had to search for or notice

Diagnosis	Clutter	Given text (high stress scenarios in *italics*; search target shown in **bold** here)	Noticing target
		Simple tasks	
Anemia	Low clutter	A 25-year-old female presents to the ED with weakness, fatigue, pallor, and shortness of breath. A complete blood count has been ordered for the patient. Your medical student has suggested, after evaluating the patient, that she is suffering from lack of sleep. You check the laboratory results for more evidence. What is the patient's mean corpuscular hemoglobin **(MCH)**?	Low hemoglobin
	High clutter	A 66-year-old male presents to the ED with fatigue and shortness of breath. His medical records show he suffers from peptic ulcers. Your medical student evaluates the patient and believes that this is a case of chronic fatigue syndrome. A blood test was performed and you open the patient's chart to check the results. What is the patient's mean corpuscular volume **(MCV)** for the last measurement?	Low hemoglobin
Myo-cardial infarction	Low clutter	*A 72-year-old male presents to the ED with chest pain, difficulty breathing, sweating, and nausea. Your medical student has examined the patient and believes that this is a case of peptic ulcers. You see that the patient has a history of high cholesterol. You open the patient's laboratory tests to check the troponin values. What is the last measured **troponin T value**?*	Elevated BNP
	High clutter	*A 77-year-old female presents to the ED with chest pain radiating to the left arm as well as vomiting. Your medical student has seen the patient, and after examining her believes this to be a case of gastric reflux. You know that the patient has a history of hypertension. You check the laboratory results for further information. What is the last measured **troponin I value**?*	Elevated myoglobin

		Difficult tasks	
Kidney stones	Low clutter	An 80-year-old female presents to the ED with extreme pain in her back, fever, and vomiting. Your medical student has evaluated the patient and believes she is suffering from kidney stones. You open the patient's chart to review her medical information. Is the current value of serum creatinine the highest ever for this patient? A 68-year-old female present to the ED with severe pain in her right side, fever, and vomiting. Your medical student evaluates the patient and believes she is suffering from kidney stones. You check the patient's laboratory results for more information. Is the **serum creatinine level** the highest ever for this patient?	High urine calcium
	High clutter	A 68-year-old female presents to the ED with severe pain in her right side, fever, and vomiting. Your medical student evaluates the patient and believes she is suffering from kidney stones. You check the patient's laboratory results for more information. Is the current **calcium level** the highest ever for this patient?	High creatinine
Liver disease	Low clutter	*A 72-year-old male presents to the ED with diarrhea, fatigue, slight abdominal pain, vomiting, and yellowish skin. Your medical student evaluates the patient and suggests that this could be some liver disease. You check the results from the laboratory tests that were ordered. Is the* ***prothrombin time*** *the highest ever for this patient?*	High AST
	High clutter	*A 76-year-old male presents to the ED with nausea, mild fever, abdominal pain, and diarrhea. He has a known history of alcoholism. Your medical student has evaluated the patient and believes that this is a case of liver disease. You open the patient's laboratory results. Is the* ***bilirubin level*** *the highest ever for this patient?*	Low albumin

Table C.3: Patient Timeline page: text of the experiment trials shown to participants and the targets that they had to search for or notice

Diagnosis	Clutter	Given text (high stress scenarios in *italics*; search target shown in **bold** here)	Noticing target
		Simple tasks	
Hypo-tension	Low clutter	A 61-year-old female is brought to the ED with trauma resulting from a car accident. Your medical student has evaluated the patient and says her temperature and blood pressure seem fine, although she is dizzy. You open the patient timeline to check the patient's progress. Was a **morphine injection** ordered after 2 pm?	Low blood pressure
	High clutter	A 93-year-old male has been in a car crash and is brought to the ED with abdominal and head pain. Your medical student has already seen the patient and mentions that despite feeling dizzy his temperature and blood pressure are fine. You open the patient timeline to check. Was his **Clonidine medication** ordered after 7:00 am?	Elevated heart rate

Unstable angina	Low clutter	*A 65-year-old male presents to the ED with chest and neck pain, breathlessness, sweating, and nausea. Your medical student has evaluated the patient and thinks this is most likely a case of unstable angina. You are looking at the patient timeline before going to see the patient. Has a* ***cardiac CT scan*** *been ordered yet?*	Elevated heart rate
	High clutter	*A 72-year-old male presents to the ED with chest pain and shortness of breath. Your medical student has seen the patient and believes that the patient might be suffering from unstable angina. You check the patient's timeline to learn more before seeing the patient. Has an* ***exercise ECG*** *test been ordered yet?*	Low blood pressure
		Difficult tasks	
Hepatitis C	Low clutter	A 57-year-old male presents to the ED with stomach pain and extreme fatigue. Blood tests show elevated ALT. Your medical student has evaluated the patient and believes this could be a case of ascending cholangitis. You check the patient timeline before going to see the patient. Has the patient's **blood pressure** decreased significantly?	Low temperature
	High clutter	A 42-year-old female presents to the ED with stomach pain, fever, and nausea. Blood test results show high AST levels. Your medical student has evaluated the patient and thinks she could be suffering from cholecystitis. You check the patient timeline before seeing the patient. Has the patient's **temperature** increased since the first measurement?	IV drug use
Compart-ment syndrome	Low clutter	*A 50-year-old male has been in a motorcycle accident and is brought to the ED with extreme pain, swelling, and numbness in his right leg. Your medical student has evaluated the patient and believes he has a broken tibia. You open the patient timeline to get a better picture. Was the* ***resting compartment pressure*** *taken after 3 pm abnormally high?*	Pain rating
	High clutter	*A 45-year-old male fell off the roof of his house and presents to the ED with severe pain, numbness, and tingling in his left forearm. Your medical student has seen the patient and believes he has broken his radius. You check the patient timeline to learn more. Was the* ***postexercise compartment pressure*** *reading after 7 am extremely high?*	Pain rating

Appendix D: Debriefing Document for Experiment 2 and Experiment 5

Principal Investigator: Dr. Nadine Sarter
Department of Industrial and Operations Engineering
University of Michigan

Participant information

Age: ___________
Resident, physician, or PA: ____________
Years in ED: _____________
EHRs used (name all that apply): ____________________________
Years current EHR has been used: ___________
How proficient would you consider yourself to be with the current EHR you use? Please circle one.

|----------|----------|----------|----------|
1 2 3 4 5
(Not proficient) (Very proficient)

Rating of clutter

Rate the following displays based on the amount of *clutter* that you believe characterizes each display (please circle one).

Display 1:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Display 2:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|

Low	1	2	3	4	5	6	7	8	9	High
clutter										clutter

Display 3:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|

Low	1	2	3	4	5	6	7	8	9	High
clutter										clutter

Display 4:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|

Low	1	2	3	4	5	6	7	8	9	High
clutter										clutter

Display 5:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|

Low	1	2	3	4	5	6	7	8	9	High
clutter										clutter

Display 6:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|

Low	1	2	3	4	5	6	7	8	9	High
clutter										clutter

Performance assessment

1. In general, you believe that your performance on these tasks was (please circle one)

|----------|----------|----------|----------|

1	2	3	4	5
(Poor)				(Excellent)

(Optional) Please explain any issues that you faced: ___________________________________

2. In general, the tasks, scenarios, and displays used in this experiment adequately represented real tasks
   1) Yes
   2) No. If no, please explain: ____

3. For all the tasks that *did not* involve a time limit, please select the point along the scales that best indicates your experience (adapted from Hart & Staveland, 1988)

Mental Demand: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc)? Was the mission easy or demanding, simple or complex, exacting or forgiving?

Low ____ High

Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the mission easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

Low ____ High

Temporal Demand: How much time pressure did you feel due to the rate or pace at which the mission occurred? Was the pace slow and leisurely or rapid and frantic?

Low ____ High

Performance: How successful do you think you were in accomplishing the goals of the mission? How satisfied were you with your performance in accomplishing these goals?

Low ____ High

Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low ____ High

Frustration: How discouraged, stressed, irritated, and annoyed versus gratified, relaxed, content, and complacent did you feel during your mission?

Low ____ High

4. For all the tasks that involved a time limit, please select the point along the scale that best indicates your experience

<u>Mental Demand</u>: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc)? Was the mission easy or demanding, simple or complex, exacting or forgiving?

Low High

<u>Physical Demand</u>: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the mission easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

Low High

<u>Temporal Demand</u>: How much time pressure did you feel due to the rate or pace at which the mission occurred? Was the pace slow and leisurely or rapid and frantic?

Low High

<u>Performance</u>: How successful do you think you were in accomplishing the goals of the mission? How satisfied were you with your performance in accomplishing these goals?

Low High

<u>Effort</u>: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low High

<u>Frustration</u>: How discouraged, stressed, irritated, and annoyed versus gratified, relaxed, content, and complacent did you feel during your mission?

Low High

Comments

Please describe any negative experience you have had in your use of EHRs where display clutter or data overload played a role

____

____

____

Please describe what you think is the main contributor to clutter and how that could be solved

____

____

____

(Optional) Feel free to add any comments about the experiment procedure or this study in general

____

____

____

Note: for Experiment 5, participants did not perform clutter ratings but answered this additional question at the end of the debriefing document:

Was the change to the display that was triggered beneficial (i.e., helps with the diagnosis) or harmful (e.g., distracting)?

____

____

____

Appendix E: Debriefing Document for Experiment 3

Principal Investigator: Nadine Moacdieh
Faculty Advisor: Dr. Nadine Sarter
Department of Industrial and Operations Engineering
University of Michigan

Rating of clutter

Rate the following displays based on the amount of *clutter* that you believe characterizes each display (please circle one).

Display 1:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Display 2:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Display 3:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Display 4:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Display 5:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Display 6:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Display 7:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Display 8:

|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
Low 1 2 3 4 5 6 7 8 9 High
clutter clutter

Performance assessment

1. In general, you believe that your performance on these tasks was (please circle one)

   |----------|----------|----------|----------|
   1 2 3 4 5
   (Poor) (Excellent)

   (Optional) Please explain any issues that you faced: ___________________________________

2. How much experience do you have with graphics/paint programs such as Adobe Photoshop®?

   |----------|----------|----------|----------|
   1 2 3 4 5
   (No experience) (Expert)

3. For all the tasks that *did not* involve collecting points, please select the point along the scale that best indicates your experience [see Appendix E for the NASA-TLX scales]

4. For all the tasks that involved points, please select the point along the scale that best indicates your experience [see Appendix E for the NASA-TLX scales]

Comments

How did the number of icons, organization of icons, and/or time pressure affect your search strategy? Please explain.

____

____

____

____

(Optional) Feel free to add any comments about the experiment procedure or this study in general

____

____

____

Appendix F: Description of EMR Pages Used in Experiment 5

Table F.1: Outline and description of the different pages implemented for Experiment 5

Page	Description	Sub-pages (tabs)	Description
Chart Review	Previous physician and nurse notes; each tab here displays a table of notes (each row is a previous note). Users can click on the rows of the table to view the note at the bottom of the page	Cardiology	Notes from previous visits with cardiologists
		Labs	Notes following the receipt of laboratory results, such as blood tests or biopsies
		Meds	List of current and previous medications; each row is a medicine and shows the start and end date for that particular medicine
		Notes	This general section shows all the notes from previous ED or other office visits (e.g., family medicine, ophthalmology, etc.) or surgical procedures (e.g., tonsillectomy)
		Radiology	This section shows any available radiology results or notes
Laboratory Results	List of all previous and currently available laboratory results, listed in a table	None	-
Patient Summary	Collection of all the previous and current (i.e., obtained during the present ED visit) medical data related to the patient	Clinical snapshot	This page shows an overview of the patient, including administered medications, nursing notes from triage, and vital signs. There is a link that takes the user to the Medical History page
		Imaging/EKG results	Any new imaging results, such as EKGs or X-rays will be displayed here
		Orders	This page shows all the current orders (e.g., blood bank, imaging, laboratory, etc.) that have been made in the ED in the current visit as well as their status (ordered or completed)
		ED medications	This shows the list of medications

		that have been ordered in the current ED visit
	Allergies	The patients confirmed allergies are listed here
	Triage summary	This is a summary of all the information collected during triage, including vital signs, the chief complaint, home medications, and the triage note

Appendix G: Details of Patient Scenarios Used in Experiment 5

Table G.1: Summary of scenarios used in Experiment 5.

Trial	Pre-trial information presented to participants			Correct answers [locations in the EMR]		
	Patient triage note	Suggested diagnosis	Search question	Noticing task 1	Noticing task 2	Correct diagnosis
			Low-stress scenarios			
1a **Low clutter**	Emma Burns is a 62-year-old female. She presents to the ED with severe abdominal pain, vomiting, and nausea.	Appendicitis	Has she had any abdominal complaints in the past?	Appendectomy [surgical history]	High alcohol consumption [notes]	Gastritis
1b **High clutter** ***(Highlight)***	Nora Terry is an 82-year-old female. She presents to the ED with extreme abdominal pain and vomiting. She is also suffering from a loss of appetite and bloating.	Cholecystitis	Does she have a history of abdominal complaints?	Cholecystectomy [surgical history]	Frequent visits to the ED for peptic ulcer [notes]	Peptic ulcer
2a **Low clutter**	Kelly Smith is a 65-year-old female. She presents to the ED with abdominal pain, loss of appetite, diarrhea, and nausea.	Inflammatory bowel disease	Does the patient have a history of gastrointestinal problems?	Recent sinusitis [medical history]	Taking antibiotics for sinusitis [medicines]	Clostirium difficile colitis
2b	Scott Vernon is a 63-	Kidney stones	Does the patient have a	Peptic ulcer [medical	Taking NSAIDS*	Peptic ulcer

High clutter ***(Panel)***	year old male. He presents to the ED with abdominal pain, bloating, and nausea.		history of kidney problems?	history]	for migraines [medicines]	
3a **Low clutter**	Jennifer Mitchell is a 52-year-old female. She presents to the ED with chest pain. You open the patient's chart to check her medical history.	Myocardial infarction	Does she have a history of cardiovascular issues?	Gastric reflux [medical history]	Frequent visits to the ED for gastric reflux [notes]	Gastric reflux
3b **High clutter** ***(Highlight)***	Rose Stoney is a 56-year-old female. She presents to the ED with a headache and blurry vision on the left side.	Stroke	Does she have any risk factors for stroke?	Migraines [medical history]	Frequent visits to the ED for migraines [notes]	Migraines
4a **Low clutter**	Mark Lawrence is a 46-year old male. He presents to the ED with dizziness and loss of balance.	Acute anemia	Does the patient have a history of blood disorders?	Recent pharyngitis [medical history]	History of pseudo-anemia [notes]	Labyrinthitis
4b **High clutter** ***(Panel)***	Katherine Willis is a 60-year old female. She presents to the ED with chest pain, cough, and difficulty breathing. The EKG is notable for a right bundle branch block.	Myocardial infarction	Does the patient have a history of heart disease?	Chronic obstructive pulmonary disease [medical history]	Patient has always had a right bundle branch block [notes]	Pneumonia
	High-stress scenarios					
5a **Low clutter**	Daniella Carter is a 23-year-old female. She presents to the ED with respiratory distress, wheezing, and nausea. She is also lightheaded and has a skin rash.	Anaphylaxis	Does she have a history of respiratory problems?	Milk allergy [medical history]	Low blood pressure [triage summary]	Anaphylaxis
5b	Mary Simpson is a	Anticoagulated	Does she have a history	Mother had a	Taking Lovenox	Anticoagulated

High clutter ***(Highlight)***	38-year-old female who was the driver in a motor vehicle collision. She is brought to the ED unconscious and with a blown left pupil. Your medical student suggests she is suffering from a subdural hematoma and will need surgery. The student then wonders if the patient is anticoagulated.		of blood clotting disorders?	pulmonary embolism [family history]	[triage summary]	
6a **Low clutter**	Ryan Woods is a 59-year old male. He is brought to the ED having collapsed while running a 5k in 100 F weather.	Stroke	Does the patient have a history of risk factors for stroke?	Alcohol abuse [social history]	Elevated creatinine kinase [laboratory results]	Rhabdomyolysis
6b **High clutter** ***(Panel)***	James Robinson is a 78-year old male. He presents to the ED with crushing substernal chest pain radiating to his neck and jaw on the left side.	Gastroenteritis	Does the patient have a history of abdominal issues?	Atrial fibrillation [medical history]	Elevated troponin [laboratory results]	Myocardial infarction
7a **Low clutter**	Hannah Jeffries is a 28-year-old female. She presents to the ED with pain and swelling in her left leg.	Deep vein thrombosis	Does she have a history of blood clotting disorders?	Diabetes [medical history]	Oral contraceptives [notes]	Deep vein thrombosis
7b **High clutter** ***(Highlight)***	Jonah Richards is a 77-year-old male who presents with 2 hours of mid-epigastric pain that is not relieved by	Myocardial infarction	Does he have a history of cardiovascular disease?	Diabetes [medical history]	Physician note about EKG being normal in the past [notes]	Myocardial infarction

	antacids. The patient has had similar "ulcer" symptoms but this time the pain is not going away and is non-radiating. An EKG was ordered and shows the presence of STEMI in inferior leads.					
8a **Low clutter**	Robert Hill is a 57-year old male. He presents to the ED complaining of nausea, vomiting, and abdominal pain which started last night. He also reports feeling weak, confused, and thirsty for the last 3 days.	Acute pancreatitis	Does the patient have a history of metabolic disorders?	Family history of diabetes [family history]	Pre-diabetes state [notes]	Diabetic ketoacidosis
8b **High clutter** ***(Panel)***	Simon Weer is a 71-year old male. He presents to the ED with dizziness and difficulty talking.	Stroke	Does the patient have risk factors for stroke?	Smokes [social history]	Has hypertension [notes]	Stroke

* Nonsteroidal anti-inflammatory drugs