UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES

School of Psychology

The Influence of Real-World Factors on Threat Detection Performance in Airport X-Ray Screening

by

Hayward James Godwin

Thesis for the degree of Doctor of Philosophy

December 2008

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES

SCHOOL OF PSYCHOLOGY

Doctor of Philosophy

THE INFLUENCE OF REAL-WORLD FACTORS ON THREAT DETECTION PERFORMANCE IN AIRPORT X-RAY SCREENING

by

Hayward James Godwin

The visual search task carried out by X-ray screening personnel has begun to be investigated in a number of recent experiments. The goal of the present thesis was, therefore, to extend previous examinations of the factors that may be detrimental to screener performance, to understand those factors in more detail, and to bring those factors to bear upon current models of visual search. It has been argued that screener performance is impaired by searching for infrequent targets (the *prevalence effect*), by searching for several targets simultaneously (the *dual-target cost*), and by the tumultuous environment in which screeners work . Over the course of six experiments, these factors, and, in some cases, the interaction between these factors, was examined. Experiments 1, 2 and 3 explored the role that the prevalence effect and the dual-target cost have upon the performance of untrained participants. Experiment 4 revealed that airport screeners are, in fact, vulnerable to both the prevalence effect and the dual-target cost, highlighting the relevance of the present work to those working in an applied environment. Experiment 5 tested the impact of ambient noise upon search performance and the dual-target cost, and found that ambient noise has no deleterious impact. Experiment 6 set the foundation for future research involving the impact of external distractions upon search performance, with the results showing that observers are slowed substantially when conducting even a simple mental arithmetic in conjunction with a search task. Based on the results from the experiments, it appears that actual screener performance could be improved by increasing the prevalence of 'dummy' items, as well as tasking with screeners to search for only a single target at any one time. Efforts could also be made to reduce sources of external distraction.

Contents

List of Tables

List of Figures

Important Note:

For security reasons, two key alterations have been made to the content of the present thesis prior to publication:

1) Appendix A originally contained images of the airport X-ray screening stimuli used in the experiments reported here; these images have now been removed
2) In the chapter that reports the results from an experiment in which airport X-ray screeners were participants (Chapter 5), the accuracy, reaction time, and signal detection measures have been hidden. Additionally, it must be noted that all graphs and figures that are presented in that chapter, aside from having their axis values hidden, also do not show the full range of accuracy values (i.e. from 0 to 1)

1

Literature Review

Developing Accounts of Visual Search and Improving the Airport X-ray Screening Task

> "Elected and appointed government officials have prescribed increased pay, more and better training, background checks, and federal employment to improve the proficiency of threat detection personnel. Seldom, however, has the fundamental problem of these systems been mentioned: *that they require people to perform tasks, under performance-degrading conditions, that they are ill-suited to perform in the first place.* Any significant improvements, therefore, must first require better integration of human operators, and technology within threat detection systems."
>
> -Harris (2002, original emphasis)

1.1 General Introduction

All too often, those involved in research and theory do not consider the applied implications of their work. Likewise, those involved in applied domains often do not consider the theoretical underpinning beneath their interests. This has been a classic problem that has plagued scientific inquiry for decades, and is still a major problem in modern psychology. The purpose of the present thesis is simple: to bring modern theories of visual search into the real world, to test them under the often-harsh light of ecological validity, and, finally, to incorporate an understanding of real-world factors into extant models and theories of search.

Bringing theoretical work based upon experiments and paradigms conducted in tightly-controlled laboratory settings into the real world is no easy task. Therefore, only an initial step will be taken in the present thesis, which will be beneficial for a number of reasons. The focal point of the research and theory presented here will be the screening task carried out by airport X-ray security officers. Security screening is a quintessential visual search task: screening personnel search X-rays of passenger baggage and must determine whether or not a threat/prohibited item is present within the bag. The extract from Harris (2002), presented above, serves as the core mandate for the work that will be described and detailed here: the task given to screeners is not an easy one, and the actual nature of the screening task is rarely, if ever, considered, as a potential source of error in their ability to detect threat items in passenger baggage.

The present Literature Review will begin by detailing previous research that has examined the performance of screening personnel. To understand how the screening task may impair threat detection performance, there will, following a description of previous applied research, be a review of previous theoretical research that will be of value to understanding the screening task, and the problems that screeners face. The theoretical framework developed during the present Literature Review will then be used as a foundation for the subsequent empirical chapters, which aim towards developing that framework considerably.

1.2 Airport X-ray Security Screening from an Applied Perspective

When stationed at a screening checkpoint, screeners must search for guns, knives, Improvised Explosive Devices (IEDs or 'bombs'), and other threat items within the X-ray images of passenger baggage. If they believe that they have located a threat or prohibited item, or feel that they cannot accurately determine whether or not such an item is present, they must highlight that bag as requiring manual inspection. Whilst this might appear to be a relatively simple process, it can be made very difficult as a result of both the screening task, and the conditions in which the task is carried out (Harris, 2002).

Searching X-ray screening displays for threat items is anything but trivial. The objects in the screening display, which are often difficult to identify even when they appear alone, can appear from any orientation, and overlap with one another in an endless variety of combinations. Guns and knives from a canonical viewpoint are more readily identified than guns and knives when viewed from above; IEDs vary a great deal in both shape and configuration, ranging from simple stand-alone devices to those embedded and hidden within other objects. Threat items can also be identified in terms of colour: guns and knives appear as a blue, blue-black, or green colour, whilst IEDs consist of both metal/electronic components, and explosive components (which appear as large masses of orange/brown). There can also be considerable random overlap between the many objects and items within an individual's baggage, making it very difficult to understand what those objects and items are in real life. Clearly, then, the visual search task involved in security screening is not an easy one to carry out effectively, with many obstacles that prevent efficient search and accurate detection. A limited set of example images are presented in Appendix A.

1.2.1 Previous Research into Airport Security Screening

Comparisons with Medical X-ray Imaging: The task carried out by airport X-ray security screeners can be compared to the task carried out by radiologists, who must search X-ray images of patients in the search for tumours. As with airport security screening, the chance of a target (a threat item or a tumour) being present is very low indeed, and the observer must search a complex image for a target of unknown appearance. The cost of missing a target in both radiology and airport security screening is very high indeed. On average, it is believed that radiologists have a tumour miss rate of somewhere between 20% and 30%, with a false positive rate of between 2% and 15% (Krupinski, 2000).

Applied research into the role of perception, visual search, and cognition in radiology began in the early 1940s (Kundel, 2006). More recently, applied research into radiographic image inspection has benefited from technological developments, and has made extensive use of eye-tracking to infer the psychological processes that are involved in the detection of tumours. Nodine and Kundel (1987) developed a model of visual search in complex radiographic displays. They argued that search begins with a global inspection of the display, during which an observer draws upon a cognitive schema to understand the image. The schema contains high-level information regarding the nature of tumours, their appearance in X-ray displays, the appearance of an X-ray that does not contain a tumour, as well as other useful knowledge and information that can aid in the efficient resolution of the task at hand. A comparison is made between the global, overall image that is first perceived, and the observer's schema. Any deviation from what is 'normal' or 'expected' within the schema is flagged by the visual system as a possible tumour location. A more detailed, focused examination is made upon each of these locations, or areas, in turn, by the observer. The detailed examination involves a decision regarding whether or not the selected area contains a tumour.

Evidence for such a set of processes comes from a number of studies using radiographic images. When an experienced radiologist is presented with an X-ray image for a very brief period of time indeed, so brief that they can not search through the image in any detailed manner, they can still achieve a high level of response accuracy, and accuracy is increased when the observer is allowed an unlimited amount of time to search the image (Kundel & Nodine, 1975). According to Nodine and Kundel (1987), this implies that the experienced observer is able to rapidly compare the image with their cognitive schema and determine the existence of any severe deviations from an X-ray that does not show the presence of a tumour. Additionally, when an experienced radiologist is presented with an image that is similar in appearance to an X-ray image (a modified black-and-white image of some clouds with a tumour superimposed somewhere within the image), but does not contain the standard structure of, for example, a chest X-ray, the radiologist then performs no better than a novice in their ability to detect a tumour. This has been taken to imply that the cognitive schema employed by radiologists is highly powerful in determining their detection performance (Hendee, 1987).

The model proposed by Nodine and Kundel (1987) was also used to generate a set of predictions regarding the errors produced by observers examining radiographic displays. Based upon eye-movement data, the authors suggest that response errors can be caused by three factors: *sampling error*, *recognition error*, and *decision-making error*. Sampling error refers to errors of omission. In essence, a sampling error occurs when an observer does not fixate on a given area of an image, when, in fact, the target tumour appears in that area of the image. A recognition error occurs when an observer attends to an actual target, but fails to recognise the area they are examining as containing a target. This can be especially the case when the target is deeply embedded in the surrounding area, causing it to be camouflaged. Observers typically require a greater period of time to recognise camouflaged targets, when compared to non-camouflaged targets, but may not maintain their fixation upon a given area for a sufficient period of time to be able to perceive a target that is camouflaged. Decision-making errors occur when an observer fixates upon various parts or portions of a target, but decides that they are viewing a non-target. Eye-movements show multiple fixations upon or around the target. The most common form of errors are decision-making errors, and are estimated to account for 60% of the errors made by radiographers, whilst 30% of the errors can be attributed to recognition errors, and 10% can be attributed to sampling errors (Nodine & Kundel, 1987).

In recent times, the research and theory into radiology developed by Nodine and Kundel (1987) has been extended successfully into the domain of airport X-ray security screening. In an initial study conducted by Gale, Mugglestone, Purdy and McClumpha (2000), a group of security screeners were engaged in a visual search for X-ray images of IEDs embedded within baggage displays. Screener eye movements were recorded. In an attempt to compare the performance of security screeners with radiologists, a replication of the procedure employed by Kundel and Nodine (1975: described above) was used. Screeners were presented with the images for either 200ms, 1s, or 6s. It was found that, as display presentation time increased, so did response accuracy, which was in agreement with the findings in the radiographic literature, and in the study carried out by Kundel and Nodine (1975). Additionally, the pattern of errors appeared to be similar to the patterns seen in radiographic image examination, with screeners most often not detecting IEDs that were present as a result of decision-making errors. That is to say, the screener spent a period of time examining the IED and the surrounding area, but, eventually, decided that an IED was not present.

Industrial Inspection and Airport Security Screening: A somewhat different approach to understanding the X-ray security screening task has been adopted by Schwaninger and colleagues (e.g. Schwaninger, 2004). As with the early radiographic image literature, screening performance is understood in terms of Signal Detection Theory. Appendix B presents a detailed account of Signal Detection Theory, but overall, Signal Detection methods are used to gain overall measures of performance, including a measure of how well a task was performed (indexed by *sensitivity* parameters such as *d'*), as well as a measure of whether participants were biased towards a particular type of response (such as 'target-present' or 'target-absent' responses, indexed by *bias* parameters such as the criterion, *c*).

Initial examinations showed that threat detection performance could be reduced by increasing the overlap between threat items and non-threat items, by changing the orientation of threat items to less canonical orientations, and by increasing the number of items within a bag (Schwaninger, Hardmeier, & Hofer, 2005). Additionally, a test of visual search ability indicated that screeners were no more skilled in visual search than novices – however, screeners were far more skilled in detecting and recognising the targets than novices, as a result of their considerable knowledge and experience with the targets (Schwaninger, et al., 2005). This is comparable to the findings described above indicating that radiologists were more skilled at detecting tumours than novices as a result of a high-level schema that can be utilised by experts to successfully resolve the task (Hendee, 1987).

Ghylin, Drury and Schwaninger (2006) have recently argued that X-ray security screening can be compared to the task of industrial inspection of sheet material, in which an observer must detect any number of infrequently-appearing faults, with a high cost for targets that are missed. An earlier, two-component model was developed by Spitz and Drury (1978) to examine the difficulties of visual search for defects in sheet material. The two component model delineates the inspection process into *visual search* and *decision-making.* This is notably similar to the models of visual search in radiographic images, and, as with the radiographic research, it has been argued that the pattern of errors produced by observers can be used to understand the processes involved in search and detection. However, unlike the radiographic literature, which focuses heavily on eye-movement data, the two-component model attempts to understand the search process via a complex analysis of response times. Ghylin et al. (2006) present data from trained screeners and naïve participants involved in the search for IEDs within baggage and show that the two-component model can be successfully applied to understand the search for threat items in security screening.

Beyond the two-component model, Ghylin et al. (2006) argue that an adaptive training regime can be highly effective in improving screener performance. Using a specialised software suite known as *X-Ray Tutor*, observers are trained in the search for threat items upon an individual basis. The software is connected to a centralised database containing detailed information regarding each observer's previous performance (Schwaninger, 2006). The data regarding previous performance is then used to determine the images that should be presented to each observer. Thus, when being trained, the target images that are presented to the observer are set at a slightly higher difficulty level than the observer is known to be able to achieve. When the observer fails to detect a target in a display, they are given detailed information regarding the nature and location of the target (Schwaninger, 2004). This drive towards detecting gradually more difficult targets is essentially an individual differences approach, and has been shown to improve relative performance considerably (Schwaninger & Hofer, 2004).

1.2.2 Discussion: Building upon the Foundations set down by Previous Applied Research

Now that some examples of applied research have been described, the present section will discuss how the previous applied research into airport security screening will be used as a foundation for gaining further insights into screener performance, and the screening task itself. It should be noted at the outset here that psychological research into the airport screening task is still relatively new, and only began properly in the early 1990s. As a result, although a great deal of work has examined the screening task since that time, there are still a very large number of factors that have been left untouched and require scrutiny. Although it would be impossible to consider all of these factors at once, at least at the moment, the present these will focus upon three key factors, namely the impact of target frequency, the impact of searching for multiple targets, and the impact of environmental distractors. Although various experiments have shown that all three of these factors can impair the search performance of observers (e.g. Han & Kim, 2004; Menneer, Barrett, Phillips, Donnelly, & Cave, 2007; Wolfe, et al., 2007), there has been very little examination of their relevance to the airport screening task.

One could argue that the ecological validity of the previous applied research is actually diminished by not having considered these factors, although, in all honesty, that would be a somewhat unfair stance to take. The previous research has been invaluable in gaining a foothold on the problems facing screeners when searching for threat items, and it now seems a fortuitous time to take the existing research and carry it forward by considering a set of additional factors that have not been previously examined. Indeed, that is the purpose of the present thesis.

1.3 Airport Security Screening from a Theoretical Perspective

The previous section outlined a series of applied perspectives that have examined the manner in which X-ray security screeners conduct their visual search task. The following review will develop a theoretical foundation for understanding the visual search task conducted by airport security screeners. As will be made clear shortly, a number of experimental paradigms that have emerged in recent years have not only begun to question extant theories and models of search, but can also be valuable to those working in an applied setting.

Why has visual search research not been applied to real world tasks more often before now? All too often, the problem lies with the constrained nature of theoretical research. For the most part, visual search experiments involve the search for simple, single targets (e.g. a blue square, a single letter, and so on) that are presented on around 50% of all trials: a markedly different situation to the screening task, yet still, not all that different from the design of the applied experiments described in the previous section, which often required screeners to search for a single target that was presented on 50% of trials as well. Therefore, connecting theoretical and applied interests will not only be beneficial to airport screening, but is also likely to test and extend current theoretical models of visual search. To begin with, however, a more basic account of the search process must be explicated and detailed.

1.3.1 Early Accounts of Visual Search

The study of visual search has been focused upon a set of key issues throughout its history. First and foremost, models and theories of search have all been aimed towards describing how the visual system is able to select and detect behaviourally-relevant target objects in the outside world (Yantis, 2000). The actual task of visual search is one which we all carry out on a very regular basis, and thus understanding visual search is not only important to esoteric research or specific applied domains, but is fundamental in understanding human behaviour and visual cognition in the real world. As the review below describes, understanding the cognitive and neural architecture that supports the visual search process has been rather problematic.

The debate regarding visual search typically begins with *Feature Integration Theory* (or FIT: Treisman & Gelade, 1980). According to FIT, the visual search process can be divided into two discrete stages. The first stage involves the parallel processing of basic elements of the display, called *features*, which can consist of shape, colour, motion, and so on. The parallel stage of processing was said to be *pre-attentive*: it operated temporally before the deployment of attention and was resource-unlimited (Treisman & Gelade, 1980). Later, it was proposed that features were realised neurologically in retinotopic maps of these different basic features (Treisman, 1998). If the target is sufficiently different to the distractors on the basis of any one of these features, it is very rapidly selected by the visual system and perceived by the observer. This is known as the *pop-out effect*. When the reaction times (RTs) of pop-out searches are plotted against the number of distractors present in the display, parallel searches typically yield flat search slopes. That is to say, the time needed to complete searches of this kind does not increase as the number of distractors increases (see the left-hand graph of Figure 1.3.1a, below).

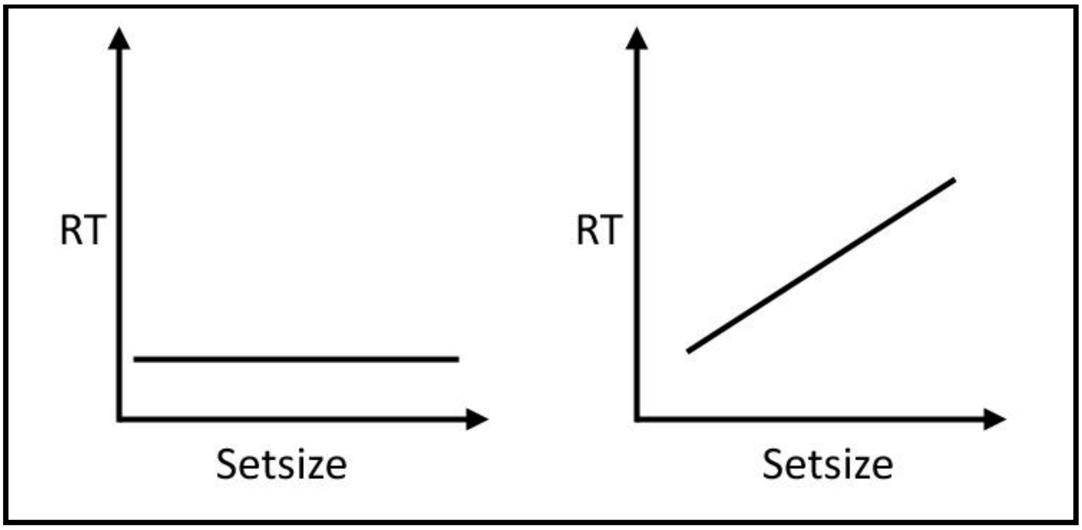


Figure 1.3.1a. Search slopes. The left-hand graph depicts 'parallel', 'efficient' or 'pop-out' search, where RT does not increase as setsize increases. The right-hand graph depicts 'serial' or 'inefficient' search, where RT does increase as setsize increases.

The serial stage of search begins once the parallel stage is complete, and involves the movement of attention around the search field. Each potential target is inspected, one at a time, until a target is found, or until there are no targets that remain to be searched. The serial stage of search is a laborious task, and requires the engagement of attention, rather like a 'spotlight' to bind the different features into a distinct perceptual unit (Treisman, 1998). Search slopes from serial search increase linearly as the number of distractors increases because increasing numbers of distractors need to be examined before the target can be located (see the right-hand graph of Figure 1.3.1, above). Although FIT has been very popular in visual search research, and remains so even to the present day (Quinlan, 2003), many of FIT's assumptions have since been discounted. To begin with, an alternative account of visual search, called *Guided Search* (or GS: Wolfe, Cave, & Franzel, 1989), argued for a modification of FIT to incorporate a flow of information between the parallel and serial stages of search. Under GS, parallel search can be used to select likely candidates for possible targets within the visual field, and pass those selections onto the serial stage, which is then automatically guided to those potential targets. This claim was based on experiments where triple conjunction targets (i.e. targets with three features) could be detected rapidly. Under the assumptions of FIT, triple conjunctions should be detected slowly. Under, GS, however, it was predicted that triple conjunctions would be detected rapidly, as there is greater featural guidance information available to the serial stage of search from the parallel stage (Treisman & Gelade, 1980; Wolfe, et al., 1989).

A revision of GS, called Guided Search 2.0 (GS2: Wolfe, 1994) incorporated aspects of Signal Detection Theory (see Appendix B) into its frame of reference. Under GS2, visual search is modelled in terms of the detection of target signals against background noise. Separate pre-attentive feature maps (for colour, orientation, and so on) register the existence of each target feature within a display and then the activation of these feature maps is summed to provide an overall map of where the target may be. The serial stage of search deploys attention, and iterates through the activation levels in this map, from the highest to the lowest, until a target is located, deemed to be absent, or a threshold is reached. This threshold is typically set so that any objects falling below it have a very minimal chance of being the actual target. A schematic of the GS2 process is depicted below, in Figure 1.3.1b.

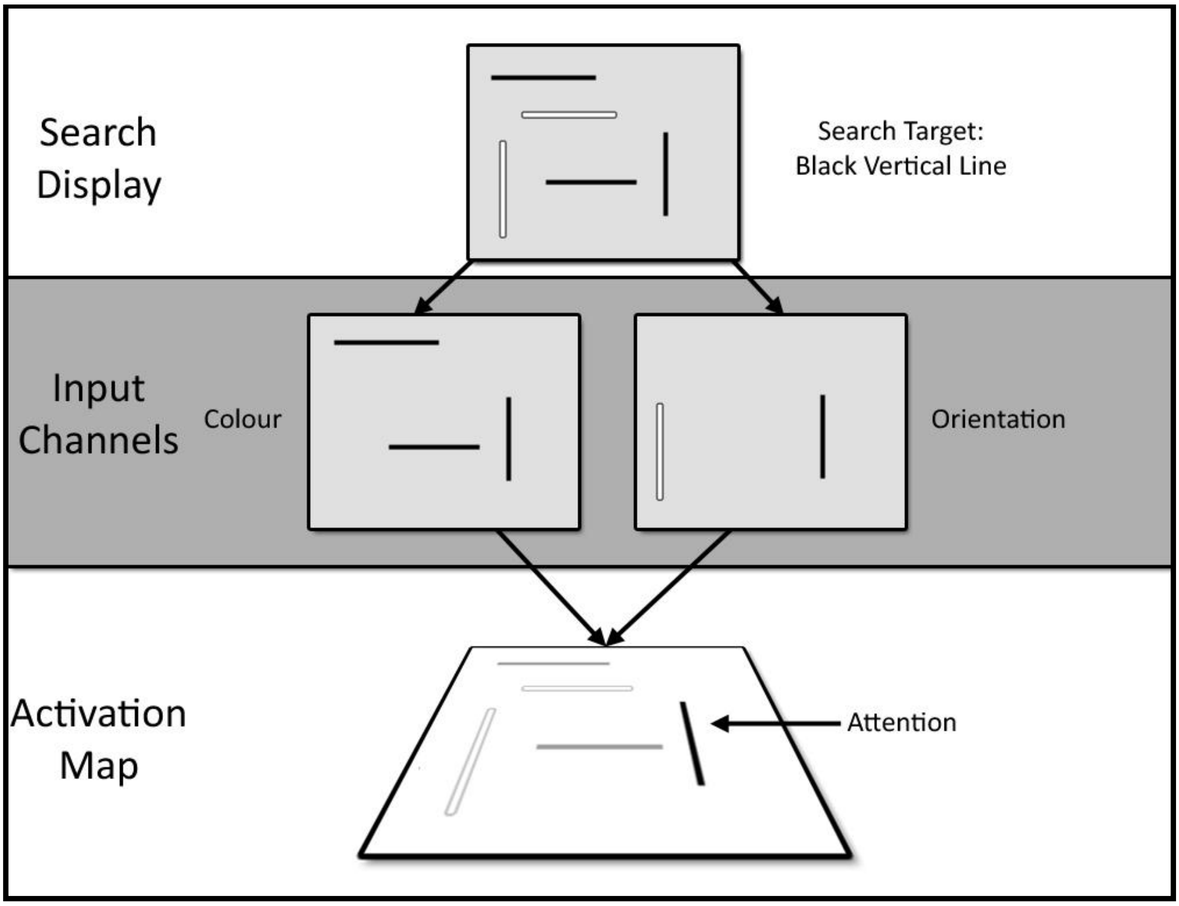


Figure 1.3.1b. Schematic of Guided Search 2.0. The observer is searching for a black vertical line. Separate Input Channels register the features being searched for (here: the colour black, and vertical lines). These registered features are then combined into an Activation Map, which highlights the potential location of the target.

1.3.2 Challenges to Early Assumptions

As new paradigms and experimental procedures emerged, it became evident that a number of the early assumptions regarding visual search that were made by FIT needed to be re-assessed. The challenges were further buttressed by technological advancements that have made eye-tracking and neuroanatomical research more readily available. The challenges and extensions to the early accounts of visual search will now be described.

Search efficiency: The first criticism of the early theories of visual search questioned the very existence of the parallel and serial stages of search. Duncan and Humphreys (1989) provided evidence to support the notion that, in fact, there were no separable stages of visual search, and that visual search performance could better be understood in terms of a continuous scale of *efficiency*. Search efficiency, they argued, was based upon the interaction between two factors: the similarity between the targets and distractors, and the similarity between the distractors themselves. Duncan and Humphreys (1989) found that, when the target was similar to the distractors, search efficiency was poor. Additionally, when the distractors varied more in their appearance (so became more *heterogeneous*), search efficiency was reduced. Search efficiency was lowest when the targets were heterogeneous, and also shared characteristics with the target. Unfortunately, the suggestions made by Duncan and Humphreys (1989) were somewhat overshadowed by the publication of the first version of GS in the same year (Wolfe, et al., 1989), along with the debates between GS and FIT that followed, and it was not until several years later that the notion of search efficiency was considered (Wolfe, Friedman-Hill, & Bilsky, 1994).

In order to further examine the notion of search efficiency, Wolfe (1998) conducted an analysis of the results from a large number of experimental studies (totalling over one million visual search trials), and found no evidence for the existence of separable stages of search. The combined search slopes across different experiments showed a *continuous* difference between response times, rather than a distinct *categorical* difference as the serial/parallel theories of visual search would predict. It was therefore suggested, in agreement with Duncan and Humphreys (1989), that search could be more accurately conceptualised in terms of a continuous scale of efficiency and inefficiency. Drawing upon the formulations of GS, parallel and serial search were re-cast as being *efficient* and *inefficient* search, with the only difference between the two being the amount of parallel guidance (i.e. the amount of information being utilised by parallel processing) involved in the search process. When there is a great deal of parallel guidance, the target is selected rapidly from the distractors, making search efficient, and yielding flat or near-flat search slopes (Wolfe, 1998). These search slopes are caused by the selection of a small number of potential targets, meaning that very few objects are inspected so search is rapid. Conversely, when there is little parallel guidance, the search is inefficient, and typically yields steep search slopes. These steeper slopes are caused by the selection of large numbers of objects as potential targets, many of which need to be inspected before the actual target is found, increasing response times significantly.

Pop-out and attention: Further criticisms of FIT came in the form of a set of experiments which provided evidence that pop-out could be primed. When a pop-out target was presented on consecutive trials, RTs to detect the target decreased significantly (Maljkovic & Nakayama, 1994, 1996, 2000). If pop-out were based upon a single, automatic process that occurs without the need for attention, then it is unlikely that pop-out could benefit from priming. Elsewhere, using a different paradigm, Joseph, Chun and Nakayama (1997) reported that, if attention were diverted to a secondary task within the same display as the search task, pop-out search was not possible, and argued that such an effect would not have been present if pop-out was a resource-unlimited parallel process.

Is search parallel? Is search serial? Treisman and Gelade (1980), in the formation of Feature Integration Theory, argued that search slopes were able to distinguish between parallel and serial visual search architectures. However, Townsend and colleagues have argued for several decades, using a variety of paradigms, that different search slopes do not necessarily imply that visual search is either parallel or serial (Townsend, 1972, 1990; Townsend & Colonius, 1997; Townsend & Wenger, 2004a, 2004b). Instead, it has been argued that search slopes which were apparently indicative of a serial visual search task could have also been generated by a parallel process that was limited in capacity, and was thus unable to process all elements of a display simultaneously.

The parallel-versus-serial debate has still not been resolved. Even in recent times, evidence is being produced, using increasingly complex modelling techniques, that search is primarily parallel in nature (Fific, Townsend, & Eidels, 2008), or that search is generally parallel with some serial components (McElree & Carrasco, 1999; Olds, Cowan, & Jolicoeur, 2000; Olds, Jolicoeur, & Cowan, 2001; Thornton & Gilden, 2007). The parallel-versus-serial debate has produced diametrically opposing accounts of cognition and processing, with, for example, van der Heijden and Bem (1997) proposing that search is unlimited in capacity, with the only problem being the limitations in input from the eyes. Tsotsos (1997) has produced a rebuttal to the claims made by Heijden and Bem (1997), highlighting the diverse and often belligerent state of the current views on visual search and visual cognition in general.

It is unlikely that the parallel-versus-serial debate will be resolved any time soon. Indeed, the current models and theories have reached such a high level of complexity that only those with a high level of mathematical prowess are able to actually make any significant contributions to the existing body of research; such a situation will naturally slow the progress (or otherwise) of accounts of visual search in the future. Still, further evidence for a mixed parallel and serial form of search can be taken from the study of eye movements.

Eye movements and visual search: The recording of eye movements have become an increasingly popular method for studying visual cognition in the last three decades (Liversedge & Findlay, 2000). Examinations of the eye movements are useful for a number of reasons. First of all, the nature of the eye leads to a number of limitations on the processing of the visual input. The fovea of the eye produces input of the highest visual acuity, and covers only the central 2° degrees of visual angle; beyond the fovea is the parafovea, which provides input of lower acuity for a further 5° degrees of visual angle (Balota & Rayner, 1983; Rayner, 1978). Finally, beyond the parafovea, visual acuity is very poor indeed, despite our conscious experience that our view of the world is clear and fully-coloured (Balota & Rayner, 1983; Rayner, 1978). Perhaps not surprisingly, the central regions that are fixated upon with the eyes are given a disproportionately large area of processing within the brain, compared to non-central regions (Carrasco, Evert, Chang, & Katz, 1995; Carrasco & Frieder, 1997; Carrasco, Mclean, Katz, & Frieder, 1998).

The existence of eye movements pose a further problem for accounts of visual search, as they impose a form of seriality whenever the objects in the search display can not be seen perfectly without actually being fixated upon (or near). Still, a number of studies of eye movements and visual search have examined overt behaviour that has been able to shed light upon existing models of search. For example, Zelinsky and Sheinberg (1997) found that the number of eye movements made during a visual search task increased linearly with increases in set size, mirroring some of the effects seen in the search slopes of classic visual search experiments (e.g. Treisman & Gelade, 1980). Thus, it was argued that, although the number of eye movements made during a search task does not match the number of objects in the search display, the search system can scan, in parallel, several objects near each fixation that is made (Motter & Belky, 1998a, 1998b; Zelinsky & Sheinberg, 1997). Additionally, in line with the notion that attention is directed towards objects in the display that are similar in some form to the target (Duncan & Humphreys, 1989; Wolfe, 1994), it has been reported that eye movements are often made towards objects in a display that share at least one feature with the target (Findlay, 1997; Findlay, Brown, & Glichrist, 2001).

Neuroscience, top-down processing and bottom-up processing: Neuroscientific research has begun to play an increasingly important role in developing models of visual search. Indeed, revised versions of FIT were partially inspired by neurological data (Treisman, 1998). Through the use of neurological research, it has been possible to partially observe the underpinnings of the visual search process. A number of neurological studies have observed that similar cortical areas are utilised in both efficient and inefficient searches (Bichot, Rossi, & Desimone, 2005; Leonards, Sunaert, Van Hecke, & Orban, 2000), lending support to the notion that previous attempts to examine search as being 'parallel' or 'serial' in nature were a mischaracterisation of the search system (Wolfe, 1998).

Neurological studies (and indeed the modern visual search literature in general) are typically concerned with *top-down* and *bottom-up* factors, and how the interplay between those two factors contributes to visual search performance (Corbetta & Shulman, 2002). Bottom-up factors are typically tied to the intrinsic *salience* of objects in a display, and a number of models have emerged which are geared towards predicting and mimicking the manner in which salience is computed within the human brain (Itti & Koch, 2000). Top-down factors are typically related to an observer's expectations or predictions, which in the case of visual search takes the form of the target *template* (Duncan &

Humphreys, 1989; Navalpakkam & Itti, 2002, 2007). At any given point in time during a visual search, an observer must select a target from a number of other objects. In some search tasks, this is a trivial process; in other search tasks, selecting the target is considerably more difficult, yet still the target can be detected, despite the task being made more difficult by increasing the number of target-similar distractors (Duncan & Humphreys, 1989). How is this all achieved neurologically?

A number of modern accounts of the neurological function of attention argue that there is considerable intrinsic *biased competition* through the various stages of visual processing (Desimone & Duncan, 1995). Input from the retinas proceeds to the early visual areas, which is then decomposed into basic parts (e.g, colour, orientation, etc.); the processing stream then proceeds to V4 where more complex attributes are processed (e.g. combinations of both shape and colour), followed by the inferotemporal and prefrontal cortices, where the shapes become organised and finally reach conscious perception (Connor, Egeth, & Yantis, 2004; Theeuwes, Itti, Fecteau, & Yantis, 2005; Yantis, 2000, 2005). As the processing proceeds from the early to the later visual areas, the receptive fields of the neurons involved in each area grows larger. This presents a problem, however, because, later processing stages can have multiple stimuli falling within their receptive fields, posing a problem for the visual system in terms of disentangling the objects from one another and perceiving them accurately.

The solution to this problem, it appears, is that selective attention is recruited to select which of the many stimuli falling into these broad receptive fields are relevant to the observer (Yantis, 2000, 2005). Consider a search scenario where an observer is searching for a blue square. If the observer is presented with a display containing a red circle and a blue square and assuming that both are of equal bottom-up salience, how does the visual system select and attend to the blue square rather than the red circle? The answer is that the search template simply biases the cortical networks to detect the target. The biasing involves an increase in the activation throughout the processing stream in favour of neurons which fire relating to the target in question (Yantis, 2000).The activation stems from top-down input.

In the example of searching for a blue target, neurons which fire when the colour blue enters their receptive field become more highly active. A number of studies have provided evidence to suggest that even early visual areas can have their activity modulated by attentional goals (Liu, Larsson, & Carrasco, 2007; Yantis, 2005), including even the lateral geniculate nucleus (O'Connor, Fukui, Pinsk, & Kastner, 2002). Thus, when bottom-up information enters the system indicating that a target may be present in a given location, the orchestration of top-down and bottom-up input increases the chance that an observer will attend to the target. In terms of a saliency map that marks locations of potential targets (as in Guided Search 2.0: see Wolfe, 1994), it thus appears that the map can be understood in terms of both top-down and bottom-up information (Fecteau & Munoz, 2005; Fecteau & Munoz, 2006).

One of the assumptions of the biased competition account is that the high-level visual areas are able to co-ordinate their processing with the low-level visual areas. More recently, examinations of such processing have been made in the form the *re-entrant* account (Di Lollo, Enns, & Rensink, 2000), which highlights the fact that traditional neurological accounts of visual processing propose that the processing operates in a *feed-forward* manner: travelling from the low level visual areas to the higher level visual areas. Di Lollo et al. (2000) note that there is considerable evidence to suggest that, in fact, within the brain, the information streams are not only feed-forward, but re-entrant in nature. Re-entrant processing involves an initially feed-forward stream, which can then loop back upon itself to further 'understand' what is being processed. Under the re-entrant account, the visual system is continuously making *perceptual hypotheses* about the nature of the objects in the outside world, and testing them, matching both low- and high-level input against existing memory structures and expectations (Hamker, 2005).

In the example of visual search, a re-entrant account posits that the visual system, having been biased towards the detection of the target, makes perceptual hypotheses about where a target may be in the display. A saccade is initiated at or near the potential target for further checking: the increased visual acuity proffered by an eye movement then allows for a higher-resolution comparison of the bottom-up input with the high-level depiction of the target (Lleras, Rensink, & Enns, 2007; van Zoest, Lleras, Kingstone, & Enns, in press). If the object under inspection matches the high-level depiction of the target, a 'present' response is made; if not, the percept of the object being inspected is updated (i.e. the object is then perceived as a distractor, and does not reach conscious awareness until that is the case). For example, in the search for the letter 'T' amongst 'L's, the visual system may have to make several hypotheses and update them before the target is found. When fixating upon what is actually an 'L', the visual system will then need to update its hypothesis, and perceive an 'L', rather than perceive a 'T' (Lleras, et al., 2007; van Zoest, et al., in press).

1.3.3 Modern Accounts of Visual Search: Dynamic Filters and Guided Search 4.0

The re-entrant account of visual processing has inspired the development of the most recent accounts of visual search, which will now be described. These recent accounts also form the most comprehensive and modern criticisms of standard models of visual search.

Perhaps the most vocal critique of the traditional models of visual search comes from Di Lollo, Kawahara, Zuvic and Visser (2001). The critique is based upon a number of factors and other findings. They argue against the existence of pre-attentive feature maps, and the existence of a two-stage, parallel/serial conception of visual search for a number of reasons. Firstly, Di Lollo et al. (2001) contend that pre-attentive feature maps can not exist for reasons of *biological plausibility*. Put simply, there are just so many features reported to exist that there is insufficient space within the human brain to process them all. Di Lollo et al. (2001) favour an account based upon a process similar to that espoused by Nakayama and Joseph (1998), in which attentional allocation in visual search can be understood in terms of a trade-off between scale and resolution. Attention can be directed to produce a low resolution sampling of a large portion of a display, or can be focused to produce a high resolution sampling of a small portion of the display (note the similarity between such an account and the radiographic accounts of Nodine & Kundel, 1987). Thus, efficient search occurs when the low-resolution sampling of attention is able to capture a single target from amongst the distractors. Conversely, inefficient (or serial) search is needed to provide a detailed examination of each object in the display.

If search is not dichotomous, and pre-attentive feature maps do not exist, how, then, are targets detected? Di Lollo et al. (2001) introduce the concept of *dynamically configurable input filters* to answer just such a question. The input filters are controlled by high-level mechanisms, and operate in a goal-directed manner. If the filters are able to prevent distractors from being attended to, search is efficient; however, if the filters are configured so that this can not occur, search becomes inefficient.

The input filter model has been extended by Wolfe and Horowitz (2004), who agreed with much of the dynamic filter theory, yet also suggested some alterations to it. Their suggestions were based upon an infusion of some of the basic concepts of GS. They proposed the existence of a high-level control module that guides attention to potential targets, rather than a set of filters that removes non-targets, primarily because filters remove information and so change what is being perceived: "Although attending to an object or location might have perceptual consequences, guidance itself should not" (Wolfe & Horowitz, 2004 p.2). Beyond this, they also attempted to understand what target attributes are able to affect this guidance system. Drawing on a great deal of previous research, they categorised different attributes on how likely those attributes are able to guide attention, and the pattern of results showed that colour, motion, orientation and size are definitely able to act as guiding attributes. Shape, however, was arguably less useful in guiding attention, although still may be able to do so in a limited fashion. For the purposes of the present discussion, Wolfe and Horowitz's (2004) high-level control module will be termed the *dynamic guided search module*.

Finally, the notion of a re-entrant network has been encapsulated in the most recent revision of Guided Search, in the form of Guided Search 4.0 (GS4: Wolfe, 2007). Under GS4, the structure of the search system can be compared metaphorically to a carwash: cars enter the carwash in serial, but a number of cars can be in the carwash simultaneously. In terms of search, this means that GS4 models the search system as having a bottleneck in terms of objects entering the search system, yet, still, the architecture allows for the processing of multiple objects in parallel (see also Wolfe, 2003). Thus, GS4 models the visual search system as having a hybrid parallel-serial architecture. In GS4, early parallel processes feed into object recognition processes via a selective bottleneck. Either one, or a limited number of, objects pass through the bottleneck; actual access to the bottleneck is controlled by selective attention, such that, only objects which bear similarity to the target are allowed to enter the object recognition processes (this is an instantiation of the selection mechanism suggested by Wolfe & Horowitz, 2004).

The object recognition stage within GS4 is modelled as an *asynchronous diffusion* process. Diffusion and random-walk models have been popular and effective in modelling real RT and performance data from a variety of visual search tasks (Nosofsky & Palmeri, 1997a, 1997b; Ratcliff, 2006). In a diffusion process, information begins at some neutral point, and further evidence is gathered as processing continues. When evidence is gathered, that evidence can lead to one of a number of decisions, which are made once the acquired evidence crosses a given threshold. In the case of visual search, there are two possible decisions: that the object currently being processed is a target, or that the object is a distractor. In GS4, a number of objects can be simultaneously processed in parallel, and finish at different times (see Figure 1.3.3a, below).

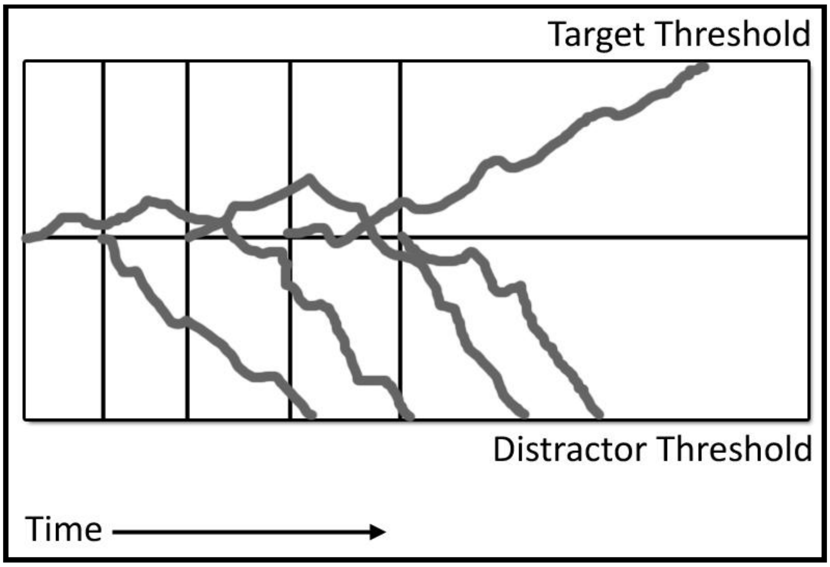


Figure 1.3.3a. Asynchronous diffusion model of Guided Search 4.0. Processing for each object selected for inspection begins at a mid-point between a Target and Distractor Threshold. As each object is examined, they progress towards either of the thresholds. Vertical lines represent an additional object entering the recognition progress. Multiple objects can be examined simultaneously.

At present, GS4 is incomplete, and its author notes a number of problems with the current state of the model (Wolfe, 2007). Still, GS4 is able to mimic and produce quantitative outputs that match a number of visual search tasks given to human observers (in terms of both reaction times and error rates). Overall, though, the present state of GS4 neatly highlights many of the developments of how visual search has come to be understood over the past three decades or so.

1.3.4 The Prevalence Effect

The models and theories regarding the nature of visual search that have been described so far all have been based upon experimental paradigms in which a target is presented on 50% of all trials. Even the most up-to-date attempts at modelling visual search, in the form of GS4 (Wolfe, 2007), are only concerned with such conditions. As such, current models have little scope for explaining or accounting for how the search process may be affected when an observer must search for a target which appears with a frequency of less than 50% (or more than 50% also).

In recent times, however, the impact of variations in target frequency have begun to be examined in detail, under the rubric of the *prevalence effect*, with *prevalence* referring to the frequency at which a target is presented (Wolfe, Horowitz, & Kenner, 2005). Put simply, it has been reported that, as target prevalence decreases, the probability that a target will be detected decreases as well (Wolfe, et al., 2005). Target prevalence has particular relevance to airport security screening because real threat items appear very infrequently. For the most part, the only threat items that screening personnel are presented with are those from a system called Threat Image Projection (TIP). This builds computer-generated images of threat items into the display on around 2% of all trials.

Initial examinations suggested that participants were engaging in a *speed-accuracy trade-off* when prevalence was reduced (Wolfe et al., 2005). Reaction times for target-absent trials were rapid in low prevalence (1% prevalence), but not in moderate prevalence (50% prevalence). It appeared that participants in low prevalence were responding “absent” before giving themselves sufficient time to detect a target. Thus, they were gaining in speed, at the cost of response accuracy. This was further supported by an eye-tracking study: participants searching for low-prevalence targets inspected fewer objects in a display before producing a response than participants searching for higher-prevalence targets (Rich, Hidalgo-Sotelo, Kunar, van Wert, & Wolfe, 2006).

Further support for a speed-accuracy trade-off account has been presented by Fleck and Mitroff (2007), who reported that allowing participants the opportunity to “correct” their target-absent responses after each trial eliminated the prevalence effect. Thus, Fleck and Mitroff (2007) argued that the prevalence effect was caused by observers producing an inappropriately rapid response, highlighting the importance of *motor priming* in the emergence of the prevalence effect.

A series of other experiments have cast doubt upon the speed-accuracy trade-off account of the prevalence effect. Using complex images from airport X-ray baggage screening, van Wert, Horowitz and Wolfe (2007) found that allowing participants to “correct” their responses attenuated, but did not eliminate, the prevalence effect. More importantly, several experiments conducted by Wolfe et al. (2007) show rather clearly that the prevalence effect is more than simply a speed-accuracy trade-off. Providing participants with ‘speeding tickets’ did not eliminate the prevalence effect. Second, preventing participants from responding before a given time period had elapsed was also unable to eliminate the effect.

Finally, and most importantly, Wolfe et al. (2007) grouped participants into pairs, and presented each pair with the *same stimuli* (participants were involved in the experiment separately). Crucially, the results indicated that the *same* targets in low prevalence tended to be missed by participants in each pair. In other words, participants *consistently* missed the same targets in low prevalence. Such a result suggests that the prevalence effect is more than just a speed-accuracy trade-off. If observers were trading speed for accuracy in low prevalence, it is unlikely that they would have missed the same targets, as they would simply be responding in a “fast and careless” manner. Couching their findings in terms of Signal Detection Theory (Green & Swets, 1966), Wolfe et al. (2007) argued that reductions in prevalence caused a *criterion shift* in decision-making (see Appendix B). In other words, low prevalence search results in observers needing stronger evidence that a target is present before they will actually respond “present”. The criterion shift that occurs in the face of low target prevalence is one of the core issues that will be examined within the present thesis, and will be examined in greater detail in subsequent chapters.

The Vigilance Decrement: In general, it has been reported that, as time-on-task increases, maintaining an accurate vigil for a rare event in a visual display becomes increasingly difficult. This is known as the *vigilance decrement* (Mackworth, 1968). The vigilance decrement is typically understood in terms of Signal Detection Theory: as time-on-task increases, so does the criterion, increasing the amount of evidence needed for the observer to report that a target is present, and thereby causing subsequent target presentations to be missed (Parasuraman, 1979). Although there has been little argument about the existence of the vigilance decrement, there has been considerable dispute and disagreement regarding the cause of the decrement (Davies & Parasuraman, 1982). The first account of the vigilance decrement was posited in behaviourist terms—the observer’s conditioned response (i.e. responding to the presence of a very low-frequency target) was gradually extinguished by repeated target-absent trials.

Thus, when the target does finally appear, the observer does not respond to the target, as that response has been made extinct (Mackworth, 1968). Despite the parsimony of such an account, it did not receive a great deal of support. For example, several studies have provided evidence to suggest that the vigilance decrement can fluctuate (in some cases, there may be a *vigilance increment*: Davies

& Parasuraman, 1982), and that performance did not decrease at a constant rate (Deese, 1955).

Since the initial studies of vigilance, a number of competing accounts have emerged to explain the effect (Baddeley, Cocchini, Della Sala, Logie, & Spinnler, 1999; Caggiano & Parasuraman, 2004; Davies & Parasuraman, 1982; Grier, et al., 2003; Helton, et al., 2005). In general, however, vigilance research has proven to be very problematic. Aside from debates and arguments regarding the cause of the vigilance decrement, numerous studies have failed to find any form of decrement whatsoever (Koelega & Brinkman, 1986).

The vigilance decrement has been of particular interest to applied researchers for a number of years. Indeed, early studies of vigilance were inspired by the finding during the Second World War that radar operators would often miss approaching enemy vessels after continuously monitoring the radar screen for a number of hours consecutively (Davies & Parasuraman, 1982). In response to the potential problems proffered by the impact of the vigilance decrement, airport screeners are only required to work on their screening task for periods of 20 minutes at any given time. However, it has been reported that, in fact, this may not be a necessary rule to follow: it has bee found that time on task had no impact upon search performance of screening personnel, even after one hour of searching X-ray screening images. Additionally, in a more recent study, Wolfe et al. (2007) showed that the vigilance decrement is separate from the prevalence effect because standard tests of fatigue and alertness, typically shown to be significant in studies of vigilance, show no impaired vigilance or fatigue in low-prevalence visual search.

1.3.5 Searching for more than one Target: the Dual-target Cost

As with the impact of target prevalence, the models and theories regarding visual search tend to have little scope for examining the impact of searching for several targets simultaneously. Again, even GS4 makes no attempt to replicate or simulate behaviour in the search for multiple targets. However, it is important to understand the impact of searching for more than one target, because X-ray screening personnel are required to search for a large number of targets when examining X-rays of passenger luggage. Is it possible for a human observer to carry out a visual search for metals and IEDs as accurately and rapidly as a search for metals and IEDs alone?

Early Research: A number of early experiments reported that observers were able to search for multiple targets simultaneously. For example, Neisser, Novick and Lazar (1963) found that, after a lengthy period of practice, the participants were able to search for ten items as rapidly as they were able to search for one. Additionally, the results indicated that the participants could search for ten targets with the same response accuracy as searching for one target. A number of other experiments echoed the finding that observers could search for a number of targets effectively with practice (Duncan & Humphreys, 1989; Kaplan & Carvellas, 1965; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977; Sperling, Budiansky, Spivak, & Johnson, 1971; Treisman & Gelade, 1980).

One of the first studies to indicate that searching for several targets may be less effective than searching for a single target comes from Sperling and Melchner (1978), who presented participants with search arrays very briefly, in order to prevent the use of eye movements. Participants were unable to carry out an effective simultaneous search for two numerals of different sizes placed within an array of letters. In all cases, accuracy for one target suffered, such that participants were effectively only searching for one of the targets. Which target suffered on any given trial varied considerably, leading the authors to suggest that participants were randomly allocating their resources between the two targets, as, when instructed to only search for one of the targets, performance was increased for that target alone.

Elsewhere, Pashler (1987) asked participants to search for two letters (C and E) and varied the number of distractors that were similar to either target (using Gs and Fs as target-similar distractors, and Xs as target-dissimilar distractors). He found that, as the number of target-similar distractors increased, so did the time taken to detect a target, as well as the time taken to respond 'absent'. The increase in RT occurred regardless of which target was actually present in the display. Pashler (1987) subsequently argued that the results were in line with an emerging theory developed by Hoffman (1979), which posited that the visual system tags objects in a display in parallel for subsequent serial checking. Hoffman's two-stage model (1979) was a predecessor of Guided Search (Wolfe, 1994, 2007; Wolfe, et al., 1989).

Can the standard models of search simply be extended to infer that searching for two targets will simply take longer to complete than searching for one target, with the increase in RT being dependent upon the number of objects in the display similar to either target? Pashler's (1987) experiments certainly seem to indicate that to be the case. However, one problem with Pashler's (1987) experiments, as well as the other early experiments, is that they involved the detection of simple digits or letters, or simple featural stimuli (Duncan & Humphreys, 1989; Kaplan & Carvellas, 1965; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977; Sperling, et al., 1971; Treisman & Gelade, 1980). To what extent do the early results extend to more difficult search tasks? D'Zmura (1991) was perhaps the first to examine such a question in detail, through the use of manipulations in colour.

D'Zmura's (1991) approach to colour in visual search was novel. He examined how search performance varied in relation to the colours of the targets and distractors, and found that search was rapid (i.e. had flat search slopes) if the target was linearly separable in colour space from the distractors. A colour space plots out colour along a range of different axes: see, for example, Figure 1.3.5a. When stimuli are plotted in a colour space, along a set of axes, D'Zmura (1991) found that, if the target could be linearly separated from the distractors, then the target would pop-out (i.e. search slopes were flat). This result is presented in Figure 1.3.5a.

Following his initial single-target search experiments, D'Zmura (1991) then examined the impact of linear separability in the search for two differently-coloured targets, finding that, when both targets were linearly separable from the distractors in colour space, then search was rapid, and would result in pop-out. D'Zmura's (1991) results have since been replicated and extended in a number of other studies (Bauer, Jolicoeur, & Cowan, 1996, 1998, 1999; D'Zmura & Knoblauch, 1998).

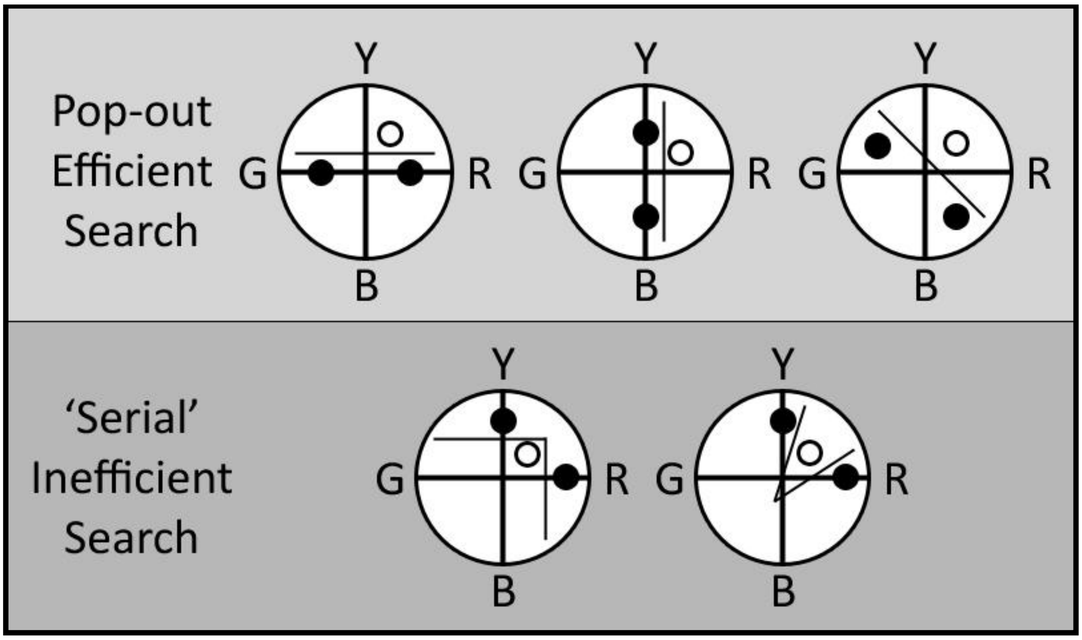


Figure 1.3.5a. Colour spaces and search performance from D'Zmura (1991). Efficient search is possible when the target (white circle) is linearly separable from distractors (black circles) in colour space (plotted here on Green-Red / Blue-Yellow axes). Search becomes inefficient when the target cannot be linearly separated from the distractors (lower panel).

Recent Research: The Dual-target Cost: For the most part, studies of linear separability were focused solely on the effects of separability, and not in the search for several targets simultaneously. As a result, Menneer, Barrett, Phillips, Donnelly and Cave (2004) developed the dual-target search paradigm, in order to examine the overall impact of search for two targets, and to allow for a comparison with performance in searching for a single target. They asked participants to search for one of two possible targets which differed from one another, and from their distractors, in colour. Each participant was engaged in three different searches. There were two *single-target* searches, one for each colour from the target pair, and a *dual-target* search, where participants were to search for both of the targets simultaneously. The results showed that there was a *dual-target cost* for both response times and response accuracy, with search efficiency and response accuracy being reduced in dual-target search, when compared to single-target search. Additionally, when searching for two differently-coloured targets, the RTs were shown to be greater than just an addition of the response times from the two component single-target searches.

The nature of the dual-target cost was later examined in more detail by Menneer, Barrett, Phillips, Donnelly and Cave (2007). There was a dual-target cost in search efficiency for targets that could not be linearly separated from their distractors in terms of colour. However, a more extreme cost was detected when searching for two targets defined by different orientations: the participants abandoned one of the targets, and searched for only one of the targets in dual-target search. The same was true when they were asked to search for two differently-shaped complex targets that were the same colour.

Following up these findings, Menneer et al. (2007) also combined the effects of shape and colour on dual-target search. In one experiment, the target pair consisted of two differently-oriented oval shapes. These shapes were coloured and differed from their distractors along the orientation dimension. Dual-target search for these shapes exhibited a cost in search efficiency (or time) that was consistent with the cost detected when searching for different-coloured targets. Additionally, a further experiment involving dual-target search for one target that was defined in terms of orientation, and a second target that was defined in terms of colour, showed no dual-target cost in terms of search efficiency, although there was a dual-target cost upon the response accuracies, with all but one participant opting to focus their search for the target that differed from the distractors upon the basis of colour.

It was argued by Menneer et al. (2007) that the pattern of results obtained in their experiments can be attributed to processing differences involved with the stimuli that were used. In particular, they noted that colour appears to be of high importance in visual search. This is due to the finding that dual-target search for targets that differ from their distractors in colour is inefficient, as was the dual-target search for targets of different orientations that were also different colours to one another. Also, when search is carried out for one target defined by orientation, and a second target defined by colour, the coloured target is focused upon whilst the orientation target is ignored. Menneer et al. (2007) argued that the salience of colour, and the manner in which colour is processed, allows it to act as a powerful factor in guiding search. Such a result is similar to the review conducted by Wolfe and Horowitz (2004), who argued that colour is one of the strongest determinants of search performance.

A second set of studies have also provided evidence to suggest that searching for numerous objects that differ in terms of their shape is very poor indeed. A novel *Visual-Search and Categorisation* (VSC) procedure that investigates *target-target* similarity across trials was used. Smith, Redford, Gent and Washburn (2005) carried out a detailed series of VSC experiments, in which participants had to detect one of several targets (defined by their shape into different categories) from amongst a set of distractors (there were always seven objects in each display). The objects in VSC displays were placed randomly, and were allowed to overlap with one another. Each trial immediately ended after six seconds. Following a considerable period of training, participants were barely able to achieve beyond chance performance in detecting the targets. When the number of target categories was reduced, performance was still not improved. However, when the different targets categories were gradually made more similar to one another, performance was greatly improved. In a second study, which involved more extensive training, it was found that introducing novel stimuli into each category reduced performance considerably, to the extent that observers appeared to be learning, over time, each image from each category, and found it difficult to generalise their knowledge of each category (Smith, Redford, Washburn, & Taglialatela, 2005).

Thus, at best, searching for several targets can be slower and less accurate than single-target search. As the search becomes increasingly difficult, one target can be ignored, with search focusing instead on the other target. At the extreme of difficulty, examined by Smith, Redford, Gent and Washburn (2005), as well as by Smith, Redford, Washburn and Taglialatela (2005), observers are unable to focus on a single target, and search appears to collapse into chance-levels of response accuracy.

One notable problem that remains to be resolved within the study of the dual-target cost is, specifically, *why* there is a cost for response accuracy and/or response times. A recent study conducted by Stroud, Menneer, Cave, Donnelly and Rayner (in preparation) is useful in understanding the causal relationship between dual-target search and impairments in performance. Participants were given the task of searching for ‘flags’ and ‘striped rectangles’, the former being a simple imitation of a gun-shape, whilst the latter was intended to resemble an IED. The targets were presented with a set of distractors that were non-linearly separable in colour space. Eye movement data were collected, and indicated that, in dual-target search, participants were less selective in their examination of the distractors. In dual-target search, participants examined distractors that they did not examine in single-target search. Interestingly, these distractors bore little resemblance to either of the targets that were being searched for, at least in terms of colour. The eye movement data are still being examined and modelled in detail, and it is likely that further insights will be available in the near future. However, for the time being, a number of potential factors that could give rise to the dual-target cost have been detected. For example, the examination of objects that are in no way similar to either target (let us call them *irrelevant objects*) could have two very direct routes to impairing the response accuracy of observers engaged in dual-target search.

First of all, the examination of irrelevant objects may cause observers to reach their internal, subjective measure of a 'quitting threshold' in terms of time. Once such a threshold is reached, the observer often either guesses or responds 'absent' if no target has been found (Chun & Wolfe, 1996). In essence, irrelevant objects may waste the time of an observer, such that they reach their quitting threshold before they have conducted a sufficiently detailed examination of the search array.

A second possibility is that, as more objects are examined in dual-target search (i.e. objects similar to both targets, as well as irrelevant objects), observers may tax their memory for objects in the display that have been examined and rejected as being distractors. Although there has been some controversy over whether or not visual search utilises any form of memory for previously-inspected objects in a display (Horowitz & Wolfe, 1998, 2001), a putative agreement is being reached which notes that search utilises a spatial memory of the display, in order to prevent repeatedly inspecting the same objects unnecessarily (Horowitz, 2006; Peterson, Kramer, Wang, Irwin, & McCarley, 2001). Unfortunately, the spatial memory is severely limited (McCarley, et al., 2006; McCarley, Wang, Kramer, Irwin, & Peterson, 2003); indeed, in GS4, memory is modelled as having a record of only the most recent three objects that have been examined (Wolfe, 2007). Thus, with more objects being inspected in dual-target search, repeat-inspections become more likely as spatial memory becomes exhausted, thus resulting in a second possible route for observers to reach their search termination threshold prematurely, and fail to detect a target, or incorrectly guess and respond 'absent' (Chun & Wolfe, 1996).

1.4 Discussion: Synthesis and Direction for the Present Thesis

Visual search is a complex task that is still not fully understood. Two major criticisms of previous work in the study of visual search are that, all too often, models and theories regarding search have not had the scope to explain or account for the impact of searching for targets which are presented infrequently, or to explain or account for the impact of searching for several complex targets simultaneously. Initially, the first steps to be taken in the present thesis will be to examine the impact of these factors upon performance, and to provide an overall account of visual search that is based upon existing models, but incorporates new data as well. In order to achieve this goal, the prevalence effect and dual-target cost need to be explored in more detail.

2

Using the Dual-target Cost to Examine the Prevalence Effect

Connecting Current Research with Existing Frameworks

> "One could argue that studies of how we scan our visual environment have been stuck in the eternal present, investigating the properties of a particular search situation without reference to what has occurred before. There is, however, increasing evidence that what we have previously viewed…has a large influence on what we see, what grabs our attention and how we organise the visual scene."
>
> -Kristjansson (2008)

2.1 Introduction

As was noted in the preceding Literature Review, current models of visual search are not easily extended to account for the prevalence effect and the dual-target cost. Additionally, there is still an ongoing debate regarding the nature of the prevalence effect. Is the prevalence effect the result of a criterion shift (Wolfe, et al., 2007), or is it the result of motor priming (Fleck & Mitroff, 2007)? In order to answer such a question, and to establish a theoretical foundation for the subsequent chapters, the present empirical chapter will examine the root cause of the prevalence effect. In doing so, it will be argued that the prevalence effect can be connected with an extant body of work relating to what is known as the *stimulus probability effect* (Estes, Burke, Atkinson, & Frankmann, 1957; Fitts, Peterson, & Wolpe, 1963). The point of connection between the prevalence and stimulus probability effects will be examined in an experiment that varies the *relative prevalence* of targets in dual-target search. Relative prevalence in dual-target search will be controlled such that one target will be presented at a higher prevalence level than the other, whilst overall target prevalence is held constant at a level of 50%. The higher-prevalence target will have a prevalence of 45%; the lower-prevalence target will have a prevalence of 5%.

The issue of relative prevalence is important to airport security screening because the true prevalence rate of real threat items is not balanced (i.e. the prevalence at which the different types of threat, such as guns, knives, or IEDs, is, of course, not equal). Indeed, the overwhelming majority of security screeners will never, throughout their entire careers, be presented with an X-ray of a passenger's bag that contains a real IED. However, since new restrictions were implemented that require passengers to not take liquids onto aircraft, screeners will often be detecting bottles of liquid (as passengers will not have become accustomed to such developments). Does the detection of one form of target that occurs frequently hamper the detection of the less-frequent target?

2.1.2 The Stimulus Probability Effect

As described earlier in the Literature Review, Fleck and Mitroff (Fitts, et al., 1963; 2007) reportedly eliminated the prevalence effect by allowing participants to 'correct' their rapid target-absent responses. The result of this was that Fleck and Mitroff (2007) argued that the prevalence effect was essentially caused by a motor priming effect. Wolfe et al. (2007) cast doubt upon such claims by reporting a series of experiments in which slowing participants' responses did not eliminate the prevalence effect, and, additionally, by finding that observers consistently missed the same targets (for more detail, refer to Section 1.3.4 of the Literature Review).

The argument between the motor priming account of Fleck and Mitroff (2007) and Wolfe et al. (2007) is interesting because it neatly mirrors an existing set of arguments regarding a related phenomenon known as the stimulus probability effect. This effect was first described by examined in a number of early studies dating back to the 1950s and 1960s (Estes, et al., 1957; Fitts, et al., 1963). Since that time, the underlying cause of the effect has been studied and debated extensively. The basics of the effect involve the finding that stimuli which are presented at a high level of prevalence are more likely to be detected correctly, and also detected more rapidly, than targets which are presented at a low level of prevalence (Erickson, 1966). Similar results are produced when prevalence is varied. Indeed, Wolfe et al. (2007) also found increased RTs to detect targets as prevalence was reduced, but made no attempts to explain such a result. What is surprising, however, is that the two separate strands of research have not been connected previously. Still, it should be noted that they are separated by many decades, as well as a markedly different lexicon.

The causes of the stimulus probability effect have been examined using a variety of different methods. For the most part, however, participants are given a set of target letters to detect, and then presented with single letters, one at a time, and required to give a response upon each presentation. The effect has been examined in terms of facilitations to perceptual and/or motor processes (e.g. Biederman & Stacy, 1974), much like the debate regarding motor priming and the prevalence effect (Fleck & Mitroff, 2007; Wolfe, et al., 2007). Various studies have examined the facilitation in terms of one of these stages independently, and often fail to examine their impact in unison. This point has been made by Gehring, Gratton, Coles and Donchin (1992), and in a more recent paradigm, they used a flanker task in conjunction with ERP data, to show that, although there was undoubtedly a motor priming aspect of the stimulus probability effect, there was also a marked perceptual effect as well.

Perhaps the most surprising and interesting point of converge between studies of the prevalence and stimulus probability effects is the fact that a criterion-shifting account of the stimulus probability effect has been suggested (Miller & Bauer, 1981). This account is similar in more than just its name to Wolfe et al.'s (2007) criterion shift account of the prevalence effect. According to the criterion-shifting account, a stronger memory trace exists for frequently-presented stimuli; this allows observers to set a lower criterion for the detection of those stimuli. Evidence for such an account comes from several experiments in which observers were presented with two targets. The key manipulation involved having one target being presented at a higher level of prevalence than the other: the lower-prevalence target was often missed by observers, and when detected, was detected less rapidly than the higher-prevalence target. However, this effect was eliminated when the targets were similar in appearance to one another (Dykes & Pascal, 1981; Miller & Bauer, 1981), and this resulted in the suggestion that a memory trace for the targets is used to guide responses, with a stronger memory trace resulting in a lower criterion (Miller & Bauer, 1981).

Despite the clear similarities between the prevalence effect and the stimulus probability effect, there are a number of marked differences between the two. First and foremost, stimulus probability experiments typically involve no visual search whatsoever: a single stimulus is presented (usually a letter), and the participant then responds to that single stimulus. As a result, despite the evidence suggesting that stimulus probability affects multiple processing stages, this may not be the case for visual search, or, instead, this may still be the case with visual search, but the processing stages may be impacted to a different magnitude.

2.1.3 Research Questions: Dual-target Search, Relative Prevalence, and Stimulus Probability

The goal of the present study, therefore, is to carry out a rigorous test of the stimulus probability effect using the dual-target search paradigm (Menneer, et al., 2007). In order to examine the role of motor priming in the prevalence effect, the rubric of the dual-target search paradigm will be used to modify a study conducted by LaBerge and Tweedy (1964). Participants were given two target letters to search for, and responded using one button when a green stimulus was presented, and a different button when either a red or a blue stimulus was presented. The red or blue stimuli were treated as targets. Crucially, one of the target stimuli were presented at a higher level of prevalence than the other. As participants were required to make a single motor response using a single button for *any* target-present trial, and overall target presence was held at 50%, motor priming was held constant. The motor priming account would have predicted that, because motor priming was held constant for both targets, there should have been no stimulus probability effect. However, this was not the case: the stimulus probability effect was still present.

A similar approach has been adopted by Wolfe et al. (2007, experiments 3 and 4), where participants were instructed to search simply for 'tools', and the actual prevalence of each type of tool was varied to be either 1%, 5%, 10%, or 34% prevalence (e.g. drills could appear on 1% of trials, whilst hammers appeared on 10% of trials, and so on). Overall, target prevalence was kept at a constant 50%, as with the present study. The results indicated that miss rates increased as prevalence decreased. Although Wolfe et al. (2007) did not discuss their findings in such terms, the effects found by varying prevalence replicate the studies conducted by LaBerge and Tweedy (1964). Indeed, this is strong evidence *against* a purely motor-priming account of the prevalence effect, suggested by Fleck and Mitroff (2007). As the participants in the Wolfe et al. (2007) study were pressing only a single response button for 'present', if the prevalence effect was entirely caused by motor priming, then the prevalence effect should not have been found when multiple targets were presented at varying prevalence levels.

Although experiments 3 and 4 of Wolfe et al. (2007) employed a design comparable to previous stimulus probability studies, there are a number of problems with their design and methodology. They failed to employ adequate control conditions: participants in multiple-target search were asked to detect 'tools', which were further broken down into four categories of different prevalence. However, the control conditions of single-target search involved only the search for a single target (e.g. a hammer) with a prevalence level of either 1% or 50%, rather than matching the various prevalence levels to the different search target categories. Additionally, no attempts were made to assess the presence of a dual-target cost. Signal detection analyses were not possible in their experiment because false alarm rates were very low. Finally, they did not report RT data, making it somewhat difficult to compare their results to the studies examining stimulus probability. It is important that the dual-target cost is taken into account when using a design of this nature, because changes in performance could be mistakenly attributed to prevalence effects, when in fact the performance changes are really a dual- (or in the case of experiments 3 and 4 or Wolfe et al., 2007, multiple-)target search costs for response accuracy.

The present study will therefore attempt to replicate the findings of Wolfe et al. (2007, Experiments 3 and 4), and LaBerge and Tweedy (1964). However, the key development here will be that single- and dual-target search will be considered. It is important that the dual-target cost is taken into account when using a design of this nature, because changes in performance could be mistakenly attributed to prevalence effects, when in fact the performance changes could be due to a dual- target search cost for response accuracy (indeed, similar concerns regarding this were also voiced by Wolfe, et al., 2007).

The previous studies failed to consider the role that the dual-target cost might have in the analyses, so it is considered in detail here. Thus, the experimental design used in the previous studies will be adapted and improved in order to cater for the dual-target cost. The problems of the dual-target cost will be built into the experimental design from the outset. In both single- and dual-target search, a target will be present on 50% of all trials; in dual-target search, the central manipulation will be that, in dual-target search, one target will be presented on nine times as many trials as the other. In other words, there will be a target present on 50% of all dual-target trials, but 45% of the dual-target trials will involve the presentation of one target, whilst 5% of the dual-target trials will involve the presentation of the second target. These relative prevalence conditions can then be compared to a standard dual-target condition where the targets appear with an equal prevalence to one another (i.e. one target present on 25% of trials, the other present on the remaining 25% of trials) to assess whether or not there is a relative prevalence effect. This design allows for a measure of the dual-target cost for both targets when prevalence is equal, and subsequent comparisons in performance when relative prevalence is varied.

2.2 Method

2.2.1 Participants

Eighteen participants (five males and thirteen females) took part in the study, with ages ranging from 19 to 53 (mean=24.5 years, SD=10.5 years). All participants were undergraduates and postgraduates, and reported normal colour vision and no previous experience with the stimuli. Participants received course credit or payment for their participation. All participants completed the study within 30 days.

2.2.2 Apparatus

The experimental software was produced using the VisionShell libraries, and was run on an Apple Macintosh G4, which presented the stimuli on a Formac ProNitron 19/600, with a refresh rate of 75 Hz and a resolution of 1600x1200 pixels. Participant responses were given using a Cedrus RB-610 button box connected via the USB port, with buttons labelled “present” and “absent”. Viewing distance was around 60cm from the monitor. The experiment took place in a moderately-lit room.

2.2.3 Stimuli

The stimuli were X-ray images of threat and non-threat items. Targets consisted of X-ray images of metal threats (guns and knives) and IEDs. Distractors consisted of X-ray images that one would normally expect to see in baggage, such as keys, sunglasses, shoes, children's toys, and so on. The objects were photographed in up to five orientations, consisting of a canonical view and rotations through 45˚ and 90˚ in both the × and y planes. In total, 200 metal images were used (100 guns and 100 knives), as well as 69 IED images, and 1302 distractor images.

In each trial, the search field contained a total of twelve separate objects. On target-present trials, only one target image was presented (even in dual-target search), resulting in one target image and eleven distractor images. Targets and distractors were selected from the image library at random, and then randomly rotated by 0˚, 90˚, 180˚, or 270˚ through the plane of the monitor screen. The objects were randomly placed on a virtual 4 × 4 grid laid out across the display. Objects were moved from the centre of each square in the grid by a randomly generated distance, in a randomly generated direction. The images were displayed in 32-bit colour, and subtended 0.5-7.0˚ of visual angle. Each square of the virtual grid subtended 8.73˚ by 8.70˚ of visual angle, with the whole display subtending 26.2˚ by 34.8˚ of visual angle.

2.2.4 Design and Procedure

Participants took part in four sessions, each lasting around 45 minutes. Before the trials began, a lengthy explanation was given concerning the nature of the targets, during which participants were guided through twenty examples of each type of threat item that they were to search for.

Each session was blocked into three different sets of trials: single-target search for metals, single-target search for IEDs, and dual-target search for metals and IEDs. Each block began with five practice trials, followed by 160 experimental trials (giving rise to 480 trials overall per session). Participants were able to take a break every 50 trials. All sessions were identical, with the exception of the training given in the practice session. The order of the blocks was counter-balanced across participants.

The study used a mixed design, with three independent variables, consisting of: Target type (metals, IEDs, absent); Search type (single-target search, dual-target search) and the dual-target prevalence condition (Ratio). The Target type and Search type factors were within-subjects factors; the Ratio factor was a between-subjects factor. In both single- and dual-target search, a target was present on 50% of trials. The dual-target prevalence condition described the relative prevalence of the two target classes in dual-target search, and consisted of *High-Prevalence Metals / Low-Prevalence IEDs* (HP-metals), *High-Prevalence IEDs / Low-Prevalence Metals* (HP-IEDs), and *Equal Prevalence* (EP). In the HP-metals and HP-IEDs conditions, the higher-prevalence target appeared on nine times as many trials as the lower-prevalence target. So, in HP-metals dual-target search, a metal target was presented on 45% of trials, and an IED was presented on 5% of trials. Conversely, in HP-IEDs dual-target search, an IED was presented on 45% of trials, whilst a metal was presented on 5% of trials. In the EP condition, targets appeared at an equal prevalence to one another, with, in dual-target search, a metal being presented on 25% of trials, and an IED being presented on 25% of trials, giving rise to an overall target prevalence of 50%. The dependent variables were response accuracy and response time.

Each trial began with the appearance of a small fixation cross at the centre of the display, followed by the presentation of the search field. There were two possible responses from the participants in any trial: “present” or “absent”. The search field remained visible until the participant made a response, which ended the current trial and began the next. When a participant gave an incorrect response, an audible tone was produced by the computer. Only one target could appear on any trial.

2.3 Results

In the following results, and for all empirical chapters within the present thesis, all *t*-tests have had their *p* values Bonferroni-corrected before being reported; additionally, the Greenhouse-Geisser *F* values, degrees of freedom, and *p* values are reported for repeated-measures ANOVA results wherever tests of sphericity are violated (i.e. Mauchly's test of sphericity shows a p value of less than .05). In all figures, error bars represent ±S.E.M.

Previous explorations of the prevalence effect have focused on signal detection measures (Wolfe, et al., 2007), such as the criterion (c) and sensitivity (d'): these measures are not computed here, because, for dual-target search, participants would have had to report which of the targets they believed to be present whenever they gave a 'target-present' response. This would have then enabled the computation of a measure of how biased participants were to respond to each of the targets. However, using such a design would have had the unfortunate consequence of modifying motor priming to not be constant across the two targets when using variations in relative prevalence, thereby rendering the paradigm unable to test the experimental hypotheses.

2.3.1 Impact of Relative Prevalence: Target-Present Trials

When prevalence was varied in dual-target search, observers shifted their performance to detect the higher-prevalence target more often than the lower-prevalence target, in a clear replication of previous experiments (LaBerge & Tweedy, 1964; Wolfe, et al., 2007). In other words, the dual-target cost was shifted to the lower-prevalence target. This effect can be seen in Figure 2.3.1a: error rates were examined using a 4 (Session: 1,2,3,4) × 2 (Target Type: Metals, IEDs) × 2 (Search Type: Single-target, Dual-target) × 3 (Ratio: HP-Metals, HP-IEDs, EP) ANOVA, with Ratio entered as a between-subjects factor. Note that here, the Target Type factor has only two levels, because the target-present trials are being examined alone. There were no main effects of either the Ratio or Target factors (both Fs<1), yet there was a significant interaction between Ratio and Target ($F(2,15)=5.7$, $p<.05$), a dual-target cost ($F(1,15)=21.9$, $p<.01$), an interaction between Ratio, Target and the dual-target cost ($F(2,15)=6.4$, $p<.05$), and an interaction between Session and Target ($F(3,45)=7.2$, $p<.01$). No other effects or interactions reached significance (Fs<1).

A set of t-tests explored the significant interactions in detail. These revealed that the dual-target cost was shifted to the low-prevalence target when prevalence was varied (i.e. there was no dual-target cost for the high-prevalence target, but there was for the low-prevalence target). Indeed, in HP-IEDs, there was a dual-target cost for metals ($t(23)=2.9$, $p<.05$), but not for IEDs ($t(23)=0.8$, $p>.05$).

Conversely, in HP-Metals, there was a dual-target cost for IEDs ($t(23)=3.1$, $p<.05$), but not for metals ($t(23)=0.9$, $p>.05$). However, as with previous experiments, when the targets appeared with an equal level of prevalence in dual-target search, there was a dual-target cost for both metals ($t(23)=5.3$, $p<.05$), and IEDs ($t(23)=5.0$, $p<.05$).

A further set of *t*-tests examined the interaction between Session and Target. These indicated that, in the first session, error rates for metals were significantly lower than those for IEDs ($t(23)=4.1$, $p<.05$); however, for all of the three remaining sessions, this was not the case (Session 2: $t(23)=1.9$, $p>.05$;Session 3: $t(23)=0.3$, $p>.05$; Session 4: $t(23)=1.4$, $p>.05$).

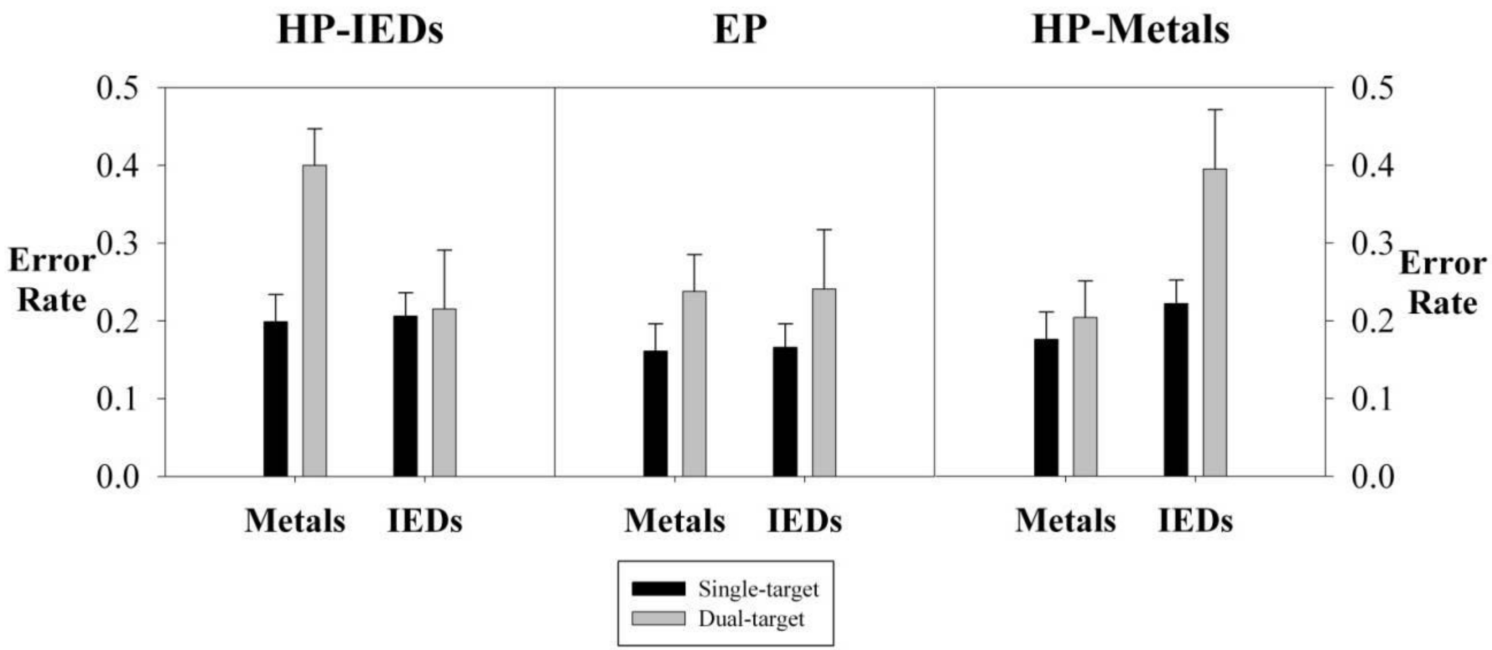


Figure 2.3.1a. Miss Error rates as a function of Ratio condition, search target (Metals or IEDs), and Single- or Dual-target Search

2.3.2 Impact of Relative Prevalence: Target-Absent Trials

In previous examinations of the prevalence effect, participants often miss targets because they become *biased* against target detection. There is typically an increase in miss error rates, but, at the same time, correct rejection rates tend to increase as well: in other words, this is a *criterion shift* (Wolfe, et al., 2007). Did the participants here show any shift in their performance in target-absent trials?
To answer such a question, a 4 (Session: 1,2,3,4) × 3 (Search Type: Single-target Metals, Single-target IEDs, Dual-target Search) × 3 (Ratio: EP, HP-Metals, HP-IEDs) ANOVA was conducted, with Ratio used as a between-subjects factor. This revealed a main effect of Session (F(3,45)=25.8, p<.001), and of Search Type (F(2,30)=10.7, p<.001), but no effect of Ratio (F<1), nor were there any other effects or interactions. As can be seen below in Figure 2.3.2a, false alarm rates on target-absent trials were lower for single-target metals than either single-target IEDs (t(71)=3.2, p<0.05), or dual-target search (t(71)=5.9, p<0.05), whilst error rates for target-absent trials in dual-target search and single-target IEDs did not differ (t(71)=1.8, p>0.05). Additionally, as the sessions progressed, false alarm rates were reduced for target-absent trials, reaching a plateau in the third session (Session 1 compared to Session 2: t(53)=4.9, p<0.05; Session 2 versus Session 3: t(53)=2.6, p<0.05; Session 3 versus Session 4: t(53)=0.4, p>0.05), see Figure 2.3.2b, below.

Given that the Ratio condition did not have any impact upon the error rates for the target-absent trials, it is clear that the Ratio manipulation in dual-target search impacted target detection rates only, and left the target-absent trial response rates unaffected.

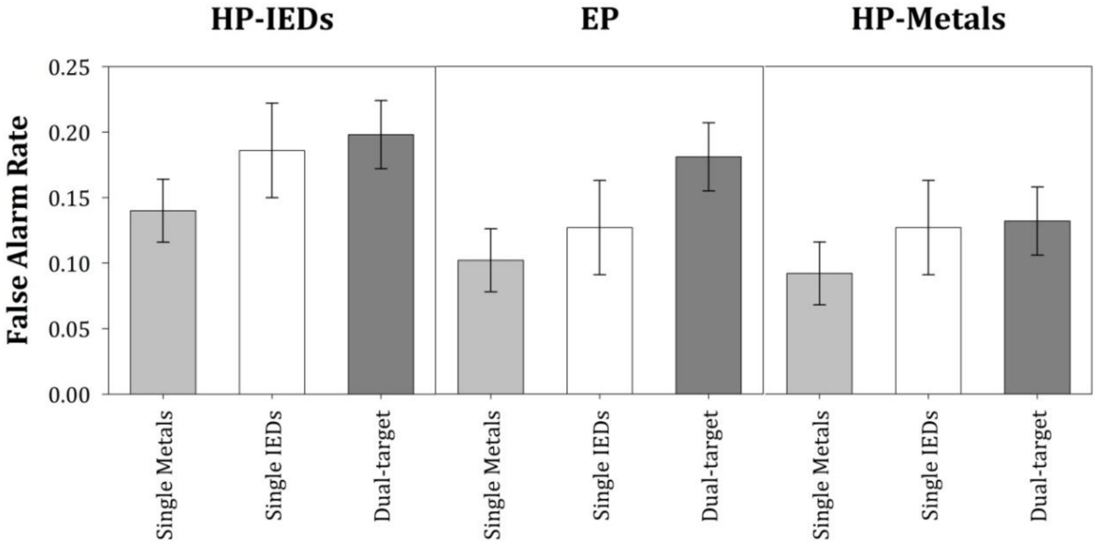


Figure 2.3.2a. False Alarm rates as a function of Search Type (single-target metals, single-target IEDs, or dual-target search) for each of the different Ratio conditions.

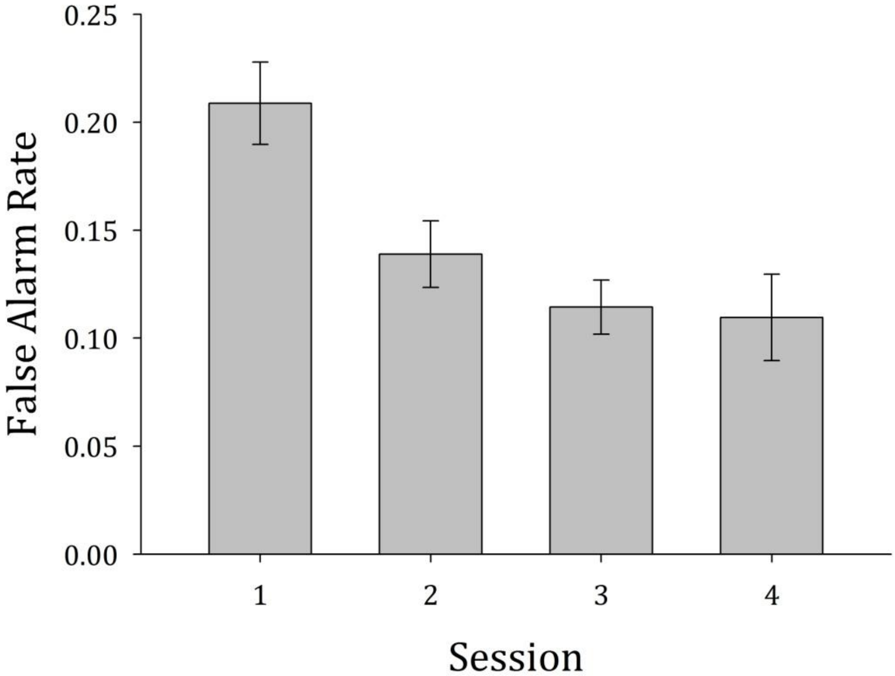


Figure 2.3.2b. False Alarm rates in each of the four Sessions. Note that these are the *overall* false alarm rates, across both single- and dual-target search.

2.3.3 Further Tests of the Stimulus Probability and Prevalence Effects: Reaction Times

Examination of the RT data confirmed that, as with studies of stimulus probability, the higher-prevalence target was detected more rapidly than the lower-prevalence target (LaBerge & Tweedy, 1964), as seen in Figure 2.3.3a, broken down by the Session factor. The RT data were examined through the use of a 4 (Session: 1,2,3,4) × 2 (Search Type: Single-target, Dual-target) × 2 (Target: Metals, IEDs) × 3 (Ratio: HP-Metals, HP-IEDs, EP) ANOVA. Ratio was entered as a between-subjects variable. This ANOVA revealed a significant main effect of Search Type ($F(1,15)=36.7$, $p<.001$), a main effect of Session ($F(3,45)=36.4$, $p<.001$), and interactions between Target and Ratio ($F(2,15)=17$, $p<.001$), between Search Type and Target ($F(1,15)=11.7$, $p<.01$), and between Target, Ratio and Search Type ($F(2,15)=7.6$, $p<.01$). The Session factor also interacted with Search Type ($F(3,45)=6.3$, $p<.01$), and with Target ($F(3,45)=8.4$, $p<.01$).

Examining RTs within each Ratio condition using t-tests, it was found that, in EP, there was a dual-target cost in detection time for both metals ($t(23)=3.8$, $p<0.05$), and IEDs ($t(23)=5.1$, $p<0.05$). Additionally, in dual-target search, metals were detected more rapidly than IEDs ($t(23)=4.5$, $p<0.05$), whilst this was not true of single-target search ($t(23)=1.2$, $p>0.05$). For HP-IEDs, however, there was a slight dual-target cost for IEDs ($t(23)=2.8$, $p<0.05$), as well as a stronger dual-target cost for metals ($t(23)=3.1$, $p<0.05$). Metals were detected less rapidly than IEDs in single-target search ($t(23)=5.9$, $p<0.05$), but this was not the case in dual-target search ($t(23)=5.9$, $p>0.05$). Finally, for HP-Metals, there was no dual-target cost for metals ($t(23)=1.6$, $p>0.05$), yet there was a dual-target cost for IEDs ($t(23)=6.1$, $p<0.05$). In single-target search, metals and IEDs were detected with equal speed ($t(23)=2.0$, $p>0.05$), whilst in dual-target search, metals were detected more rapidly than IEDs ($t(23)=5.8$, $p<0.05$).

As the sessions progressed, overall RTs were reduced, for both single- and dual-target search. Examining the Session × Search Type interaction, a set of t-tests examined the dual-target cost for each session separately. These revealed the presence of the dual-target cost over all four sessions (Session 1: $t(35)=5.2$, $p<0.05$; Session 2: $t(35)=4.7$, $p<0.05$; Session 3: $t(35)=4.3$, $p<0.05$; Session 4: $t(35)=3.7$, $p<0.05$). Additionally, they also revealed that, for single-target search, RTs decreased between Sessions 1 and 2 ($t(35)=6.3$, $p<0.05$), and between

Sessions 2 and 3 ($t(35)=3.3$, $p<0.05$), but not between Sessions 3 and 4 ($t(35)=2.6$, $p>0.05$). However, for dual-target search, RTs decreased between Sessions 1 and 2 ($t(35)=4.8$, $p<0.05$), but not between Sessions 2 and 3 ($t(35)=1.1$, $p>0.05$) or Sessions 3 and 4 ($t(35)=1.7$, $p>0.05$).

Examining the Session × Target interaction, it was found that, overall RTs for metal detection rates were faster in Session 1 than detection rates for IEDs ($t(35)=4$, $p<0.05$); yet this was not significant for the remaining sessions (Session 2: $t(35)=1.7$, $p>0.05$; Session 3: $t(35)=0.7$, $p>0.05$; Session 4: $t(35)=0.1$, $p>0.05$). Additionally, RTs to detect metal targets decreased between Sessions 1 and 2 ($t(35)=3$, $p<0.05$), but not between the remaining sessions (Session 2 versus 3: $t(35)=1$, $p>0.05$; Session 3 versus 4: $t(35)=1.1$, $p>0.05$). The same was true for IEDs, with detection speed becoming faster between Sessions 1 and 2 ($t(35)=5.6$, $p<0.05$), and reaching a plateau thereafter (Session 2 versus 3: $t(35)=1.7$, $p>0.05$; Session 3 versus 4: $t(35)=2$, $p>0.05$).

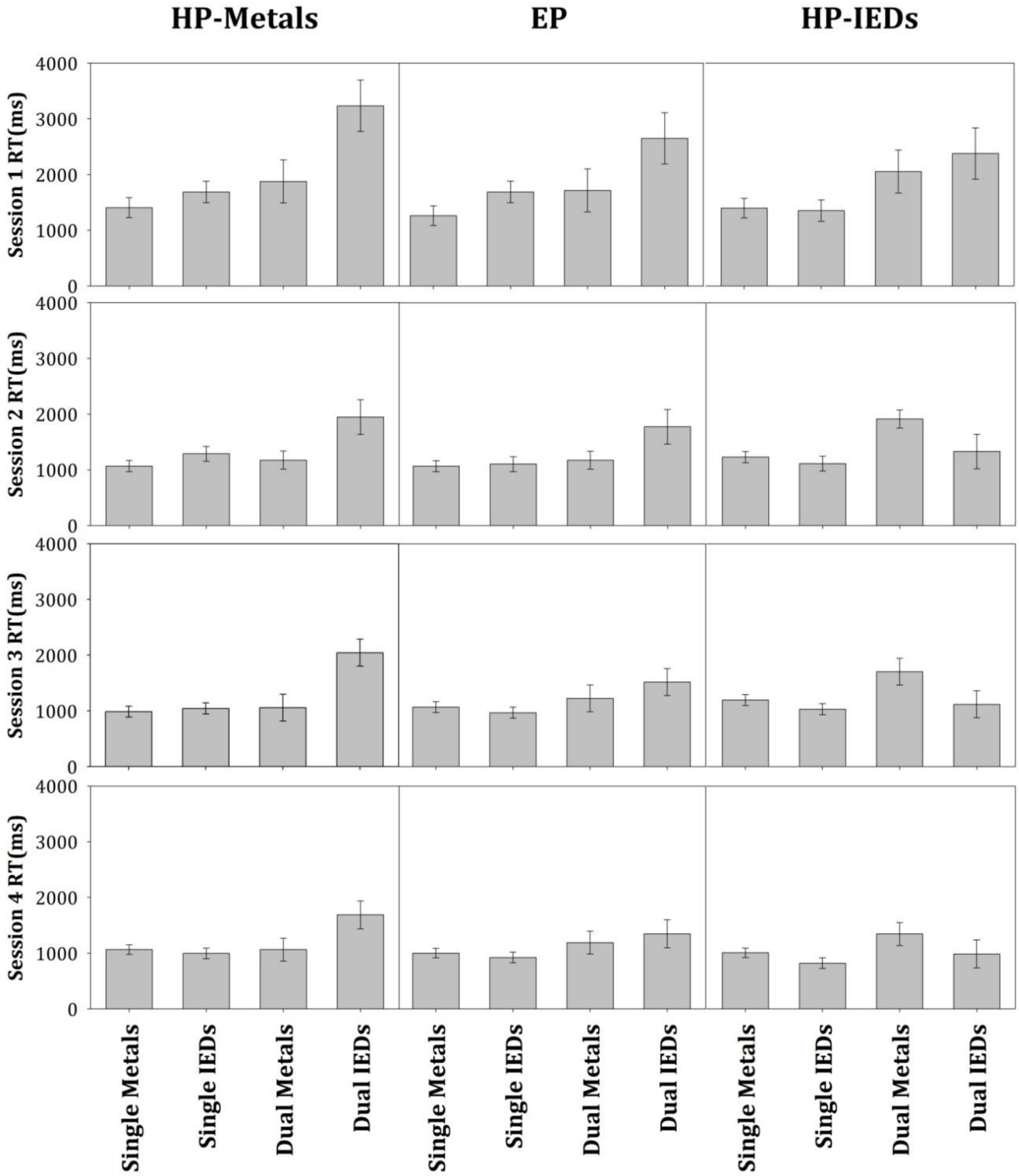


Figure 2.3.3a. Reaction times for the different search types and Relative Prevalence conditions broken down as a function of the different Sessions, for correct responses to target-present trials only.

2.3.4 Were Lower-Prevalence Targets Missed due to Rapid 'Absent' Responses?

In previous examinations of the prevalence effect, it has been reported that target-absent responses typically become very rapid. Thus, the target-absent response times were examined in order to determine if the lower-prevalence targets in dual-target search were missed as a result of rapid 'absent' responses. Unlike previous examinations of the prevalence effect, it was apparent that the target-absent responses in the present study were not rapid: in fact, they did not vary across the different Ratio conditions.

Target-absent trials were examined using a 4 (Session: 1,2,3,4) × 3 (Ratio: HP-Metals, HP-IEDs, EP) × 3 (Search Block: Metals, IEDs, Dual-target) ANOVA, with Ratio entered as a between-subjects factor. As can be seen clearly in Figure 2.3.4a, below, all main effects and interactions involving the Ratio factor failed to reach significance (all Fs<1), implying that target-absent RTs were not more rapid when the relative prevalence of the targets was varied. However, there was an impact of Search Block ($F(2,30)=14.5$, $p<.001$), as well as of Session ($F(3,45)=34.5$, $p<.001$). A set of t-tests revealed that dual-target search had slower target-absent responses than either metals ($t(71)=5$, $p<0.05$) or IEDs ($t(71)=7.6$, $p<0.05$); additionally target-absent RTs for IEDs than for metals ($t(71)=2.8$, $p<0.05$). Examining the impact of the Session factor revealed that target-absent RTs decreased between Sessions 1 and 2 ($t(53)=8.6$, $p<0.05$), as well as between Sessions 3 and 4 ($t(53)=1.5$, $p<0.05$), but not between Sessions 2 and 3 ($t(53)=1.6$, $p>0.05$). Figure 2.3.4b presents the RTs for target-absent trials as a function of the Session factor. Given the lack of any impact of the Ratio factor, it appears that the effects seen in the error rates for the varied-prevalence targets was not due to rapid target-absent responses when prevalence was varied in dual-target search.

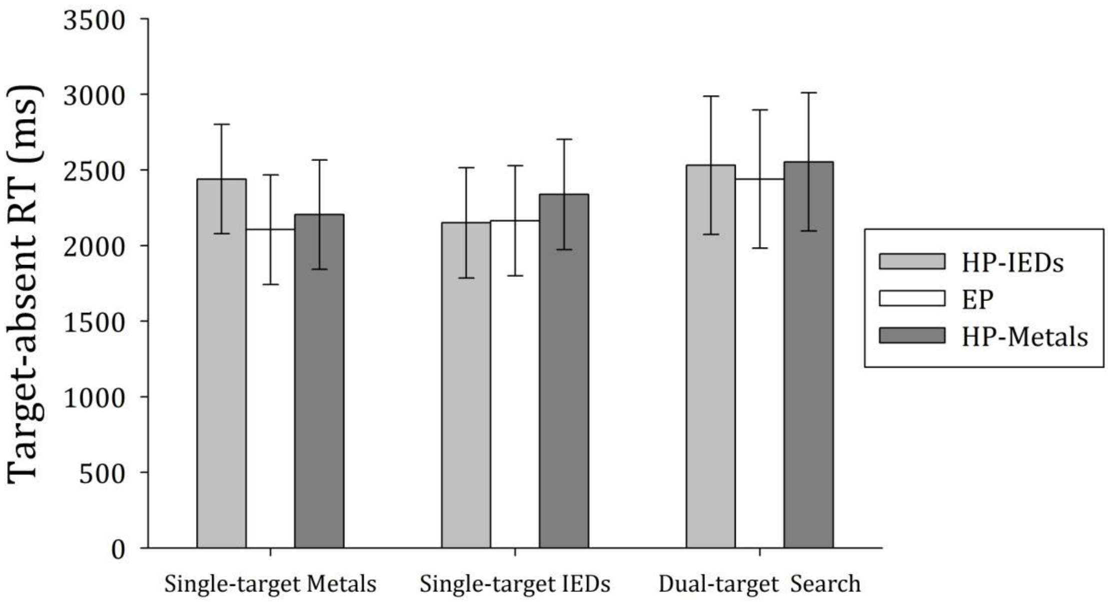


Figure 2.3.4a. Target-absent Reaction Times in each of the different Ratio conditions, for Single-target Metals, Single-target IEDs, and Dual-target Search.

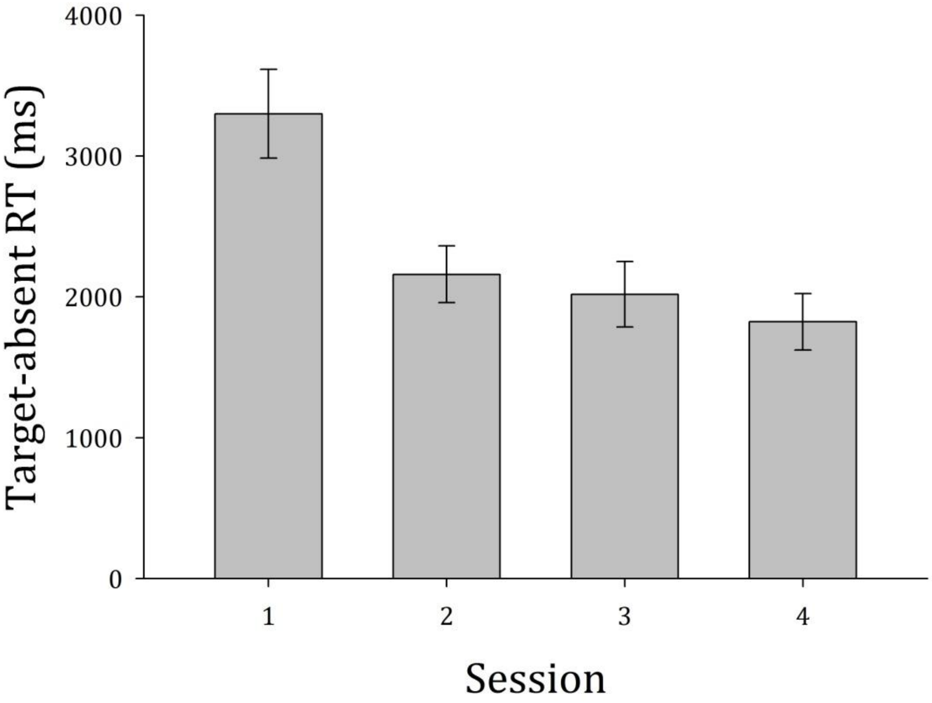


Figure 2.3.4b. Target-absent Reaction Times in each of the four Sessions. Note that these are the *overall* reaction times, across both single- and dual-target search.

2.4 Discussion

The purpose of the present study was to explore the relationship between the prevalence effect, the stimulus probability effect, and the dual-target cost in an experiment that varied the relative prevalence of targets in dual-target search. When targets appeared at an equal prevalence to one another in dual-target search, they were both impacted by the dual-target cost. When one target was presented at a higher level of prevalence than the other, the dual-target cost was eliminated for the higher-prevalence target, and amplified for the lower-prevalence target. There was also a general pattern of RTs for the lower-prevalence target being slowed in comparison to single-target search.

2.4.1 Using the Dual-target Cost to explore the Relationship between the Effects of Prevalence and Stimulus Probability

The predictions from the stimulus probability studies have been replicated (Erickson, 1966; LaBerge & Tweedy, 1964; Miller & Bauer, 1981): detection rates for higher-prevalence targets were higher than those of lower-prevalence targets, and, furthermore, RTs for higher-prevalence targets were more rapid than those to lower-prevalence targets. One caveat in regards to the RTs is that IEDs showed an overall tendency towards slower detection speeds than metals. Indeed, even in EP, dual-target IED RTs were slower than those for metals. This is not really very surprising. IEDs are notoriously difficult targets to detect, and are highly complex; additionally, participants will enter the study with no prior knowledge of the nature of IEDs, yet will have seen images of guns and knives before taking part. Still, in HP-IEDs, the metals showed a stronger dual-target cost for RTs than IEDs, whilst in HP-Metals, metal detection RTs were faster than those for IEDs.

The results presented here can also be used to echo the argument made by Wolfe et al. (2007) against a motor priming account of the prevalence and stimulus probability effects (Fleck & Mitroff, 2007), because participants were using a single 'present' button to respond to the presence of a target. If the prevalence effect was entirely caused by motor priming, as argued by Fleck and Mitroff (2007), then, here, when, participants were equally primed to respond 'present' as they were to respond 'absent', there should have been no prevalence effect whatsoever. However, this was not the case: there was a marked increase in the error rates for the lower-prevalence target in dual-target search. It is rather clear, therefore, that the prevalence and stimulus probability effects are intricately related to one another. This is a novel connection between two related strands of literature; indeed, it is rather surprising that studies of the prevalence effect have not previously considered the extant stimulus probability literature.

2.4.2 General Discussion and Relevance to Airport X-ray Security Screening

The memory trace account taken from the stimulus probability research (Miller & Bauer, 1981) is a viable explanation of the results in the present study. If detections of the higher-prevalence target resulted in a stronger memory trace for that target, it is easy to see how that target would then be detected rapidly, and exhibit high hit rates, whilst the lower-prevalence target would often be missed and detected less rapidly. A logical corollary that would follow from the memory trace account, and strengthen support for it, would be that targets that are similar in appearance, yet presented at different relative prevalence levels, would have a reduced effect of target prevalence or stimulus probability. Indeed, in two separate experiments, this has already been reported (Dykes & Pascal, 1981; Miller & Bauer, 1981).

It would be of interest to determine whether or not observers in the varied-prevalence conditions are still examining irrelevant distractors in the displays, as is seen in typical examinations of the dual-target cost (Stroud, et al., in preparation). Given that RTs to the higher-prevalence target were equal to those of single-target search, it seems plausible to suggest that, in actual fact, observers were no longer scanning irrelevant distractors in the display, or else they would not have been detecting the higher-prevalence targets so rapidly. It is an open question as to whether or not the lower-prevalence targets were scanned *later* in the search process, or whether they simply required more time to examine and process (i.e. was it a strategic decision to look for higher-prevalence targets first or if it was perceptual effect only, vis-à-vis criterion shifting). Future studies involving eye movements would be very revealing on this issue. Furthermore, it would be of interest to examine the time needed to identify a target once the eyes had fixated upon it. Criterion-shifting accounts (Miller & Bauer, 1981; Wolfe, et al., 2007) would predict that more time would be needed, and this seems plausible given the available data.

In terms of relevance to airport security screening, a number of salient points can be drawn from the present results, as well as the previous research regarding the stimulus probability effect. First of all, it is apparent from the experiment conducted here that the dual-target cost can not be eliminated: it can only be shifted by varying the relative prevalence of the targets. One could argue that the cost could be eliminated following extensive practice. This does not seem to be the case: Menneer, Cave and Donnelly (under review) engaged participants in eleven sessions of visual search, each containing several hundred trials and lasting around one hour, and reported that the dual-target cost was not eliminated, even in the final session. The impact of variations in relative prevalence have some important implications for real screeners, because the relative prevalence levels of real threat items is not balanced in any way. As already noted, a large number of bottles and liquids now have to be removed from passenger baggage, in line with changes to procedures and regulations. On the surface, one might expect this to cause real problems for the detection of actual threat items, because, based upon the present results, it could be predicted that screeners will focus more on the search for bottles and liquids, at the expense of the other targets. However, given the finding that the stimulus probability effect is attenuated when targets of varying prevalence are presented when those targets have a similar appearance to one another, it may actually be the case that the high prevalence rate of bottles and liquids actually *increases* the probability that a screener will detect a real IED, should one actually be presented. This is because liquids, being organic in nature, have a similar appearance to the explosive components of real IEDs (see Appendix A). Still, it is important to remember that, in such situations, the ability to detect other targets (e.g. metal threat items, such as guns and knives, which are of different colour and shape to both liquids and IEDs) may be impaired.

3

The Dual-target Cost across a variety of Prevalence Levels

Extending the Criterion Shift Account of the Prevalence Effect

3.1 Introduction

In the previous chapter, the relationship between the dual-target cost and prevalence effect was explored in an experiment where overall target prevalence was capped at 50%. However, no previous studies have examined how the dual-target cost is impacted as *overall* prevalence itself is varied. Furthermore, although previous examinations of the prevalence effect have explored the effect in terms of *low* prevalence, there have been no attempts made to understand *high* prevalence (i.e. when a target is presented on more than 50% of trials).

If, as was suggested in the previous chapter, the prevalence effect is the result of a criterion shift in responding (Wolfe, et al., 2007), then the effect should also be extensible to conditions in which a target is presented very high level of prevalence indeed, with the criterion shifting to become more liberal as prevalence increases. In other words, in a high prevalence scenario, participants should become biased towards responding 'present', thereby increasing the hit rate, whilst also increasing the false alarm rate (Macmillan & Creelman, 2005). Thus, the goal of the present chapter is to explore two new areas of interest that have not previously been investigated: first, the manner in which the prevalence effect and dual-target cost interact (if at all), and second, the extent to which the prevalence effect can be extended to account for conditions of high prevalence.

The present chapter is intended as a companion to the next chapter. Together, the two chapters examine a range of prevalence levels and explore the prevalence effect and dual-target cost in detail. Here, the investigation will begin by initially covering a relatively narrow range of prevalence levels (prevalence levels of 20%, 50%, and 80%). In the following chapter, the results will be extended using a second experiment which examines a much broader range of prevalence levels (prevalence levels of 2%, 24%, 50%, 76%, and 98%).

3.1.1 Research Questions I: Specific Tests of the Criterion Shift Account

By varying overall prevalence levels, the experiment that will be detailed in the present chapter will enable some further tests of the *criterion shift* account of the prevalence effect (Wolfe, et al., 2007). As already noted, thus far, studies of prevalence have only really been concerned with single-target search for targets appearing on between 2% and 50% of trials (Fleck & Mitroff, 2007; Wolfe, et al., 2005; Wolfe, et al., 2007). If it is the case that the prevalence effect can be explained in terms of a criterion shift, then high prevalence should see the criterion continuing to shift towards a more liberal position (see Appendix B for further details on Signal Detection Theory and criterion shifts). The criterion shift account was described briefly in the preceding Literature Review, but will be examined in detail here.

In a series of experiments, Wolfe et al. (2007) found that observers exhibited a higher level of sensitivity (measured by *d′*), as well as a higher criterion (measured by *c*) in low prevalence (2% prevalence), when compared to higher prevalence levels (50% prevalence). The concomitant increase in sensitivity and criterion should not normally occur if the standard assumptions of Signal Detection Theory are met (see also Appendix B). Typically, Signal Detection Theory assumes that the noise and variance distributions that are being used to make a decision are normally distributed. If *d′* is varying alongside *c*, then this implies that one or both of the distributions are *not* normally distributed (Macmillan & Creelman, 2005; Wickens, 2002). If that is the case, then it can not be said that true sensitivity is varying as the criterion varies: the apparent change in sensitivity can then be ascribed to a quirk of the mathematics of Signal Detection Theory.

One very simple method that can be used to test just such a possibility is to plot the *z*-transformed hit and false alarm rates upon a Receiver Operating Characteristic curve (as the rates are z-transformed, this is called a *zROC* curve). If the Signal and Noise distributions are both Gaussian in nature, then the slope of the zROC curve will be equal to one. In line with this, Wolfe et al. (2007) found that, contrary to the assumptions of Signal Detection Theory, plotting variations in prevalence produced a zROC curve with a slope of around 0.6. However, in order to generate their zROC curve, Wolfe et al. (2007) only used two prevalence levels: 2% and 50%, in three separate experiments using similar stimuli, and detected a strong agreement in the zROC curve slope across the experiments. Given that these levels do not extend beyond 50%, it is unclear whether or not a zROC curve plotted between prevalence levels of 20%, 50% and 80% will still exhibit a slope of around 0.6.

Finally, Wolfe et al.'s (2007) criterion shift account contains the claim that, in order to set the position of *c*, observers attempt to equate their raw number of miss errors with their raw number of false alarm errors. Such a claim can be easily tested. Taken as a direct theory of the processing of visual stimuli, the criterion shift account essentially argues that observers require less evidence to suggest that a target is present before actually responding 'present'. In Miller and Bauer's (1981) shifting-criterion account of the stimulus probability effect, which was shown in the previous chapter to be related to the prevalence effect, the threshold for detecting highly probable stimuli is lowered, when compared to infrequently-presented stimuli. As noted in the previous chapter, in experiments where one target is presented with a higher level of prevalence than another, there is no stimulus probability effect if those targets are similar in appearance (Dykes & Pascal, 1981). Additionally, when the visibility of the stimuli has been reduced by lowering the contrast of the display, participants are still able to detect frequently-presented stimuli, yet there is a significant decrease in the detection rates of infrequently-presented stimuli (Miller & Pachella, 1973). Thus, when visibility is reduced, a target can still be detected if less evidence is required to detect that target: this is the embodiment of a criterion shift.

In order to provide both a replication and extension of the criterion shift account of the prevalence effect (Wolfe et al., 2007), a number of questions will be addressed in the present chapter:

1. Can the finding that decreases in target prevalence result in an increase in *d'*, coupled with an increase in *c* be replicated?
2. Do observers operate upon a zROC curve with a slope of around 0.6 using a different set of stimuli to that used by Wolfe et al. (2007)?
3. If so, can this be extended to high levels of prevalence?
4. Furthermore, is this pattern replicated in dual-target search?

3.1.2 Research Questions II: The Dual-target Cost from a Signal Detection Perspective

Given the use of Signal Detection Theory in the criterion shift account of the prevalence effect, how might Signal Detection Theory be used to explore the dual-target cost? Previous studies of the dual-target cost have typically focused on response accuracy rates (i.e. hit and correction rejection rates only), and have not considered the dual-target cost in terms of Signal Detection Theory (D. M. Green & Swets, 1966). To pre-empt the discussion below, in the present chapter, and those that follow, dual-target search is envisaged as being less *sensitive* (i.e. showing a reduction in *d'*) when compared to single-target search, with no change in criterion (*c*).

The Unidimensional Signal Detection Approach: The core problem with understanding dual-target search performance from a signal detection perspective is one of representation. Consider the standard interpretations of Signal and Noise distributions (see Appendix B). In the early experiments involving signal detection, such as involving judgements of the length of a line, where the experimenter controls the Signal and Noise distributions in an *a priori* manner, it is easy to see how these distributions can be represented internally by an observer (Creelman, 1965; Creelman & Donaldson, 1968). Problems begin to arise, however, when dealing with more complex stimuli, primarily because it is not clear how those stimuli are varying within the decision space (i.e. the overall dimension upon which the decision is being made). One could argue that, for example, a gun should be represented along a number of dimensions: colour, shape, and orientation, to name but a few. Although representing the gun targets along such a set of dimensions would be interesting in some senses, it would be very difficult to achieve for practical reasons. The colour dimension may be relatively easy to discern, yet how would the shape or three-dimensional orientation of a gun be represented along a continuous dimension? What about the Noise? How would the Noise be characterised along a set of separate dimensions? Such questions are not easily answered, because control over complex real-world stimuli is somewhat difficult to achieve, and furthermore, because there is little direct insight into how the variation in those stimuli is processed internally.

Fortunately, the signal detection approach has a solution, and indeed, it could be argued that the solution is one of the approach's main strengths. Although the signal detection approach acknowledges the existence of multiple dimensions within a stimulus, it is assumed that the *overall* decision about whether or not to respond 'Signal' or 'Noise' (or 'present' or 'absent') is based upon a *single dimension* (this is known as the *unidimensional* approach: Wickens, 2002). The single dimension is based upon input from the many dimensions and features of the stimuli involved in the task. It is accepted that we may never truly have a detailed insight into all of the dimensions upon which a decision is based; instead, the ‘black box’ problem of internal processing is solved by a set of rigorously-examined mathematical assumptions, based upon the highly robust General Linear Model (Wickens, 2002).

An example from Wickens (2002) highlights the strength of the unidimensional approach. Consider an experiment in which observers are presented with a series of faces. The observers are then presented with a second series of faces, and instructed to report whether or not each face in the second series was presented as part of the first series. One approach would be to consider the various features of each face (eyes, nose, mouth, etc.) along different dimensions. This would be a *multidimensional* approach: the decision could be made using multiple criteria, one for each feature, in the various dimensions involved. One popular example of the multidimensional signal detection approach is General Recognition Theory (e.g. Ashby, 2000).

Unfortunately, the multidimensional approach is not appropriate with the stimuli used in the present thesis. As already noted, the stimuli vary considerably, and examining them in their constituent dimensions (e.g. colour, shape, form, orientation, viewpoint) is not really possible, and, furthermore, is not really of interest here. Thus, the unidimensional approach can be useful to aggregate the various dimensions on which a judgement is made into a single dimension, such as, in the example of the face task given above, a feeling of ‘familiarity’. Here, the single dimension would be along the lines of ‘target-ness’ (i.e. the degree to which an image is similar to a real target).

Finally, it needs to be noted that there are some further practical issues to consider. The mathematics of multidimensional Signal Detection Theory requires a condition in which the Signal stimuli are presented, and observers are instructed to discriminate the different signals from one another (Macmillan & Creelman, 2005). The measure of sensitivity between the two signal stimuli is then used to examine the overall pattern of the distributions at work in the decision. In the case of the stimuli used here, it is unlikely that a measure of sensitivity could be produced in such a manner. Sensitivity calculations require performance to not be at either ceiling or at floor levels, and it strains credulity to expect that asking individuals to determine whether a single image is an IED or a metal threat item will produce anything other than perfect performance, because the images are simply so different from one another.

Overall, the purpose of the signal detection approach is to understand and examine how 'present' and 'absent' responses are balanced between each other in a single task. The balance between responses is used to produce a measure of how well observers can perform a task (using sensitivity parameters, such as *d'*, which increases when either hit rates increase or false alarm rates decrease), and whether or not observers are biased towards 'present' or 'absent' responses (using measures of the criterion, such as *c*). It can thus be very useful in examining overall performance when search becomes more or less difficult (i.e. via changes in sensitivity), and when observers become biased towards or away from a particular response (i.e. 'present' or 'absent' responses).

Unidimensional Signal Detection and Dual-target Search: Extending the unidimensional approach to account for dual-target search, may, at first, seem to be somewhat complex and opaque. For example, when searching for two targets, do observers essentially create two Signal distributions, one for each of the targets? In previous research, and within the present thesis, it is, in fact, assumed that dual-target search involves the use of a lone Signal distribution, which is then compared to a Noise distribution, as with single-target search. Indeed, in the experiments upon which the criterion shift account was based, Wolfe et al. (2007) in fact required participants to search for both guns and knives (a convention also used here, to group both target types under the category of 'metals'). So, in other words, others have already assumed that, no matter how many targets are being searched for, the individual targets (despite the similarities in colour shared by guns and knives, they still differ very clearly in shape) can still be aggregated onto a single dimension. After all, the participants in the studies conducted by Wolfe et al. (2007), and those presented here, require participants to report 'present' or 'absent'. Participants are not asked to report *which* of the two targets they believe that they are detecting when a number of targets could appear in the search display.

Previous examinations of the dual-target cost have focused on increased error rates and increased RTs in dual-target search, when compared to single-target search (Menneer, et al., 2007). In signal detection terms, previous research has not examined whether or not sensitivity and criterion change in dual-target search. The criterion typically only changes when the prevalence of the target changes, or when payoff regimes are introduced (see Appendix B): in dual-target search, a target is presented on as many trials in single-target search, so therefore it can be predicted that there will be no change in criterion between single and dual-target search.

However, in terms of response sensitivity, the dual-target cost is likely to be reflected as a decrease in *d'* compared to single-target search. This is because error rates are typically higher in dual-target search than in single-target search. As noted already, any increase in error rates will result in a decrease in *d'*. Still, from a purely behavioural standpoint, there is strong evidence pointing to such a possibility already. When searching for two targets of different colours that are non-linearly separable from the distractors in the display, observers examine *irrelevant* objects that are not similar to either of the targets currently being searched for (as described in more detail in the preceding Literature Review). Given that Signal Detection Theory argues that observers gather information from a stimulus or display, and then make a decision based upon that information (Wickens, 2002), it seems reasonable to assume that, if the collection of that information is less sensitive, and adds irrelevant noise to the decision process, then the resultant decision which is made based upon that information will also be less sensitive.

Thus, a combination of effects between dual-target search and the prevalence effect can be predicted for the present study. First of all, a criterion shift will be predicted as prevalence decreases, coupled with, if the findings of Wolfe et al. (2007) can be replicated, an increase in sensitivity. Alongside such effects, it can be predicted that single-target search will be more sensitive than dual-target search, with no differences in the criterion placement between the two. Finally, it should be noted that, despite the somewhat unusual behaviour exhibited by *d'* when prevalence is varied that has been reported by Wolfe et al. (2007), this does not necessarily invalidate the use of *d'* to examine the dual-target cost. Thus far, *d'* has only been shown to produce unexpected results when prevalence is varied, and thus, comparing sensitivity within a given prevalence condition (i.e. comparing single-target search sensitivity at each prevalence level with dual-target search sensitivity at each prevalence level) is still a viable, and valuable, route to follow.

3.2 Method

3.2.1 Participants

Twenty-seven participants (four males and twenty-three females) took part in this study, with an age range of 18 to 56 (mean=21.7, SD=7.4). All participants were undergraduates and postgraduates, and reported normal colour vision, and took part either for course credit, or for payment. Participation was completed in less than 30 days for all but one participant, who completed the four sessions within 35 days.

3.2.2 Apparatus and Stimuli

The apparatus and stimuli were the same as in Chapter 2.

3.2.3 Design and Procedure

Participants took part in four sessions, with each session lasting around 45 minutes. Before the trials began, a detailed explanation was given to the participants regarding the nature and appearance of the targets, with participants being guided through twenty examples of each type of threat item that they were to search for.

The sessions were each blocked into three different sets of trials: single-target search for metals, single-target search for IEDs, and dual-target search for metals and IEDs. The blocks began with five practice trials, followed by 160 experimental trials. Participants were given the opportunity to take a break every 50 trials. All sessions were identical, and the order of the blocks was counter-balanced across participants.

Each of the trials began with a small fixation cross at the centre of the display, followed by the presentation of the search field. Participants were then given an unlimited amount of time to then respond either “present” or “absent”. After doing so, the current trial ended and the next began. When an incorrect response was given, an audible tone was produced by the computer.

The study used a mixed design, with three independent variables, consisting of: Target Type (metals, IEDs, absent), Search Type (single-target search, dual-target search), and target Prevalence (20%, 50%, 80%). Target Type and Search Type were within-subjects factors, and target Prevalence was a between-subjects variable. Target Prevalence described the percentage of trials on which a target was presented to participants in both single- and dual-target search. Note that, unlike the previous chapter, the relative prevalence of the targets in dual-target search was held constant at a 1:1 ratio. In other words, in 50% prevalence dual-target search, metals were presented on 25% of trials, and IEDs were presented on 25% of trials. In 20% prevalence dual-target search, metals were presented on 10% of trials and IEDs were presented on 10% of trials. Likewise, in 80% prevalence dual-target search, metals were presented on 40% of trials and IEDs were presented on 40% of trials. Only one target could appear on any trial. The dependent variables were response accuracy and response time.

3.2.4 Important Note: Methodological and Computational Issues

It must be noted here that, for the present, and subsequent chapters, a number of factors need to be considered in the examination of the dual-target cost. To maintain consistency across the different experiments, the results for each experiment will be examined using the same process.

To begin with, overall error rates will be examined. Error rates for misses and false alarms in single-target metals, single-target IEDs, and dual-target search will be compared. Initially, only the *overall* error rates will be used (i.e. the total error rate for dual-target search, rather than the proportion of trials on which a metal was missed, or the proportion of trials on which an IED was missed). Normally, it would be preferable to examine the error rates for dual-target search in terms of which target was missed, yet, the first set of analyses will not do so, because it is upon the overall error rates that the signal detection measures for *d'* and *c* are based. This is because participants were asked to state only whether or not they believed a target to be present or absent, rather than being asked to state *which* target they believed to be present. Had the participants been asked to state which target they believed to be present, then a different set of signal detection formulae, based upon the three-alternative forced choice model (3AFC), would be used.

Essentially, requiring participants to respond “present” or “absent” can easily give a measure of hit rates, but can only give a unified measure of false alarm rates, rather than false alarm rates for each choice (so, for example, false alarms occurring when a participant reported that a metal was present, when actually an IED was present), which would be needed for a 3AFC calculation. Thus, a solution to this is to use the overall miss rate for dual-target search. Still, there is reason to believe, based upon the results of Chapter 2, that IEDs are somewhat more difficult to detect than metals. Therefore, rather than ignore this issue, once the overall analyses of the error rates is complete, the impact of the dual-target cost upon each target will be examined, focusing upon the miss error rates for each of the targets in dual-target search. Although it would be preferable to reduce the number of analyses being carried out, the approach here is to explore the changes in the error rates that give rise to changes in the signal detection measures, and to explore how search performance for the individual targets is impacted by the experimental manipulations. This process is used throughout the remaining chapters of the present thesis.

3.3 Results

In the following results, all *t*-tests that were conducted had their *p* values Bonferroni-corrected before being reported. Furthermore, Greenhouse-Geisser *F* values, degrees of freedom, and *p* values are reported for ANOVA results where tests of sphericity are violated (i.e. Mauchly’s test of sphericity shows a *p* value of less than .05). In all figures, error bars represent ±S.E.M.

A number of related questions were explored in the analyses of the results from this experiment. To begin with, error rates in search were be examined, in order to test the impact of the prevalence effect and dual-target cost upon search performance. Following on from the error rates are analyses of the RT data, conducted in order to compare the results with previous examinations of the prevalence effect and dual-target cost (Menneer, et al., 2007; Wolfe, et al., 2007).

Finally, specific tests of the criterion shift account of the prevalence effect will be made using signal detection parameters.

3.3.1 Replicating the Prevalence Effect and the Dual-target Cost: Error Rates

Figure 3.3.1a shows the basic replication of the prevalence effect and the dual-target cost. The left panel shows miss errors, the right panel shows false alarm errors. Miss and false alarm rates were examined using two separate 3 (Search Type: single-target metals, single-target IEDs, dual-target search) × 4 (Session: 1, 2, 3, 4) × 4 (Setsize: 4, 8, 12, 16) × 3 (Prevalence: 20%, 50%, 80%) repeated-measures ANOVAs, with Prevalence entered as a between-subjects factor. To begin with, *overall* error rates in dual-target search were examined, because, as already noted, it is upon such overall error rates that the signal detection parameters rest.

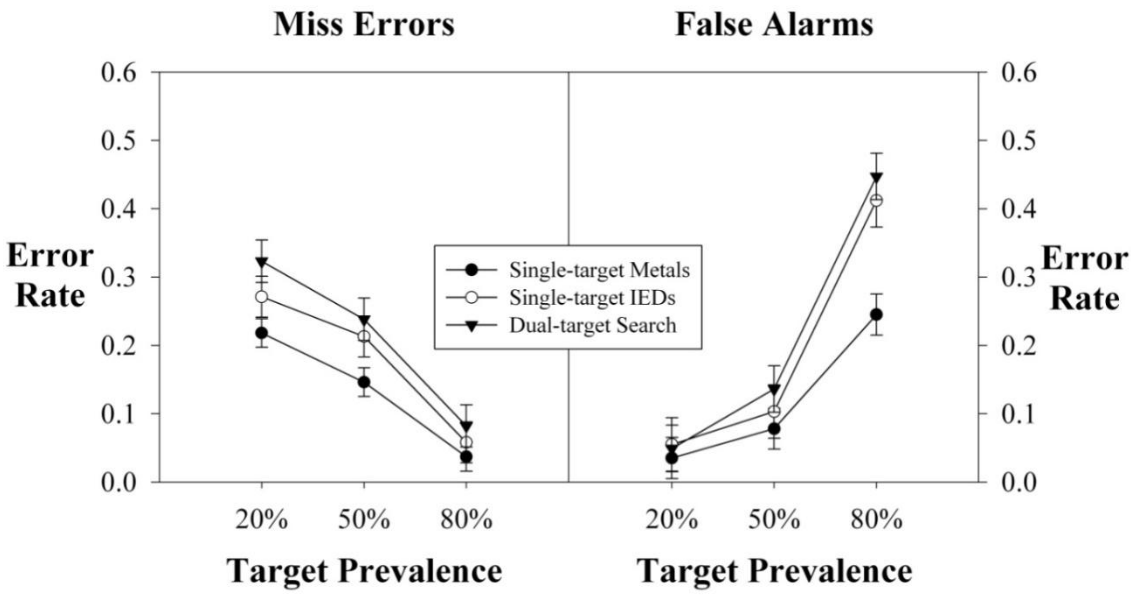


Figure 3.3.1a: Error rates as a function of Target Prevalence and Search Type.

Miss rates: Examination of the miss rates replicated the dual-target cost and prevalence effect. More miss errors were made in dual-target search than in single-target search; additionally, miss errors increased as prevalence decreased. Miss rates showed no main effects of Setsize, nor any interactions between Setsize and the remaining factors (Fs< 2.2). Thus, to simplify the rather complex and lengthy analyses presented in the present chapter, subsequent examinations of miss rates were based upon mean error rates averaged across the setsizes. This reduced the

ANOVA design to consisting of a 3 (Search Type: single-target metals, single-target IEDs, dual-target search) × 4 (Session: 1, 2, 3, 4) × 3 (Prevalence: 20%, 50%, 80%) repeated-measures ANOVA, with Prevalence entered as a between-subjects factor. Under this ANOVA, miss rates showed a significant main effect of Prevalence ($F(2,24)=16.8$, $p<.001$), as well as a main effect of Search Type ($F(2,48)=20.8$, $p<.001$). However, there was no evidence of an interaction between Search Type and Prevalence ($F<1.6$). The Session factor also failed to reach significance, either as a main effect, or as an interaction (all Fs<1.5).

Subsequent t-tests revealed that, as Prevalence decreased, the miss error rate increased. Miss error rates increased from 0.06 (S.E.M.=0.01) at 80% prevalence to 0.19 (S.E.M.=0.01) at 50% prevalence ($t(214)=12.5, p<.001$); miss error rates also increased to 0.27 (S.E.M.=0.01) at 20% prevalence (comparing 50% versus 20% prevalence: $t(214)=4.3, p<.001$; comparing 20% versus 80% prevalence: $t(214)=15.9, p<.001$). Thus, as a target was presented with a lower level of prevalence, the likelihood that the target would be missed increased significantly: this is a clear replication of the standard prevalence effect (Wolfe, et al., 2007).

Comparing the different search types revealed the presence of a dual-target cost for miss rates (the Search Type factor was examined separately from the Prevalence factor because the Search Type × Prevalence interaction did not reach significance, as described above). Miss rates increased significantly from 0.13 (S.E.M.=0.01) for single-target metals and 0.18 (S.E.M.=0.01) for single-target IEDs to 0.21 (S.E.M.=0.01) for dual-target search (single-target metals versus dual-target: $t(107)=8.3, p<.001$; single-target IEDs versus dual-target: $t(107)=4.3, p<.001$). Additionally, single-target metals exhibited overall lower miss rates than single-target IEDs ($t(107)=5.1, p<.001$).

False Alarm Rates: Analyses of the false alarm rates produced a substantially larger number of effects than the miss rates. Broadly speaking, false alarm rates increased as Prevalence increased (i.e. the opposite effect of how prevalence impacted miss rates). Additionally, there was some evidence that the dual-target cost was present in false alarm rates (it was detected in comparisons between single-target metals and dual-target search, but not between single-target IEDs and dual-target search).

As with the miss rates, the false alarm rates exhibited an effect of Prevalence ($F(2,24)=24.9$, $p<.001$). As Prevalence increased, so did the false alarm rate. Additionally, there was a three-way interaction between Setsize, Search Type and Prevalence ($F(12,144)=2.6$, $p<.01$: see Figure 3.3.1b), in which was embedded a number of other main effects and interactions, including main effects of Search Type ($F(2,48)=26.2$, $p<.001$), and of Setsize ($F(1.4,33.8)=33.9$, $p<.001$), as well as interactions between Search Type and Prevalence ($F(4,48)=11.7$, $p<.001$), between Setsize and Prevalence ($F(6,72)=10.8$, $p<.001$), and between Search Type and Setsize ($F(3.5,84)=5.5$, $p<.01$).

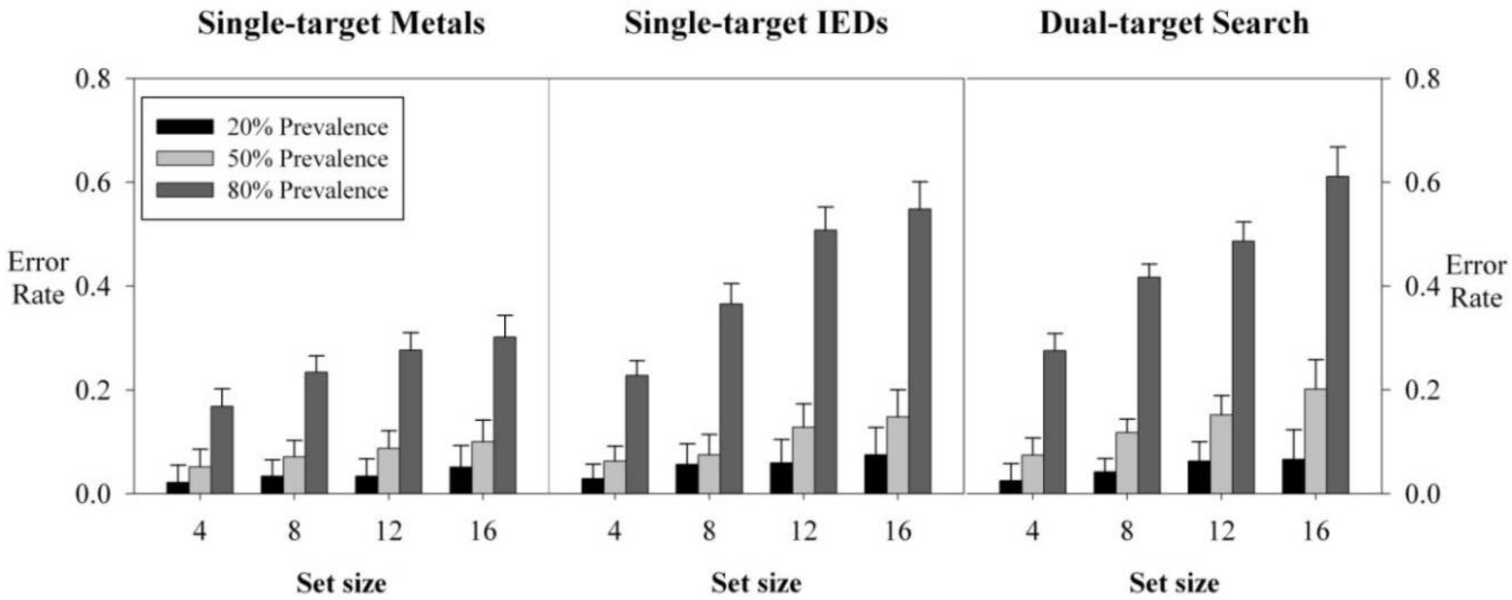


Figure 3.3.1b: False alarm rates as a function of Setsize, Search Type and Prevalence

In order to explore the three-way Search Type × Prevalence × Setsize interaction, false alarm rates for the three Search Types were examined using a series of further ANOVAs conducted upon each Search Type separately. These revealed that there was a Setsize effect for false alarms in single-target metals ($F(3,72)=5.5$, $p<.05$), which did not interact with Prevalence ($F<1.4$). For single-target metals, increases in Setsize increased the probability that a false alarm would occur (comparing Setsize=4 with Setsize=16: $t(26)=2.7$,$p<.05$). Still, there was an effect of Prevalence for single-target metals ($F(2,24)=12.9$, $p<.001$). Comparing false alarm rates at 20% Prevalence with those at 50% Prevalence in an additional ANOVA, it was found that the effect of Prevalence narrowly missed significance ($F1,16)=4.1$,$p=.059$). Next, comparing false alarm rates at 50% Prevalence with those at 80% Prevalence revealed that false alarm rates increased between 50% and 80% Prevalence for single-target metals ($F(1,16)=9.96$,$p<.01$). Mean false alarm rate for single-target metals at 20% Prevalence was 0.035, for 50% Prevalence was 0.078, and for 80% Prevalence was markedly higher, with a false alarm rate of 0.256.

Next, single-target IEDs saw increasing false alarm rates with increases in Setsize ($F(1.8,44)=39$, $p<.01$), yet also showed an interaction between Setsize and Prevalence ($F(6,72)=14.9$, $p<.01$). This interaction was caused by the fact that there was no difference in false alarm rates between 20% and 50% Prevalence across the setsizes, whilst 80% Prevalence showed elevated false alarm rates compared to both 20% and 50% Prevalence. Table 3.3.1c gives details of the independent-samples *t*-tests that were carried out to examine this interaction. Finally, dual-target search resulted in higher false alarm errors as a function of Setsize ($F(1.7,39)=21.4$, $p<.01$), which interacted with Prevalence ($F(6,72)=6.6$, $p<.01$). As with single-target IEDs, this interaction was caused by there being no differences between 20% and 50% Prevalence across the setsizes (aside from Setsize=8), whilst false alarm rates were higher in 80% Prevalence than in both 50% and 20% Prevalence across all of the setsizes. Table 3.3.1c gives the results of the *t*-tests examining these effects.

In order to test for the presence of the dual-target cost upon false alarm rates, two further ANOVAs were used to compare the three-way Search Type × Prevalence × Setsize interaction. In these ANOVAs, false alarm rates in dual-target search were compared with each of the single-target search conditions separately (i.e. dual-target search versus single-target metals, followed by dual-target search versus single-target IEDs). These ANOVAs revealed that the false alarm rates showed a dual-target cost for metals ($F(1,24)=38.4$, $p<.001$), but not for IEDs ($F<2.3$). The dual-target cost for false alarm errors was amplified for high levels of Prevalence, yet only for metals (there was a significant dual-target cost × prevalence interaction for metals versus dual-target: $F(2,24)=16$, $p<.001$, but not for IEDs: $F<1.6$). Indeed, there was no dual-target cost for metals at 20% prevalence ($F<1$), yet there was for 50% ($F(1,8)=21.8$, $p<.01$) and 80% ($F(1,24)=25$, $p<.01$) prevalence levels. The effect size (measured by partial eta squared, η^2) grew from $\eta^2=.731$ at 50% prevalence to $\eta^2=.757$ at 80% prevalence.

Finally, a main effect of Session narrowly reached significance for the overall false alarm rates ($F(2.1,51.3)=4.2$, $p=.049$), and did not interact with any of the other factors (Fs<1.5). The main effect was caused by an overall reduction in false alarm rates across the sessions. A set of ANOVAs revealed that error rates decreased significantly between the first and second sessions (i.e. there was a main effect of Session: $F(1,24)=9.9$,$p<.01$), but not between the second and third sessions, or third and fourth sessions. Furthermore, in all of these ANOVAs, Session did not interact with any of the other factors (all Fs<1.7). The main effect of Session is displayed graphically in Figure 3.3.1d.

Table 3.3.1c
Results of Independent-Samples t*-tests comparing False Alarm Error Rates at Different Prevalence Levels and Setsizes, for Single-Target IEDs and Dual-Target Search*

Prevalence Comparison	Setsize	*t*	*df*	*p*	*t*	*df*	*p*
		Search Type					
		Single-target IEDs			Dual-target Search		
20% versus 50%	4	2.2	16	ns	2.4	10.1	ns
20% versus 50%	8	0.6	16	ns	3.4	16	<.05
20% versus 50%	12	1.9	16	ns	2.9	16	ns
20% versus 50%	16	1.8	16	ns	3.3	10	ns
50% versus 80%	4	3.4	8.3	<.05	3.5	9.8	<.05
50% versus 80%	8	4.4	8.5	<.05	6.9	10.4	<.05
50% versus 80%	12	4.8	9.1	<.05	4.8	16	<.05
50% versus 80%	16	4.3	9.3	<.05	3.9	16	<.05
20% versus 80%	4	3.9	9.2	<.05	4.5	8.2	<.05
20% versus 80%	8	4.4	10.1	<.05	8.7	9.8	<.05
20% versus 80%	12	5.5	9.7	<.05	6.4	16	<.05
20% versus 80%	16	4.9	9.3	<.05	5.2	8.2	<.05

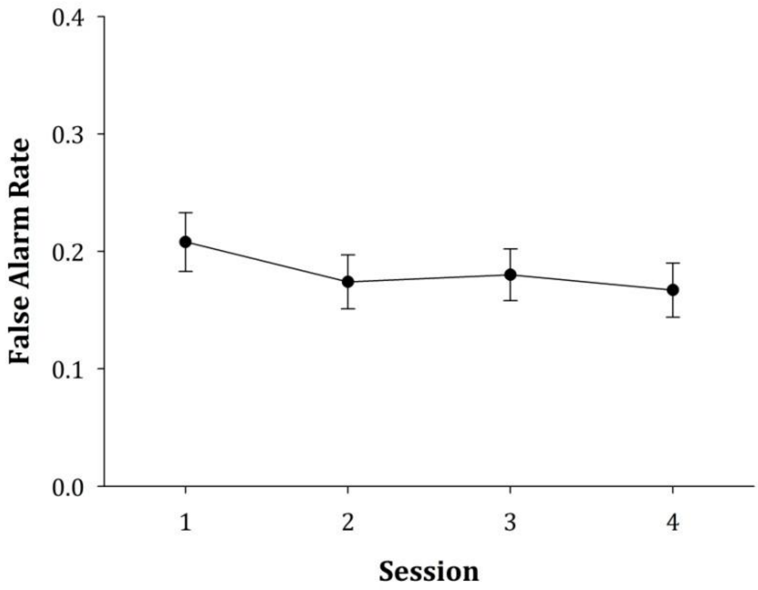


Figure 3.3.1d: Overall false alarm rates in each Session.

3.3.2 Impact of Prevalence on Each Target in Dual-target Search

The preceding analyses examined the dual-target cost and prevalence effect on overall error rates in dual-target search, in order to lead in naturally to the signal detection examinations that are presented below. However, an important issue needs to be addressed based upon the detection performance of both individual targets: was the dual-target cost amplified for one target more than the other as prevalence varied?

In order to answer such a question, a 2 (Target Type: metals, IEDs) × 2 (Search Type: single-target search, dual-target search) × 4 (Session: 1, 2, 3, 4) × 3 (Prevalence: 20%, 50%, 80%) repeated-measures ANOVA was carried out. Again, Prevalence was entered as a between-subjects factor. Note that, as with the previous chapter, the Target Type factor here has only two levels, because the target-present trials are being examined alone. The dual-target cost was present (i.e. there was a main effect of Search Type: $F(1,24)=27$, $p<.001$), with single-target search having a mean error rate of 0.16 (S.E.M=0.01), and dual-target search having a mean error rate of 0.22 (S.E.M.=0.02). There was also a main effect of Target Type ($F(1,24)=10.3$, $p<.01$), with metals showing lower error rates than IEDs (metals mean miss rate=0.15, S.E.M.0.01; IEDs mean miss rate=0.22, S.E.M.=0.02).

Additionally, a main effect of Prevalence was detected ($F(2,24)=15.9$, $p<.001$). Although miss errors were no different between 20% prevalence and

50% prevalence ($t(16)=1.7, p>.05$), miss errors did decline between 50% and 80% prevalence ($t(16)=4.4, p<.01$), and between 20% and 80% prevalence ($t(16)=5.9, p<.01$). This is a somewhat surprising result, especially given that, in the previous section, it was reported that miss errors were significantly higher in 20% prevalence than in 50% prevalence. The most likely cause for this difference is the inclusion of the individual performance levels of each target in dual-target search. As can be plainly seen from Figure 3.3.2a, the standard error of the miss rates is quite large in the search for IEDs in dual-target search. This alone could contribute to a weakening of the differences between the Prevalence groups, especially between 20% Prevalence and 50% Prevalence. Indeed, a three-way interaction between Search Type, Target Type and Prevalence marginally neared significance ($F(2,24)=2.7, p=0.089$). No other effects or interactions were significant (all Fs<2). Despite the irregularities, these analyses show that, although metals were detected more regularly than IEDs, there was no impact of Prevalence upon the dual-target cost for miss error rates, and target prevalence did not impact IEDs more than metals.

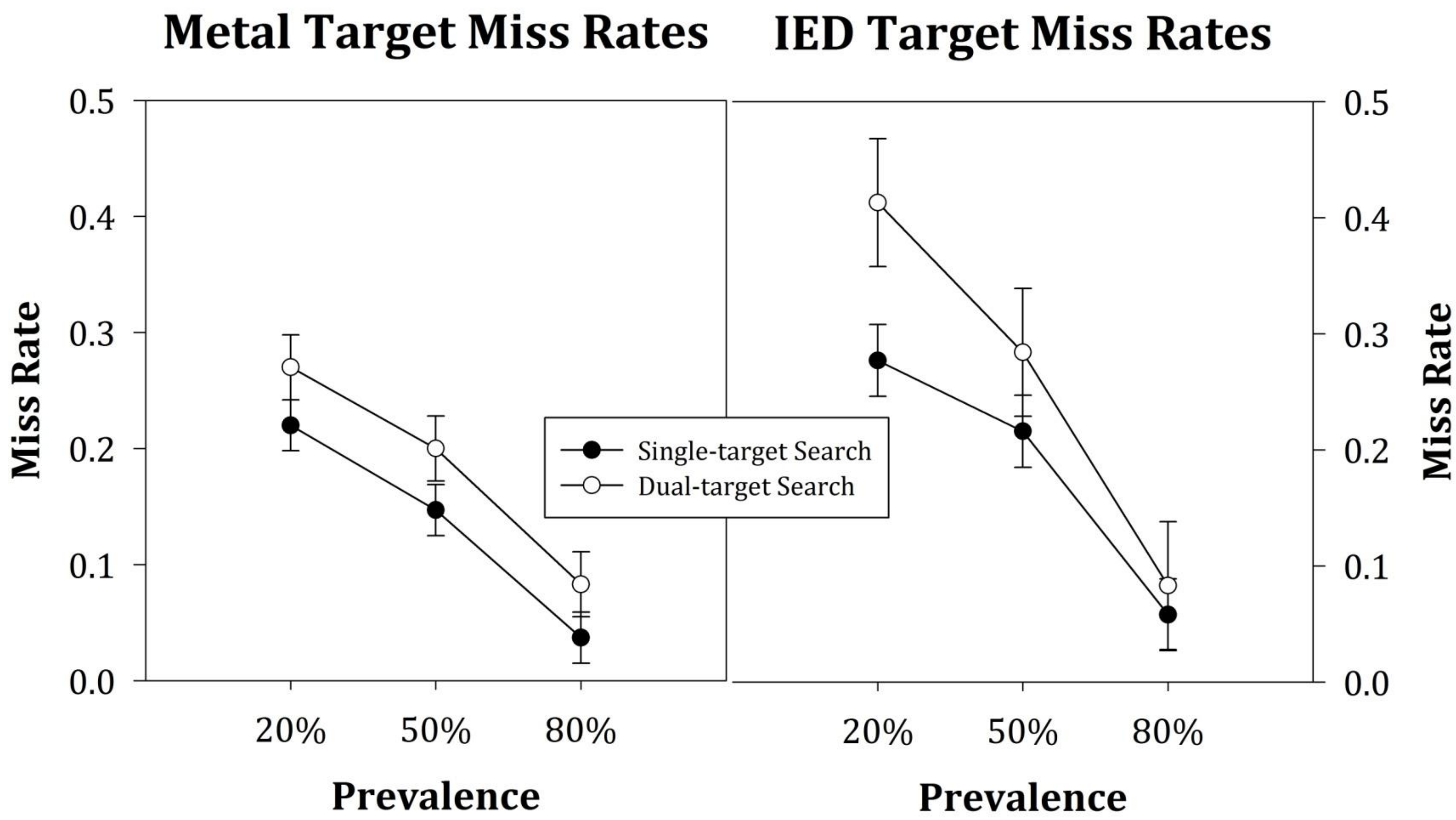


Figure 3.3.2a: Miss error rates as a function of Prevalence and single- or dual-target search, for both metals and IEDs

3.3.3 Reaction Times

Previous studies of the prevalence effect have reported that target-absent trials are responded to as rapidly as target-present trials when prevalence is low (Wolfe et al., 2007). The present study did not fully replicate this pattern. The three Prevalence levels were examined separately using a set of 4 (Session:1,2,3,4) × 4 (Setsize:4,8,12,16) × 2 (Search Type: Single-target, Dual-target) × 2 (Presence: Present, Absent) ANOVAs, conducted upon the metal and IED trials separately. These revealed significant main effects of Presence for both metals and IEDs, but not for every Prevalence level (see Table 3.3.3a), suggesting that target-absent trials were being responded to less rapidly than the target-present trials, even when prevalence was low.

The most likely reason for these comparisons to not perfectly mirror the previous studies of the prevalence effect is that the prevalence levels used here were considerably higher than used previously (Wolfe et al., 2007 used prevalence levels as low as 2%).

Table 3.3.3a

Results of ANOVA Comparisons Examining the Reaction Times for Target-present and Target-absent Responses for each of the three Prevalence Levels

Target Prevalence	*df*	*F*	*p*	η^2	Mean Present RT (S.E.M.)	Mean Absent RT (S.E.M.)
			Metals			
20%	1,8	12.9	<.01	0.617	1228 (62)	1752 (190)
50%	1,8	33.7	<.001	0.808	1018 (71)	1702 (139)
80%	1,8	76.8	<.001	0.906	1099 (65)	2842 (256)
			IEDs			
20%	1,8	1.5	ns	0.157	1574 (98)	1752 (190)
50%	1,8	20.5	<.01	0.719	1207 (54)	1702 (139)
80%	1,8	61.8	<.001	0.885	1225 (57)	2842 (256)

Despite the fact that the initial examinations of the RT data were not able to perfectly mirror previous experiments, there were a number of further questions that needed to be addressed:

1. First of all, was there a dual-target cost for RT?
2. Second, were participants slower to detect targets in low-prevalence search (as would be predicted by the stimulus probability effect: Estes, et al., 1957; Simpson & Voss, 1961)?
3. Third, were participants responding ‘absent’ more rapidly as prevalence decreased?
4. Finally, did the effects of the dual-target cost interact with those of any prevalence effects?

In order to answer the above questions, a 2 (Search Type: Single-target, Dual-target) × 3 (Trial Type: Metals, IEDs, Absent) × 4 (Setsize: 4,8,12,16) × 4 (Session: 1,2,3,4) × 3 (Prevalence: 20%, 50%, 80%) ANOVA was carried out upon the RT data, with Prevalence entered as a between-subjects factor. This ANOVA revealed a number of main effects and interactions, which can be broadly grouped into two categories, namely a *dual-target and prevalence* category, as well as a *practice effects* category. The overall pattern of results is depicted below, in Figure 3.3.3b.

It is important to note that, in dual-target search, absent trials occur when *both* targets are absent, meaning that there are no separable IEDs-absent, or metals-absent trials in dual-target search. Thus, to allow for a direct comparison between RTs for single-target search and dual-target search absent trials, the single-target metals and single-target IEDs RTs were mean-averaged. This was shown to be permissible through the use of a 2 (Search Type: Single-target metals, Single-target IEDs) × 4 (Setsize: 4,8,12,16) × 4 (Session: 1,2,3,4) × 3 (Prevalence: 20%,50%,80%) ANOVA, with Prevalence used as a between-subjects factor. Under this ANOVA, Search Type showed no significant main effects or interactions with any of the other factors (all Fs<2).

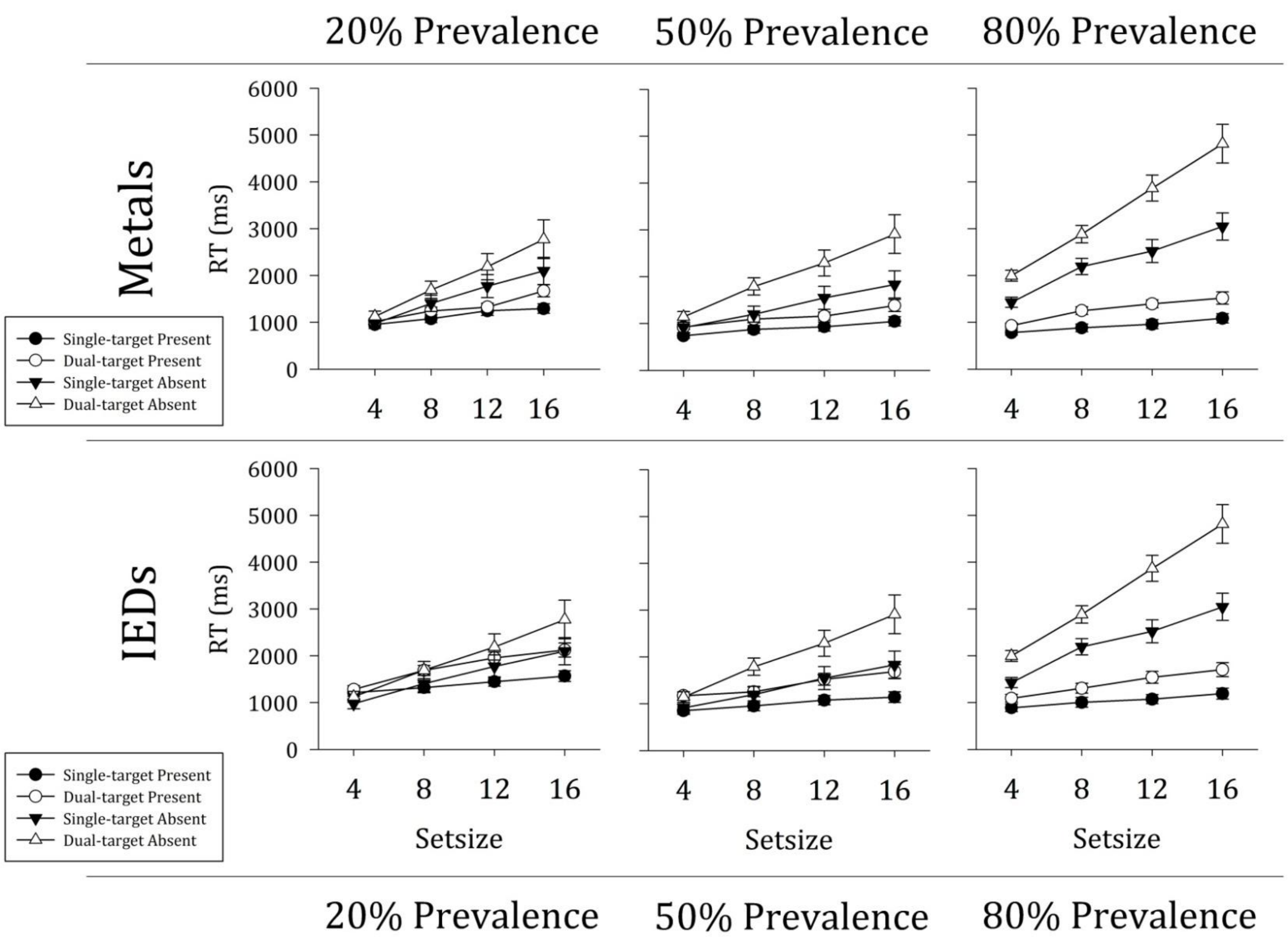


Figure 3.3.3a: Reaction times in the three Search Types as a function of Prevalence and Setsize. Note that target-absent RTs have been placed upon both Metals and IEDs, in order to show how 'absent' responses vary as Prevalence decreases

3.3.4 Reaction Times: Dual-target and Prevalence Effects

Examinations of the RT data revealed that there was evidence of a dual-target cost for metals, IEDs and target-absent trials, and, furthermore, that there was evidence of variations in the time needed to make a response for IEDs and target-absent trials only. Additionally, increases in setsize appear to amplify the magnitude of the dual-target cost.

There was a significant four-way interaction between Setsize, Search Type, Trial Type and Prevalence ($F(3,73)=6.5$, $p<.05$), which encapsulated a number of the other main effects and interactions that were detected. The interaction between Setsize, Search Type, Trial Type and Prevalence was examined for the presence of the dual-target cost as follows. To begin with, each Trial Type was analysed using a further ANOVA that was identical in design to the initial ANOVA conducted upon the RT data, except that the Trial Type factor was removed.

Time taken to detect metal targets: Examination of metal trials revealed that the magnitude of the dual-target cost increased with increases in Setsize (dual-target cost × Setsize interaction: $F(3,72)=6.7$,$p<.001$). The effect size of the dual-target cost tended to increase with increases in Setsize: $\eta^2=.416$ at Setsize=4; $\eta^2=.490$ at Setsize=8; $\eta^2=.463$ at Setsize=12; $\eta^2=.562$ at Setsize=16. However, Prevalence did not reach significance as either a main effect ($F(2,24)=2.6$,$p=.099$), or as an interaction with any of the remaining factors (all Fs<1.5).

Time taken to detect IEDs: As with metals, the IEDs showed an increase in the dual-target cost with increases in Setsize (dual-target cost × Setsize interaction: $F(2.3,54.5)=7.8$,$p<.01$). The magnitude of the dual-target cost tended to increase with Setsize: $\eta^2=.288$ at Setsize=4; $\eta^2=.478$ at Setsize=8; $\eta^2=.618$ at Setsize=12; $\eta^2=.500$ at Setsize=16. However, unlike metals, IEDs showed evidence of being detected less rapidly as Prevalence reduced. There was a main effect of Prevalence ($F(2,24)=8.1$,$p<.01$). At 20% Prevalence, mean detection RTs for IEDs was 1573ms, at 50% prevalence, mean RT detection decreased to 1207ms (this was detected using an ANOVA isolating 20% versus 50% prevalence: $F(1,16)=10.8$,$p<.01$). However, target detection RTs in 80% prevalence were no faster than in 50% prevalence (80% prevalence mean detection RT=1224: $F<1$).

Time taken to respond 'absent': Finally, target-absent trials also showed a dual-target cost for RTs which increased with increases in Setsize (dual-target cost × Setsize interaction: $F(1.4,34.1)=20.4$,$p<.001$), $\eta^2=.583$ at Setsize=4; $\eta^2=.588$ at

Setsize=8; η^2=.703 at Setsize=12; η^2=.597 at Setsize=16. There was also a main effect of Prevalence ($F(2,24)=10.3, p<.010$) and an interaction between Search Type and Prevalence ($F(2,24)=4.1, p<.01$).

The Search Type × Prevalence interaction was examined in a series of two additional ANOVAs, one comparing 20% and 50% Prevalence, and the second comparing 50% and 80% Prevalence. These revealed that target-absent responses were no different in their speed between 20% and 50% prevalence, yet were significantly slower in 80% prevalence. The first of these, comparing RTs in 20% Prevalence with RTs in 50% Prevalence, revealed no main effect of Prevalence, nor any interactions between Prevalence and the remaining factors (Fs<1.7). However, comparing 50% Prevalence with 80% Prevalence revealed that there was now a main effect of Prevalence: $F(1,16)=15.3, p<.01$. Thus, although RTs for target-absent trials were no different between 20% Prevalence (mean RT=1753ms) and 50% Prevalence (mean RT=1703ms), the RTs were longer in 80% Prevalence (2842ms) than 50% Prevalence for target-absent trials.

3.3.5 Reaction Times: Practice Effects

A number of practice effects emerged from analyses of the RT data. As would be expected, when faced with a difficult task, participants became more rapid in their responses as the sessions progressed. Based upon the initial ANOVA conducted upon the RTs, there was a main effect of Session ($F(2,39)=53$, $p<.001$), and interactions between Setsize and Session ($F(3,62)=2.6$, $p<.001$), between Trial Type and Session ($F(2.5,60)=13.7$, $p<.001$), and between Trial Type, Session and Prevalence ($F(5,60)=3$, $p<.05$).

The interaction between Trial Type, Session and Prevalence is displayed graphically in Figure 3.3.5a. For all three trial types, the Session factor reached significance (metals: $F(1.8,3.6)=29.1, p<.001$; IEDs: $F(1.9,3.7)=64.6, p<.001$; target-absent: $F(1.5,35.2)=36.4, p<.001$). These main effects were the result of participants responding more rapidly as the sessions progressed (comparing the first and final sessions: metals, $F(1,24)=42.1, p<.001$; IEDs, $F(1,24)=92.2, p<.001$; absent, $F(1,24)=45.8, p<.001$). For all three sets of comparisons (metals, IEDs, target-absent), the Session factor failed to interact with Prevalence (Fs<1). RTs for metals trials decreased from an average of 1445ms (S.E.M.=79) in Session 1 to an average of 934ms (S.E.M.=29.8) in Session 4. RTs for IEDs-present trials decreased from an average of 1903ms (S.E.M.=88) in Session 1 to an average of 1034ms (S.E.M.=51) in Session 4. RTs for target-present trials decreased from an average of 2897ms (S.E.M.=190) in Session 1 to an average of 1629ms (S.E.M.=112) in Session 4.

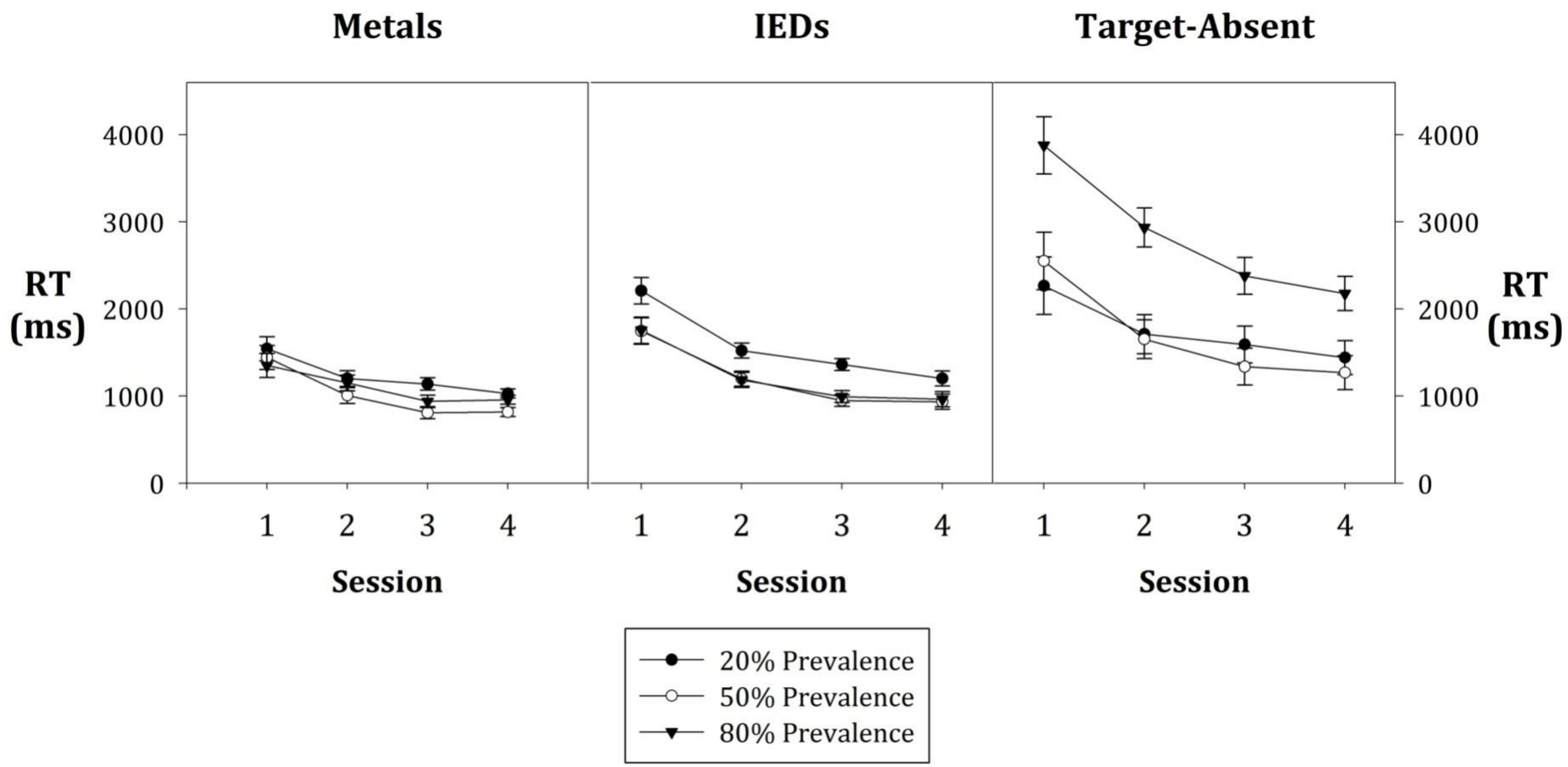


Figure 3.3.5a: Reaction times for each Trial Type as a function of Prevalence and Session

3.3.6 Signal Detection Theory Parameters

For the analyses of Signal Detection Theory parameters, hit rates and miss rates were averaged across the setsizes. This was a method used previously by Wolfe et al. (2007), and is necessary because doing so reduces the number of cells in the analysis where performance is at ceiling or at zero. Signal Detection Theory analyses can not function upon 0% or 100% scores, and, although corrections are possible, it is safer to aggregate across setsizes in this instance, or else a substantial number of cells will need correction.

Sensitivity and Criterion: Hit and false alarm rates were utilised to compute Signal Detection Theory parameters for sensitivity (d') and criterion (c). These are shown in Figure 3.3.6a. As can be seen from Figure 3.3.6a, there was a decline in sensitivity as prevalence increased for IEDs and dual-target search, but not for metals, partially replicating the results of Wolfe et al. (2007). An overall ANOVA examining sensitivity with a 3 (Search Type: single-target metals, single-target IEDs, dual-target search) × 4 (Session: 1, 2, 3, 4) × 4 (Setsize: 4, 8, 12, 16) × 3 (Prevalence: 20%, 50%, 80%) design showed an interaction between Prevalence and Search Type ($F(6,72)=4$, $p<.01$), with a main effect of Search Type ($F(2,48)=66.8$, $p<.001$), and a strong trend for a main effect of Prevalence ($F(2,24)=3$, $p=.074$). This Prevalence × Search Type interaction was caused by there being no effect of Prevalence upon single-target metals ($F<1$), yet there was a significant main effect of Prevalence for single-target IEDs ($F(2,24)=3.3$, $p=.054$) and dual-target search ($F(2,24)=5.5$, $p<.05$).

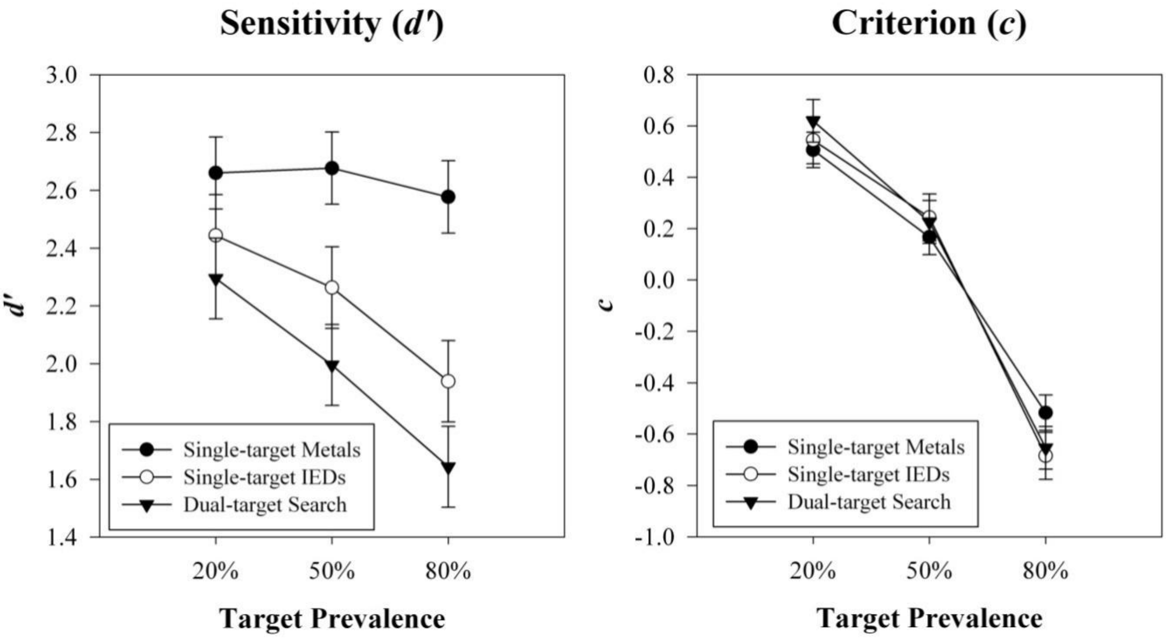


Figure 3.3.6a: Sensitivity and criterion parameters for the various prevalence levels and search types.

Single-target IEDs and dual-target search were then compared separately, with tests for an effect of Prevalence in a series of further ANOVAs. For IEDs, there was no effect of Prevalence between the 20% and 50% prevalence conditions (F<1), or between the 50% and 80% prevalence conditions (F<2.9). However, there was a reduction in sensitivity between 20% and 80% prevalence ($F(1,16)=5.8, p<.05$). Mean sensitivity in single-target IEDs at 20% Prevalence was 2.4, for 50% Prevalence was 2.2, and finally, for 80% Prevalence, was 1.9. For dual-target search, there was a similar pattern. Sensitivity did not vary between 20% and 50% Prevalence, or between 50% and 80% Prevalence (Fs<2.9), yet did drop significantly between 20% and 80% Prevalence ($F(1,16)=10, p<.01$). Mean sensitivity in dual-target search was 2.3 for 20% Prevalence, 2.0 for 50% Prevalence, and 1.6 for 80% Prevalence.

Finally, there was a main effect of Session ($F(3,72)=10, p<.001$), and an interaction between Session and Search Type ($F(6,144)=2.5, p<.05$: see Figure 3.3.6b, below). The Session × Search Type interaction was examined using a series of t-tests conducted upon the first and final sessions. The results of these tests are displayed in Table 3.3.6c. Overall, they indicated that, in the first session, single-target Metals search was more sensitive than either single-target IEDs or dual-target search. However, in Sessions 2-4, sensitivity for IEDs rose, and became higher than sensitivity in dual-target search. By the final Session, IEDs were equal in sensitivity to metals, and both were higher in sensitivity than dual-target search.

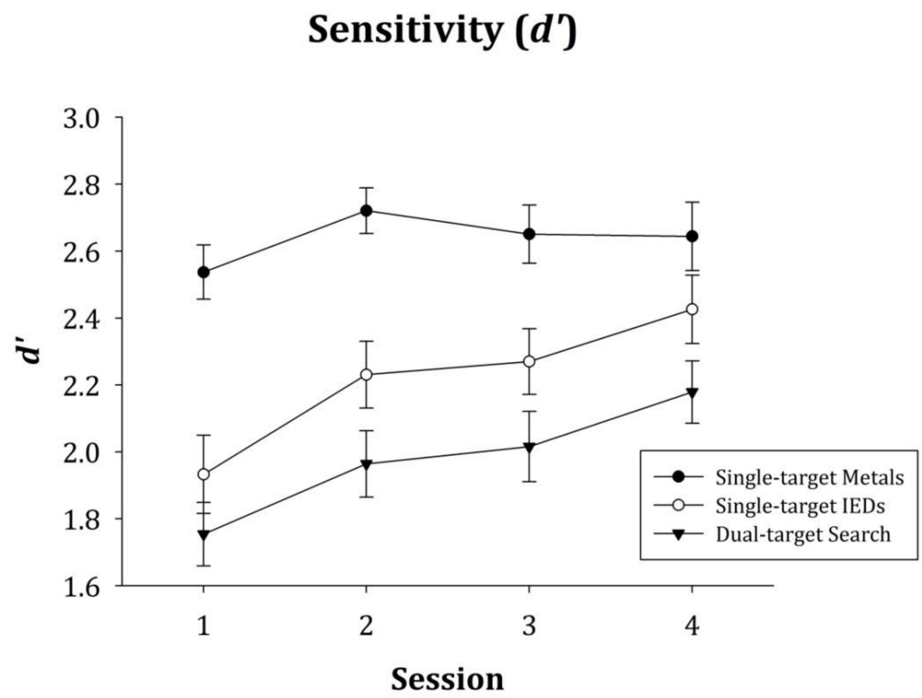


Figure 3.3.6b: Sensitivity as a function of Search Type and Session.

Table 3.3.6c

Results of t*-test comparisons between Sensitivity for the different Search Type factors in each Session*

Comparison	Session	*t*	*df*	*p*
Single-target Metals versus IEDs	1	5.4	26	<.05
Single-target Metals versus Dual-target	1	9.2	26	<.05
Single-target IEDs versus Dual-target	1	1.5	26	ns
Single-target Metals versus IEDs	2	5.0	26	<.05
Single-target Metals versus Dual-target	2	7.3	26	<.05
Single-target IEDs versus Dual-target	2	3.8	26	<.05
Single-target Metals versus IEDs	3	4.4	26	<.05
Single-target Metals versus Dual-target	3	5.9	26	<.05
Single-target IEDs versus Dual-target	3	3.9	26	<.05
Single-target Metals versus IEDs	4	2.1	26	ns
Single-target Metals versus Dual-target	4	4.4	26	<.05
Single-target IEDs versus Dual-target	4	3.8	26	<.05

The criterion parameter was also examined using an ANOVA with a 3 (Search Type: single-target metals, single-target IEDs, dual-target search) × 4 (Session: 1, 2, 3, 4) × 4 (Setsize: 4, 8, 12, 16) × 3 (Prevalence: 20%, 50%, 80%) design. This showed a closer replication of Wolfe et al.'s (2007) findings than the sensitivity data reported above. Overall, the criterion became more liberal (i.e. reduced), as prevalence increased. Although there was no main effect of Search Type (F<1), there was a main effect of Prevalence ($F(2,24)=65$, $p<.001$), and an interaction between Prevalence and Search Type ($F(4,48)=4$, $p<.01$). There was also apparent evidence of an interaction between Prevalence and Session ($F(6,72)=2.7$,$p<.05$), but the effect was so weak that post-hoc comparisons failed to find evidence of any significant differences within the interaction. No other effects or interactions reached significance (all *F*s<1.7).

The interaction between Prevalence and Search Type was explored using a series of smaller ANOVAs, focusing on each Prevalence level in turn. There were no effects of Search Type at either 20% Prevalence ($F<2.1$) or 50% Prevalence ($F<2.1$), yet there was at 80% prevalence ($F(2,16)=3.7$, $p<.05$). In 80% Prevalence, the criterion for single-target IEDs was slightly more liberal than search for metals ($F(1,8)=13.3$,$p<.01$), yet was unchanged between single-target metals and dual-target search ($F<3.1$,$p>.110$), or between single-target IEDs and dual-target search ($F<1$).Mean criterion for single-target metals was -0.52 (S.E.M.=0.08), for single-target

zROC Curve Slopes. Aside from reporting a concomitant increase in sensitivity and criterion as prevalence was reduced, Wolfe et al. (2007) also plotted variations in prevalence levels upon a zROC curve, and reported that the slope of the curve was roughly equal to 0.6. This was an important finding as it allowed Wolfe et al. (2007) to explain why sensitivity (*d'*) in their study was exhibiting the rather odd behaviour of increasing alongside increases in the criterion (*c*). In the present study, these effects were, for the most part, replicated. In order to carry out their examination, Wolfe et al. (2007) averaged the zROC points across all participants and setsizes in a given experiment for each prevalence level. Here, zROC co-ordinates for the hit rate and false alarm rates were averaged across the setsizes, sessions, and across the participants, in order to replicate the method used by Wolfe et al. The resultant zROC curves are presented in Figure 3.3.6e. For single-target metals, the slope of the zROC curve was 0.9; thus it is not surprising that the single-target metals condition showed changes in criterion, but no changes in sensitivity. Participants searching for metals alone were operating at essentially a unit-slope ROC curve. However, the single-target IEDs condition showed a somewhat different slope of 0.67, whereas dual-target search was shallower still, with a slope of 0.60. Thus, it appears that, as with Wolfe et al. (2007), the present results suggest that, when *d'* increases alongside increases in the criterion, then this is the result of a non-unit zROC slope underlying the decision-making processes involved in the experiment. If this were not the case, then we would be forced to accept the rather perplexing possibility that observers can perform *better* at a task (i.e. have higher sensitivity) whilst also being *biased against detecting targets* (i.e. had a conservative critierion).

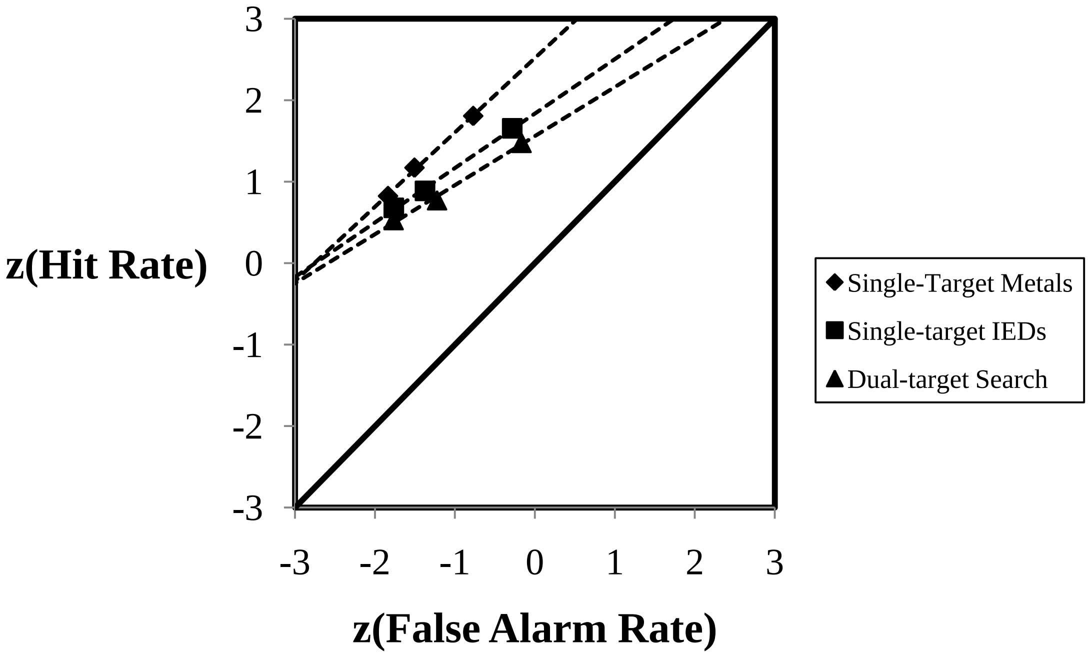


Figure 3.3.6e: zROC curve for the different search types across variations in prevalence

3.3.7 Matching False Alarms and Misses

One further claim made by Wolfe et al. (2007) was that participants set a criterion in a given condition based upon an attempt to equalise their number of false alarms with their number of misses. To test this claim, aggregated miss and false alarm counts were produced for each search type (single-target metals, single-target IEDs, dual-target search) and participant involved in the experiment. A set of paired-samples *t*-tests, and mean miss and false alarm counts are presented in Table 3.3.7a. The results of these *t*-tests were not entirely clear. Out of the nine comparisons that were made, five of them supported Wolfe et al.'s (2007) claims, making it somewhat uncertain that participants were actually attempting to equalise their raw number of false alarms with their number of misses. Additionally, the *t*-tests which failed to support Wolfe et al.'s (2007) claims were spread across all three prevalence levels, making it unlikely that the inability to support their claims were stemming from just one group of the participants engaging in an abnormal response strategy. These tests will be returned to shortly in the discussion, below.

Table 3.3.7a

Matching Miss Errors with False Alarm Errors using paired t*-tests. Mean Error Counts are presented, with Standard Error of the Mean presented in brackets*

Prevalence (%)	Search Type	Mean Errors (SEM)		Number of Trials		*t*	*df*	*p*
		False Alarms	Misses	Absent	Present			
20	Dual-target Search	25 (6)	41 (5)	512	128	1.7	8	ns
20	Single-target IEDs	28 (11)	35 (5)	512	128	0.5	8	ns
20	Single-target Metals	18 (2)	28 (3)	512	128	2.6	8	<.05
50	Dual-target Search	45 (7)	66 (3)	320	320	3	8	<.05
50	Single-target IEDs	34 (4)	59 (7)	320	320	2.9	8	<.05
50	Single-target Metals	25 (6)	40 (6)	320	320	2.1	8	ns
80	Dual-target Search	56 (6)	42 (6)	128	512	1.6	8	ns
80	Single-target IEDs	51 (7)	29 (4)	128	512	2.4	8	<.05
80	Single-target Metals	31 (6)	19 (2)	128	512	1.8	8	ns

3.4 Discussion

The present study had a number of goals that were intended to develop an understanding of the task given to airport X-ray security screeners, and also to develop previous research that has investigated the target prevalence effect and dual-target cost. These will now be discussed in turn, with their relevance to theory and practice.

3.4.1 The Effects of Prevalence and Imbalances in Human Perception

The key finding from the present study is that human perception of target prevalence is *imbalanced* when comparing high (80% prevalence) with low (20% prevalence) prevalence conditions. As is rather strikingly clear from Figure 3.3.1a, it appears that observers are more willing to make false alarms at high prevalence than they are willing to miss targets at low prevalence. This imbalance in the effects of prevalence has not been reported in previous studies, which have only considered low levels of prevalence (i.e. <50% prevalence), and is thus a novel effect.

Why did variations in prevalence have such an imbalanced impact upon miss and false alarm error rates? Examinations of the error rates as a function of setsize showed negligible effects upon miss rates in terms of setsize, yet for false alarms, aside from the main effect of prevalence, there were substantial increases in false alarms as setsize increased. Were these errors caused by guessing on behalf of the participants? It may have been the case that, in 80% prevalence, participants were rapidly responding 'present', and giving a false alarm, just as Fleck and Mitroff (2007) suggested that, in low prevalence search, participants rapidly respond 'absent' and producing a miss error. However, analyses of the RT data make this seem unlikely: for target-absent trials, as prevalence increased, the time taken to respond 'absent' increased substantially. Had participants merely been guessing in the 80% prevalence condition, then one would expect these RTs to have been low. This was not the case for target-absent trials, and, for target-present trials, RTs were no different to those in the 50% prevalence condition.

One key goal of the present study was to examine the claims made by the criterion shift account of the prevalence effect produced by Wolfe et al. (2007). The results here can be used to extend the criterion shift account in a number of ways.

Under the criterion shift account, it was argued that there are no *true* changes in sensitivity as prevalence varies; instead, the apparent changes in sensitivity are a result of the non-unit zROC slope employed by observers. Wolfe et al. (2007) further argued that observers operate upon a zROC slope of around 0.6 as prevalence goes from low (2% prevalence) to medium (50% prevalence) levels. Here, there were no changes in sensitivity for metal targets across the different prevalence levels, yet still a change in criterion. This was not the case for IEDs or dual-target search, which both showed a concomitant decrease in sensitivity and criterion as prevalence increased, in line with the results of Wolfe et al. (2007). The slope for the zROC curve in the search for metals was essentially at unity (0.9), whilst the slope for IEDs was closer to 0.7 (it was 0.67), and the slope for dual-target search did, in fact, reach 0.6.

Why do the zROC curve slopes produced in the present study differ to the claims of Wolfe et al. (2007) that observers will be operating using a zROC curve slope of 0.6? It seems likely that, in fact, the zROC curve slope is related to the difficulty of the search that is used to generate the zROC curve in the first place. In Wolfe et al.'s (2007) experiments, the task was somewhat more difficult to that which was used in the present study. In an effort to make their task more comparable to airport security screening, they used overlapping and transparency in their displays, making the task somewhat more difficult to complete effectively. Here, no overlapping or transparency was used. Thus, it seems plausible to consider the curve of the zROC curve as being related to task difficulty: the more difficult the task, the more shallow the slope becomes (for further discussion regarding the interpretation of non-unit slope zROC curves, see Macmillan & Creelman, 2005).

One very obvious question could thus be: what makes a search task 'difficult'? In signal detection terms, this can be answered by stating that any factor that increases the noise and thereby decreases the sensitivity will make a search task more difficult. Relating this to visual search, task difficulty can be broadly divided between *bottom-up* and *top-down* factors. Bottom-up factors include salience and conspicuity (Itti & Koch, 2000; Navalpakkam & Itti, 2007). Top-down factors include object knowledge and the target template (Connor, et al., 2004; Theeuwes, et al., 2005; Yantis, 2000, 2005). IEDs suffer compared to metals in terms of both bottom-up and top-down factors. Metals are simple targets that can be defined in most cases by a well-known shape (gun-shaped or knife-shaped) and a single colour range (blue – blue-black). IEDs are complex and varied targets which can be defined in terms of two somewhat more diffuse shapes (a large 'block' of explosives, coupled with some form of electronic detonator) and two broader colour ranges (the explosives are orange / orange-brown; the detonators are blue, green, and black for the wiring). IEDs can be embedded within any number of objects, including radios, laptops, clothing, shoes, and so on. At a most basic level, this means that IEDs are more difficult to recognise compared to metal threats. Furthermore, participants entered the study having seen images of both guns and knives beforehand, and thus had a pre-existing knowledge of their appearance. This was not the case for IEDs. Indeed, in the studies conducted by Wolfe et al. (2007), IEDs are not used, because of their complexity. Here, it has been shown that IEDs can be detected to a reasonable degree of accuracy, but still, are detected with a lower degree of sensitivity compared to metals. Putting the results of Wolfe et al. (2007) together with those of the present study, it would appear that metals can be detected with a unit slope zROC curve when the search task is easy, but not when it is made more difficult. IEDs, on the other hand, can not be detected with a unit slope zROC curve, at least with the stimuli used here.

Finally, as a further test of the model of prevalence presented by Wolfe et al. (2007), a comparison between the raw number of false alarms and misses was conducted. This revealed some rather confusing effects: as shown in Table 3.3.7a, five of the eight sets of *t*-tests supported Wolfe et al.'s (2007) claims, whilst three did not. It is thus somewhat difficult to comment on the accuracy behind their suggestion that the criterion is set based upon an attempt to equalise the raw number of miss errors with the raw number of false alarms errors. As Wolfe et al. (2007) argue that this attempt to equalise the types of errors is fundamental to the setting of the criterion, it seems that the present results can not entirely support their claims. In their experiments, they made greater efforts to give feedback to participants than in the present study. Here, whenever a participant made an error, the computer produced an audible tone, and no other feedback was given. Wolfe et al. (2007) gave participants feedback after every response that they gave, either correct or incorrect, and encouraged participants to perform effectively.

3.4.2 The Dual-target Cost

The Signal Detection approach adopted here was very useful in exploring the dual-target cost and connecting an understanding of the dual-target cost with current signal detection work surrounding the prevalence effect. In terms of sensitivity, the dual-target cost interacted with prevalence, being weaker at low levels of prevalence. Examining the miss and false alarm rates, it appears that the cause of this is the compression of false alarm rates at low levels of prevalence (i.e. between 20% and 50% prevalence; see Figure 3.3.2), as part of the overall imbalance in the prevalence effect. As was expected, the sensitivity of dual-target search was *lower* than that for IEDs and metals. Given the impact of irrelevant objects in dual-target search (Stroud, et al., in preparation), this result is not really surprising.

However, what was surprising was that the dual-target cost did not interact with target prevalence, and become attenuated or amplified. Instead, the dual-target cost appears to be a flat decrease in sensitivity (i.e. d'), and increase in RT.

3.4.3 General Discussion and Relevance to Airport X-ray Security Screening

The present experiment developed and tested the criterion shift account of Wolfe et al. (2007) in a number of ways. First, the results showed quite clearly that the assumption that variations in prevalence are examined along a zROC curve with a slope of around 0.6 could not be replicated perfectly. Instead, a clarification was made: the slope of the zROC is a function of task difficulty, and the slope becomes more shallow when the task is made more difficult. Additionally, it seems unlikely that the criterion adopted by observers is based upon an attempt to equate miss error rates with false alarm error rates. Still, it appears that the basic framework set down by Wolfe et al. (2007) can be extended to high levels of prevalence.

Perhaps the most vital development from the present experiment is the finding that variations in prevalence are responded to in an imbalanced manner across miss and false alarm rates. This may have some important benefits for airport security screening. Security screeners are rarely ever presented with real targets, as one might expect, yet, still, false images are artificially implanted by the screening equipment software occasionally to ensure that screeners are performing their task efficiently. The false images are called TIP (Threat Image

Projection) images, and are presented on around 2% of all passenger bags (Hofer & Schwaninger, 2005).

Based upon the compression of false alarms between 20% and 50% prevalence, and the lack of a compression of miss error rates between 20% and 50% prevalence, it may be the case that, in real security screening, TIP rates can be increased safely in order to increase the hit rate, and decrease the miss rate. Such a procedure would, if the normal assumptions of Signal Detection Theory held true, also increase the false alarm rate (Macmillan & Creelman, 2005). However, that seems to not be the case here, because of the imbalances in perception as prevalence varies. As a result, it may be possible to increase the TIP prevalence, whilst increasing the TIP hit rate, and also the real threat item hit rate, whilst, at the same time, *not* increasing the false alarm rate. It should be noted that, in the screening environment, false alarms are both costly and time-consuming (of course, less so than missing a target), and, as such it is desirable to avoid increases in false alarms wherever possible. The next chapter explores such possibilities in more detail, and extends the range of prevalence levels under examination.

4

Further Extensions of the Dual-target Cost and Prevalence Effect

The Dual-target Cost across a wide Gamut of Prevalence Levels

4.1 Introduction

The previous chapter explored the prevalence effect in terms of Wolfe et al.'s criterion shift account (2007). Several of the account's claims were replicated (a concomitant increase in sensitivity and criterion as prevalence decreased), yet some claims were not replicated (the main aspect of the criterion shift account did not hold true for metal threat items, and participants did not always appear to be equalising false alarm errors with miss errors). Still, when *d'* did increase with criterion, as was the case with single-target IEDs and dual-target search, the zROC curve that was constructed between the prevalence levels was around the 0.6 mark, as predicted by the criterion shift account. Thus, in agreement with Wolfe et al. (2007), it can be said that the sensitivity levels in low-prevalence search were not truly elevated: in other words, the apparent increase in *d'* when prevalence was low was not a true effect, and was merely the result of the fact that the assumptions upon which *d'* calculations are based were violated.

With that in mind, is it at all possible to use a different measure of sensitivity, one which does not have its assumptions violated in the search for threat items? Measures of sensitivity using *d'* assume that the underlying distributions of Signal and Noise are normally distributed, however, the finding that the zROC curve slopes were less than 1 indicate that the actual Signal and Noise distributions in the search task that was used are *not* normally distributed. In some senses, this is rather 'circumstantial' evidence, and it would be of value to use an alternative index of sensitivity, which can be compared statistically in the different prevalence conditions. Fortunately, there exist measures of sensitivity that are *distribution-free* or *non-parametric* (Macmillan & Creelman, 2005). One such measure will be used in the present chapter to explore whether or not a true change in sensitivity actually occurs when prevalence varies.

Therefore, the present chapter further extends the criterion shift account, using a broader range of prevalence levels (the experiment reported in the previous chapter used prevalence levels of 20%, 50% and 80%; the present experiment uses prevalence levels of 2%, 24%, 50%, 76% and 98%), as well as to test the imbalanced variations between hit and false alarm rates as prevalence varies that were reported in Chapter 3. Additionally, non-parametric signal detection parameters will be used to determine whether or not sensitivity truly changes in conjunction with changes in target prevalence.

4.1.1 Prevalence Effects and Receiver Operating Characteristic Curves

The alternative measure of sensitivity that will be used in the present chapter is A_z, otherwise known as the area under the Receiver Operating Characteristic (ROC) curve. ROC curves were discussed in the previous chapter, and are described in some detail in Appendix B. ROC curves plot false alarm and hit rates along a continuum. The higher the 'arch' of the ROC curve, the higher the sensitivity. However, on any given ROC curve, sensitivity is invariant (recall from Appendix B that ROC curves are also known as *isosensitivity* curves, with *iso* meaning 'same').

Therefore, if Wolfe et al. (2007) are correct in their assumption that variations in prevalence leads to a shift in criterion but not in sensitivity, then, if ROC curves are constructed for different levels of prevalence, those curves should essentially be the same. In other words, different levels of prevalence should show the same sensitivity, despite changes in criterion. As prevalence increases, the only change that occurs should be that observers move along the ROC curve showing more confident responses that targets are present, rather than moving onto a different ROC curve entirely (see also Mueller & Weidemann, 2008). Referring to Figure 4.1.1a, below, when prevalence increases, observers should move from left to right along the ROC curve, becoming more confident that a target is present (see also Appendix B). Wolfe et al. (2007) assumed that this was occurring when plotting their zROC curves, and, in the previous chapter, the construction of the zROC curves also made the same assumption. Essentially, the present experiment tests this assumption.

The assumptions of the criterion shift account are important to consider in light of a number of recent claims that the prevalence effect only exists in the laboratory. Gur and colleagues (Gur, Rockette, Armfield, et al., 2003; Gur, Rockette, Warfel, Lacomis, & Fuhrman, 2003) examined confidence ratings from a group of radiologists searching for abnormalities in X-rays, and reported that A_z remained unchanged for different levels of prevalence. However, they failed to report (or even test) for changes in criterion (concern with relation to a similar study has been noted by Wolfe et al., 2007). Thus, the present study aims to resolve this dispute by exploring *both* sides of the argument, using the area under the ROC curve, and standard measures of the criterion.

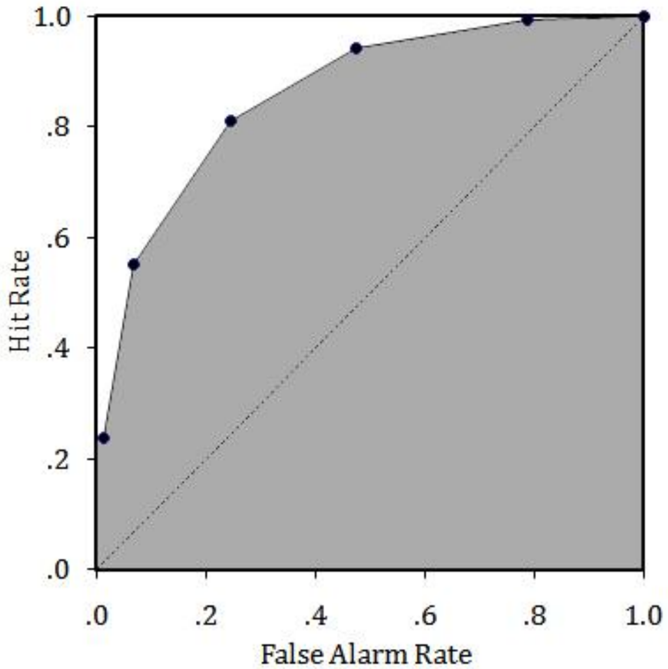


Figure 4.1.1a: ROC Curve. Shaded area represents area under the ROC curve (A_z), used as an alternative measure of sensitivity to *d'*. Each point on the curve represents a hit rate / false alarm rate pair for each confidence rating. Confidence increases from left to right along the curve.

4.1.2 Research Questions: Extending the Prevalence Effect

As with the previous chapter, the focus of the experiment described here will be on examining the relationship between the prevalence effect and the dual-target cost, and extending the criterion shift account of the prevalence effect (Wolfe et al., 2007). A broad range of prevalence levels are used here, in order to examine the prevalence effect across a wide range (2%-98%). The use of ROC curves requires confidence ratings to be given on each trial, and, as a result, a substantial number of trials will be needed for each participant; this is exaggerated by the extreme prevalence levels used here. Thus, a limited number of participants were engaged in a series of sessions: although in some senses this will constitute a loss of power, it does mean that alternative measures of sensitivity can be generated from the data, and, in addition, further examinations of the prevalence effect can be carried out. As with the previous chapter, the key research questions here can be re-iterated (and altered slightly, based upon the results from Chapter 3):

1. Can the finding that decreases in target prevalence result in an increase in *d'*, coupled with an increase in *c* be replicated?
    a. Does this only occur with *d'*? Will it also occur with a non-parametric measure of sensitivity, A_z?
    b. Will the reduction in sensitivity occur with IEDs and dual-target search again, as in the previous chapter? Will metals now be affected in a similar manner, when lower levels of prevalence are used?
2. Do observers operate upon a zROC curve with a slope of around 0.6 using a different set of stimuli to that used by Wolfe et al. (2007)?
3. If so, can this be extended to high levels of prevalence?
4. Furthermore, is this pattern replicated in dual-target search?

4.2 Method

4.2.1 Participants

Nine participants took part in the study (seven females and two males). Ages ranged from 18 to 27 years (mean age=20.6 years, SD=3.2 years). Participants were undergraduates and postgraduates, and received either course credit or payment for their participation. All participants reported normal colour vision and no previous experience with the stimuli. All participants completed the study within 30 days.

4.2.2 Apparatus

The experimental software was programmed using *Presentation*, and was conducted using a PC running Windows XP Professional with Service Pack 2 installed. The PC had a 1.7 GHz Intel processor, 512 MB of RAM, and the stimuli were presented on a 17" Relisys CRT Monitor, with a refresh rate of 75 Hz and a resolution of 1024x768 pixels. Participants responded using the same Cedrus RB-610 button box as in the previous experiments reported in this thesis. No head restraints were used, and viewing distance was around 60cm from the monitor. The experiment took place in a moderately-lit room.

4.2.3 Stimuli

The stimuli were the same set of images as used in the previous experiment. Due to changes in monitor size, however, the image sizes were reduced using Adobe Photoshop CS3 in order to maintain the overall shape and ratio between the height and width of the images. As with the previous experiments, the images were presented on a virtual 4 × 4 grid drawn out across the display. Selection of the images to be used was based on the same criteria (randomisation of selection, position, and orientation) as in the previous experiments.

4.2.4 Design and Procedure

Participants took part in eight sessions, with each session lasting around 45 minutes. Before the actual trials began, a detailed explanation was given to the participants regarding the nature and appearance of the targets, with participants being guided through twenty examples of each type of threat item that they were to search for.

The sessions were each blocked into three different sets of trials: single-target search for metals, single-target search for IEDs, and dual-target search for metals and IEDs. The blocks began with five practice trials, followed by 200 experimental trials. Participants were given the opportunity to take a break every 50 trials. All sessions were identical, and the order of the blocks was counter-balanced across participants.

Complete counter-balancing of the search blocks was not possible with such a small number of participants. The three search blocks counterbalanced perfectly would give a total of six different block orders that would have to be counterbalanced. With only three participants in each prevalence condition, this resulted in only three of the six block orders being used. The same set of block orders was thus used for all prevalence groups. One participant in each prevalence condition was first presented with single-target metals, followed by single-target IEDs, followed by dual-target search. A second participant was first presented with dual-target search, followed by single-target metals, followed by single-target IEDs. A final participant was presented with single-target IEDs first, followed by dual-target search, followed by single-target metals. Thus, across the three participants in each condition, each of the three blocks was presented as the first, second, and third blocks. For further discussion of some of the problems facing the counterbalancing of experiments investigating target prevalence, see the supplementary material from Wolfe et al. (2005).

Each of the trials began with a small fixation cross at the centre of the display for 1000ms, followed by the presentation of the search field. Participants were then given an unlimited amount of time to respond with a rating of how confident they were that a target was present in the display. Ratings were given using a response box, with six potential ratings in total. After a rating was given, the trial ended and the next began. When an incorrect response was made, an audible tone was produced by the computer. Only one target could appear on any trial.

Before the trials, began, it was explained to participants that the ratings that they gave were on a scale, going from "Certain Present" to "Certain Absent". The button box that was used to give responses was labelled "present" near the "Certain Present" button, and was labelled "absent" near the "Certain Absent" button. Additionally, it was explained that ratings between 1 and 3 signified belief that a target was present, with ratings between 4 and 6 signifying that a target was absent. This is summarised below in Table 4.2.4a.

Table 4.2.4a

Summary of Confidence Ratings used in the Present Experiment

Status	Target Present			Target Absent		
Rating	1	2	3	4	5	6
Confidence	Certain	Moderately Certain	Slightly Certain	Slightly Certain	Moderately Certain	Certain

The study used a mixed design, with three independent variables, consisting of: Target Type (metals, IEDs, absent), Search Type (single-target search, dual-target search), and target Prevalence (2%, 26%, 50%, 74%, 80%). Target Prevalence was a between-subjects variable, and described the regularity with which targets were presented to participants in both single- and dual-target search. Dependent variables were response time and the rating given on each trial. The rating was converted to a third dependent variable, namely response accuracy. This was achieved by treating ratings 1-3 as a “present” response, and ratings 4-6 as an “absent” response.

4.3 Results

As with the previous chapter, the results will first be examined in terms of error rates (miss errors and false alarm errors), after which a set of Signal Detection Theory parameters will be computed and examined in detail. This will enable further extensions and tests of the criterion shift account of the prevalence effect (Wolfe et al., 2007). Proportion of error rates were based upon responses in which participants responded “present” or “absent”, with all three degrees of certainty (e.g. “weak confidence present” was simply treated as “present”, and so on). As with the previous chapters, *t*-tests show Bonferroni-corrected values, and

Greenhouse-Geisser degrees of freedom and *p* values are used whenever sphericity is violated. In all figures, error bars represent ±S.E.M.

An initial examination of the results suggested that one of the participants in the 50% Prevalence condition performed poorly at the search task. In several cases, their response accuracy on target-absent trials was greater (often double) their target-present trial response accuracy. To correct the noise introduced by this participant, any cells in which they scored an error rate higher than 2.5 standard deviations from the mean of the other two participants for the given cell, session, and search target, were corrected with the mean error rate for the other two 50% prevalence participants for that respective cell. Overall, twenty-six of the participant's cells were corrected in this manner (out of a total of forty-eight cells overall).

4.3.1 Replicating the Prevalence Effect and the Dual-target Cost: Error Rates

Miss and False Alarm Rates. Examinations of the miss and false alarm rates replicated the prevalence effect, but failed to replicate the dual-target cost. Figure 4.3.1a shows the impact of target prevalence and dual-target search upon error rates, with the left panel presenting miss errors, and the right panel showing false alarm errors. As prevalence decreased, miss rates increased and false alarm rates decreased. Error rates were examined using a 2 (Target Presence: present, absent) × 3 (Search Type: single-target metals, single-target IEDs, dual-target search) × 8 (Session: 1, 2, 3, 4, 5, 6, 7, 8), × 5 (Prevalence: 2%, 26%, 50%, 74%, 98%) repeated-measures ANOVA. This ANOVA revealed a main effect of Search Type ($F(2,20)=13.8$, $p<.001$), as well as of Prevalence ($F(4,10)=6.1$, $p<.01$). Additionally, there was a three-way interaction between Search Type, Presence, and Prevalence ($F(8,20)=5.9, p<.01$) which encapsulated a number of other interactions, namely between Presence and Prevalence ($F(4,10)=52, p<.001$), and between Search Type and Presence ($F(2,20)=5.9, p<.05$). Finally, there was also a main effect of Session ($F(7,70)=2.9, p<.05$), as well as an interaction between Presence, Prevalence and Session ($F(28,70)=1.7, p<.05$).

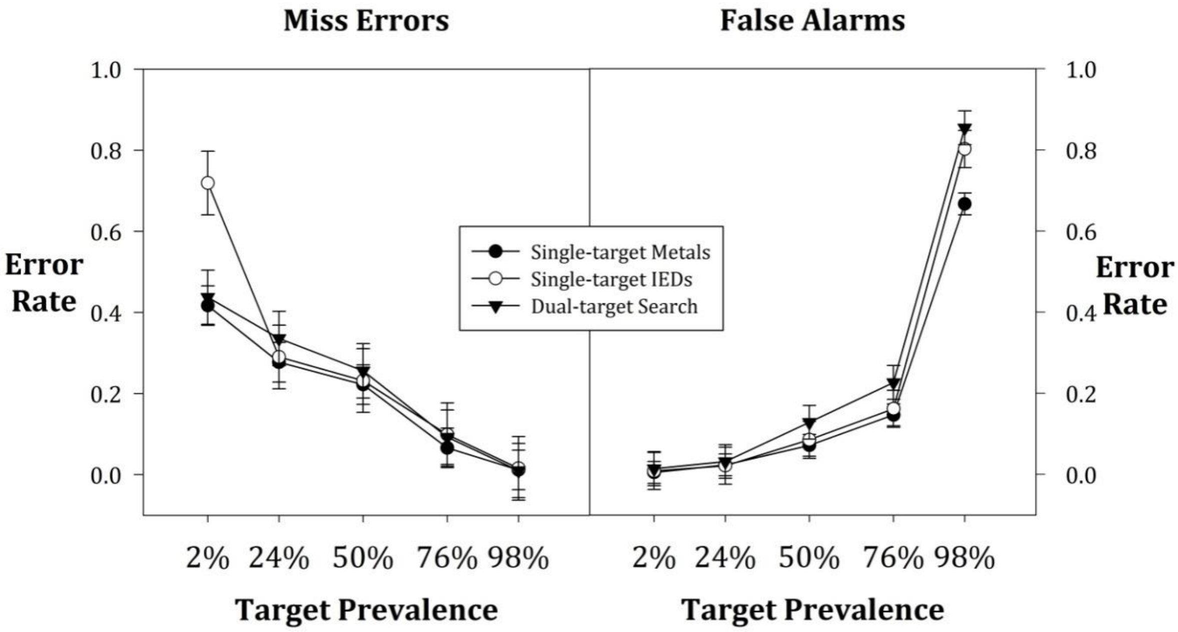


Figure 4.3.1a: Miss and false alarm error rates as a function of Prevalence and Search Type.

Miss Rates: Miss rates increased as prevalence decreased. Examination of the miss error rates indicated a main effect of Prevalence ($F(4,10)=10.9$, $p<.01$), as well as a main effect of Search Type ($F(2,20)=6.4$, $p<.01$), and an interaction between the two ($F(8,20)=5.6$, $p<.01$), plus an interaction between Session and Prevalence ($F(28,70)=1.7$,$p<.05$). Additional analyses were conducted with two separate goals: testing for the presence of the dual-target cost, and testing for the presence of the prevalence effect. These will now be discussed in turn.

Miss Rates: Testing for the Dual-target Cost: The present study failed to replicate the dual-target cost for miss rates. Further examinations of the miss rates upon each Prevalence level revealed that the Search Type factor only reached significance in the 2% Prevalence condition ($F(2,4)=7.3$,$p<.05$), and was non-significant, both as a main effect, and as an interaction, in the remaining Prevalence levels (Fs<1.8, ps>.15).

Search Type was then examined in the 2% Prevalence condition only, using paired comparisons of each of the Search Type factors. Surprisingly, the paired comparisons failed to reveal any significant effects, suggesting that the main effect of Search Type was weak to begin with. Still, what is perhaps more surprising is that the dual-target cost was not detected for the miss rates.

Miss Rates: Testing for the Prevalence Effect: Despite the lack of a dual-target cost for miss rates, the miss rates did show evidence of a prevalence effect. Single-target metals showed a main effect of Prevalence ($F(4,10)=11.2, p<.01$), as did single-target IEDs ($F(4,10)=12.1, p<.01$), and dual-target search ($F(4,10)=6.8, p<.01$). The Prevalence levels in each Search Type were examined using Tukey's HSD tests. The results of these tests are reported in Table 4.3.1b. It is important to consider that Prevalence is likely to have a continuous effect upon miss rates, rather than a strictly categorical effect, so consecutive levels of the Prevalence factor are not always likely to show significant differences between one another, especially when the separation between those levels is quite small, as it was here. However, in general, as Prevalence decreased, miss rates increased, thereby replicating the standard prevalence effect.

False Alarm Rates: The false alarm rates mirrored the miss rates with variations in prevalence. As prevalence increased, the false alarm rate increased. Additionally, as with the miss rates, the false alarm rates showed little evidence of a dual-target cost. For false alarms, there was a main effect of Prevalence ($F(4,10)=75.9$, $p<.001$), as well as a main effect of Search Type ($F(2,20)=22.4$, $p<.001$), and an interaction between the two ($F(8,20)=5.9$, $p<.01$). These effects will now be explored in detail.

False Alarm Rates: Testing for the Dual-target Cost: Each Prevalence level was examined separately, in order to examine the Search Type × Prevalence interaction, to test for the dual-target cost. There was no evidence of a dual-target cost for 2%, 24%, or 76% Prevalence (Fs<3.1, ps>.14). However, there was a dual-target cost 50% Prevalence ($F(2,4)=30.9, p<.01$) and for 98% Prevalence ($F(2,4)=13.6, p<.05$). Further comparisons revealed that false alarm rates were higher for dual-target search than for single-target metals in 50% Prevalence and 98% Prevalence (50% Prevalence: $F(1,2)=30.5, p<.05$; 98% Prevalence: $F(1,2)=36, p<.05$). For both 50% and 98% prevalence, false alarm rates did not differ between single-target metals and single-target IEDs, or between single-target IEDs and dual-target search (Fs<3.9,ps>.1).

False Alarm Rates: Testing for the Prevalence Effect: As with the miss rates, false alarm rates overall showed an interaction between Search Type and Prevalence (see above). To examine the impact of Prevalence, each of the Search Types were examined separately. In each Search Type, there was a main effect of

Prevalence (single-target metals: $F(4,10)=103.7, p<.001$; single-target IEDs: $F(4,10)=53.6, p<.001$); dual-target search: $F(4,10)=69.8, p<.001$). As with the miss rates, Tukey's HSD post-hoc tests were carried out to examine these main effects. The results are presented in Table 4.3.1b. Again, it is important to remember that Prevalence is likely to have a continuous effect upon false alarm rates, so consecutive levels of the Prevalence factor are not always likely to show significant differences between one another. Still, there was a general increase in false alarm rates as Prevalence increased.

Table 4.3.1b

Results of post-hoc tests comparing Prevalence Error Rates for Misses and False Alarms

Prevalence Comparison		Miss Error Rates			False Alarm Rates		
		Single-target Metals	Single-target IEDs	Dual-target Search	Single-target Metals	Single-target IEDs	Dual-target Search
2	24	ns	<.05	ns	ns	ns	ns
	50	ns	<.05	ns	ns	ns	ns
	76	<.05	<.05	<.05	<.05	ns	<.05
	98	<.05	<.05	<.05	<.05	<.05	<.05
24	50	ns	ns	ns	ns	ns	ns
	76	ns	ns	ns	ns	ns	<.05
	98	<.05	ns	<.05	<.05	<.05	<.05
50	76	ns	ns	ns	ns	ns	ns
	98	ns	ns	ns	<.05	<.05	<.05
76	98	ns	ns	ns	<.05	<.05	<.05

4.3.2 Reaction Times

Reaction times showed that the dual-target cost for RT was persistent throughout the experiment, that the dual-target cost was amplified for target-absent trials compared to metal or IED trials, and that, as Prevalence increased, target-absent RTs slowed considerably.

In the 2% Prevalence condition, there was an exceedingly low number of target-present trials (4 per block per Session); likewise, in the 98% Prevalence condition, there was an exceedingly low number of target-absent trials (again, 4 per block per Session). Thus, overall comparisons between target-present and target-absent RT performance were restricted to the 24%, 50% and 76% Prevalence conditions.

As was noted in the previous chapter, dual-target search absent trials occur when *both* targets are absent: thus, there are no separable IEDs-absent, or metals-absent trials in dual-target search. To allow for a direct comparison between RTs for single-target search and dual-target search absent trials, the single-target metals and single-target IEDs RTs were mean-averaged. This was shown to be permissible through the use of a 2 (Search Type: Single-target metals, Single-target IEDs) × 8 (Session: 1,2,3,4,5,6,7,8) × 3 (Prevalence: 24%,50%,76%) ANOVA, with Prevalence used as a between-subjects factor. Under this ANOVA, Search Type showed no significant main effect, or significant interactions with any of the other factors (all Fs<1.9).

Thus, RTs were examined using an 8 (Session: 1,2,3,4,5,6,7,8) × 3 (Search Type: Single-target, dual-target) × 2 (Trial Type: Metals, IEDs, Absent) × 3 (Prevalence: 24%, 50%, 76%) ANOVA, with Prevalence entered as a between-subjects factor.

There was an interaction between Search Type and Trial Type ($F(2,12)=9.6, p<.01$), which was caused by the fact that the dual-target cost for RT was largest for target-absent trials, and, although detected for metals and IEDs, was of a smaller magnitude (this interaction is depicted in left-hand graph of Figure 4.3.2a).For all three of the Trial Type factors, there was a dual-target cost (metals: $F(1,6)=6.3, p<.05$; IEDs: $F(1,6)=12.4, p<.05$; absent: $F(1,6)=25.4, p<.01$). It is of note that the effect size of the dual-target cost was highest in target-absent trials ($\eta^2=0.809$), and was reduced for IEDs ($\eta^2=0.675$), and was lowest for metals ($\eta^2=0.514$).

There was also an interaction between Search Type and Session ($F(7,42)=2.9, p<.05$). This was caused by practice effects: as can be seen in Figure 4.3.3a, in the right-hand panel below, RTs decreased as the sessions progressed, as would be expected. However, the dual-target cost was not eliminated, even in the final Session: there was a dual-target cost for RT in both the first Session ($F(1,6)=6.6, p<.05$) and the final Session ($F(1,6)=18.4, p<.01$), highlighting that the cost for RT cannot be eliminated even with considerable practice.

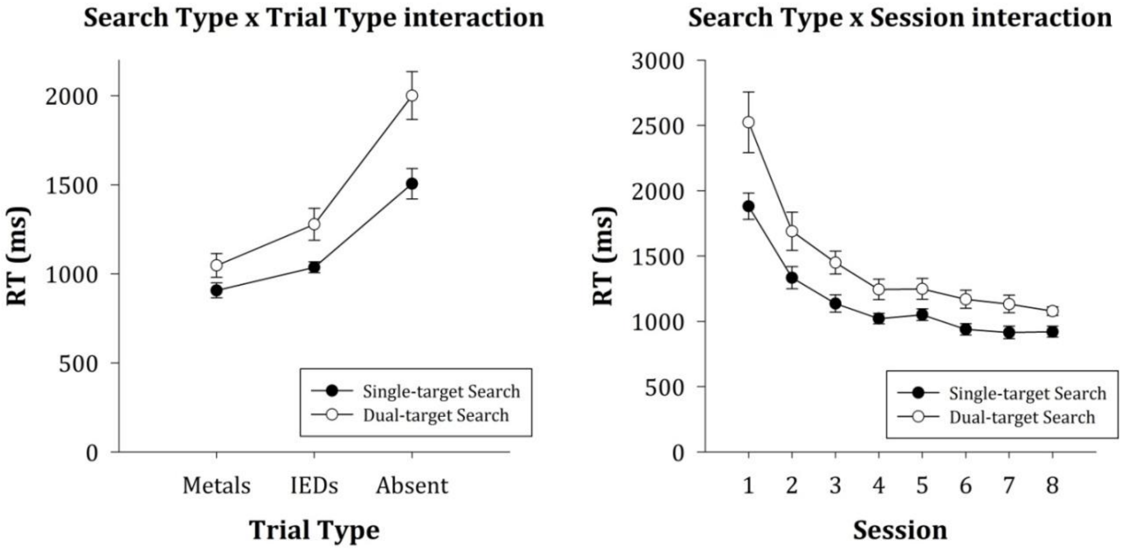


Figure 4.3.2a: Left-hand graph: Examination of the Search Type × Trial Type interaction, showing RTs for different Search Types and Trial Types; Right-hand graph: Examination of the Search Type × Session interaction, showing RTs for single- and dual-target search as a function of Session.

Finally, there was an interaction between Trial Type, Prevalence and Session ($F(28,84)=3.9, p<.001$), see Figure 4.3.2b, below. This interaction was caused by the fact that there were persistent differences between the target-absent and target-present trials for 76% Prevalence, yet not for 24% and 50% Prevalence. In 76% Prevalence, target-absent trials had longer RTs than the target-present trials in both the first and final sessions (first Session: $F(1,2)=264, p<.001$; final Session: $F(1,2)=242.4, p<.001$). This was not the case for 50% Prevalence, or 24% Prevalence, which both showed signs of having longer target-absent than target-present RTs in the first session (50% Prevalence: $F(2,4)=8.9, p<.05$; 24% Prevalence: $F(2,4)=8.9, p<.05$), but not in the final session (Fs<1). As noted in the previous chapter, when prevalence is low, RTs for target-absent trials become as rapid as those for target-present trials.

Thus, it was not surprising that there were no differences in the RTs between target-absent and target-present trials in the 24% Prevalence condition. What is surprising, however, is that there were also no differences in RT between the target-present and target-absent trials in the 50% Prevalence condition, because it is typically the case that, in 50% Prevalence, RTs for target-absent trials are longer than those for target-present trials (Wolfe et al., 2007). Still, despite this, there was evidence of a persistent pattern in the 76% Prevalence condition, with

RTs being longer for target-absent trials than target-present trials. Apparently, participants spent a great deal of time searching for targets in the 76% Prevalence condition, suggesting that they were not simply pressing ‘present’ without attempting to complete the trials accurately when prevalence was high.

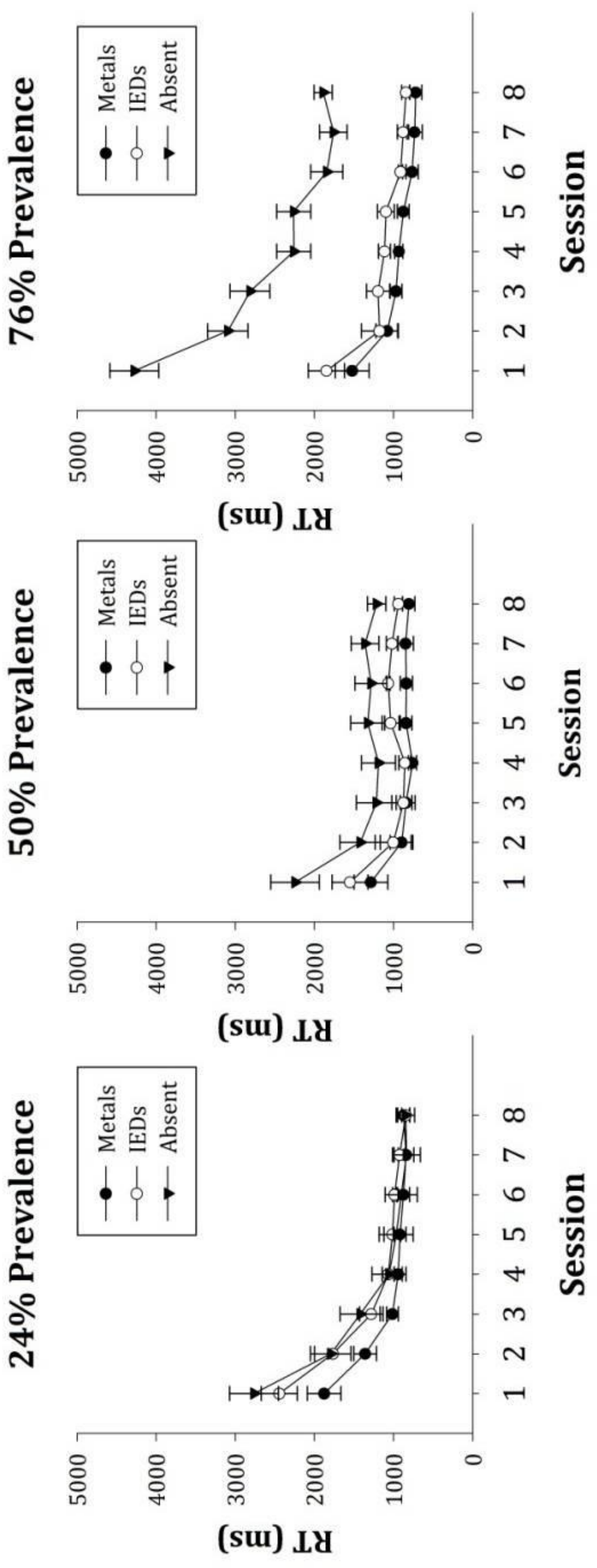


Figure 4.3.2b: RTs for the different Trial Types as a function of Session

4.3.3 Signal Detection Theory Parameters: d' and c

Hit and false alarm rates were utilised to compute Signal Detection Theory parameters for sensitivity (*d'*) and criterion (*c*). As with the previous chapter, the aberrant effects with *d'* observed by Wolfe et al. (2007) were replicated: as prevalence increased, the criterion decreased, yet *d'* also decreased. Figure 4.3.3a presents these results graphically. An 8 (Session: 1,2,3,4,5,6,7,8) × 3 (Search Type: single-target metals, single-target IEDs, dual-target search) × 5 (Prevalence: 2%, 24%, 50%, 76%, 98%) repeated-measures ANOVA was used to examine both the sensitivity and criterion parameters.

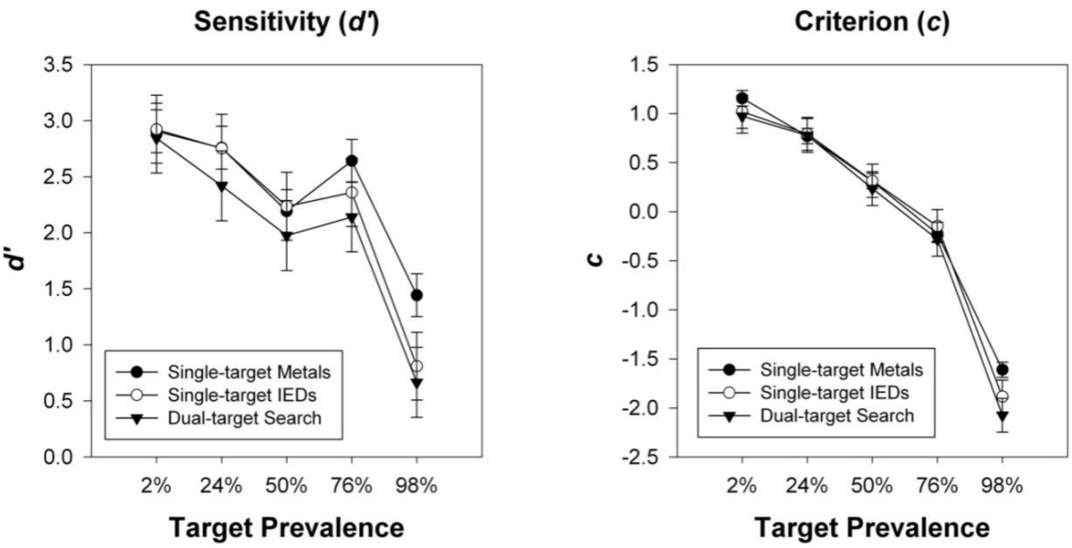


Figure 4.3.3a: Sensitivity and Criterion parameters as a function of Prevalence and Search Type. Note that there is no significant Search Type × Prevalence interaction for either sensitivity or criterion.

For *d'*, there was a main effect of Prevalence ($F(4,10)=9.2$, $p<.01$), which was explored using a series of main effects comparisons. These revealed that 98% Prevalence had a lower level of sensitivity than the other prevalence levels (*p*s<.05); however, there were no differences between the sensitivity for the other Prevalence levels (*p*s>.05).

Additionally, there was a main effect of Search Type ($F(2,20)=7.9$, $p<.01$), which was caused by the presence of the dual-target cost for sensitivity. Although there was no difference in sensitivity between single-target metals and single-target IEDs ($F<1$), sensitivity for dual-target search was lower than single-target metals ($F(1,10)=14.8, p<.01$); likewise, sensitivity was also higher in single-target IEDs than in dual-target search ($F(1,10)=8.8, p<.05$).

Finally, there was a main effect of Session ($F(7,70)=6.4$, $p<.01$), which was caused by the fact that sensitivity increased as the sessions progressed (ANOVA comparing Sessions 1 and 8: $F(1,10)=18.3, p<.01$). Mean sensitivity in Session 1 was 1.8 (S.E.M=0.13), and in Session 8 was 2.3 (S.E.M.=0.15).

The criterion parameter was also examined using an 8 (Session: 1,2,3,4,5,6,7,8) × 3 (Search Type: single-target metals, single-target IEDs, dual-target search) × 5 (Prevalence: 2%, 24%, 50%, 76%, 98%) ANOVA. This revealed to key points: first, there was a prevalence effect, with the criterion becoming more liberal as prevalence increased; second, there was evidence of differences between single-target and dual-target search, with dual-target search being more liberal than single-target search.

There was a main effect of Prevalence ($F(4,10)=79$, $p<.001$), which also interacted with the Session factor ($F(12,31)=2.6$, $p<.01$). The interaction was examined using a set of ANOVAs conducted upon each Session (see Figure 4.3.3b, below). Differences between Prevalence levels in each Session were examined using Tukey's HSD tests: the results of these tests are presented in Table 4.3.3c. Overall, there was no difference in criterion between the 2% and 24% Prevalence conditions; for sessions 1-4, the criterion was more liberal for 50% Prevalence than 24% Prevalence, but not for session 5-8; there were no differences in the criterion between 50% Prevalence and 76% Prevalence; and finally, 98% was consistently more liberal than all of the other Prevalence levels throughout the entire experiment.

Additionally, there was a main effect of Search Type ($F(2,20)=4.2$, $p<.01$). Although there was no difference between the criterion placement in single-target metals and single-target IEDs ($F<1$), yet the criterion was more liberal in dual-target search than single-target metals ($F(1,10)=7.4, p<.05$), and was also more liberal in dual-target search than single-target IEDs ($F(1,10)=5.6, p<.05$). Mean criterion placement for single-target metals was 0.08 (S.E.M.=0.04), for single-target IEDs was 0.02 (S.E.M.=0.08), and for dual-target search was -.08. (S.E.M.=0.08). The more liberal criterion for dual-target search may be able to account for some of the problems with detecting a dual-target cost for response accuracy in the present study. If participants here were adopting a more liberal criterion in dual-target search, then this could cause an increase in hit rates (as they would be responding 'present' at an elevated rate) and thereby attenuate the dual-target still. Still, it is important to note that a dual-target cost *was detected* for both the RTs and the sensitivity measure (*d'*).

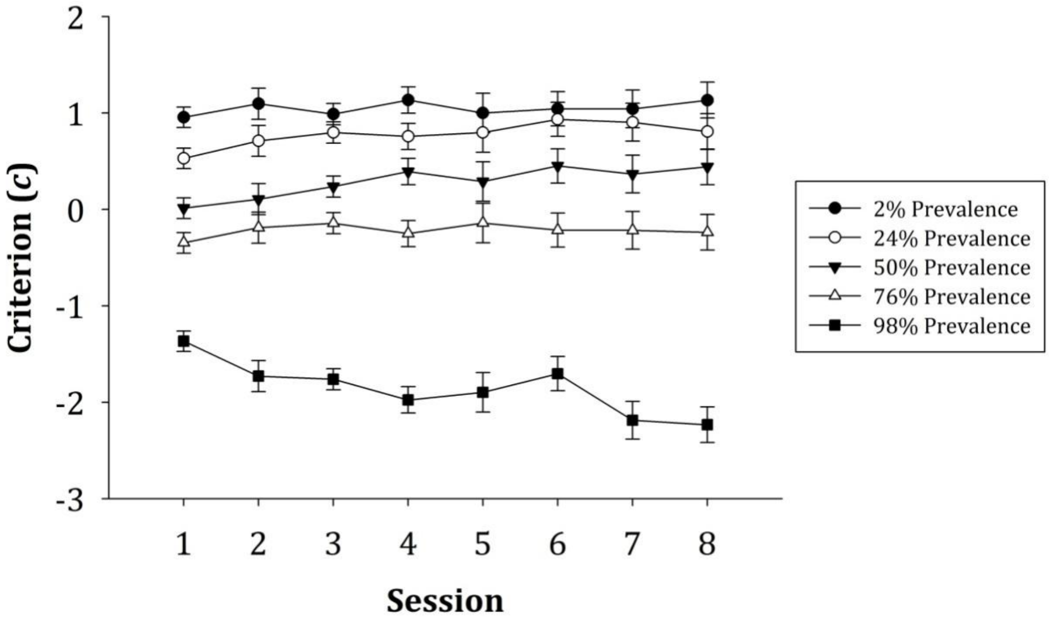


Figure 4.3.3b: Criterion level as a function of Session and Prevalence.

Table 4.3.3c
Results of post-hoc Tukey's HSD tests examining the different Prevalence Levels in each Session

Prevalence Comparison		Session							
		1	2	3	4	5	6	7	8
2%	24%	ns	ns	ns	ns	ns	ns	ns	ns
	50%	<.05	<.05	<.05	<.05	ns	ns	ns	ns
	76%	<.05	<.05	<.05	<.05	<.05	<.05	<.05	<.05
	98%	<.05	<.05	<.05	<.05	<.05	<.05	<.05	<.05
24%	50%	<.05	ns	<.05	ns	ns	ns	ns	ns
	76%	<.05	<.05	<.05	<.05	ns	<.05	<.05	<.05
	98%	<.05	<.05	<.05	<.05	<.05	<.05	<.05	<.05
50%	76%	ns	ns	ns	<.05	ns	ns	ns	ns
	98%	<.05	<.05	<.05	<.05	<.05	<.05	<.05	<.05
76%	98%	<.05	<.05	<.05	<.05	<.05	<.05	<.05	<.05

4.3.4 Signal Detection Parameters: Area under the ROC Curve

In order to examine whether or not sensitivity truly varies with variations in prevalence, the non-parametric signal detection parameter A_z was calculated for each prevalence level and search type. In the low- and high- prevalence conditions (i.e. 2% and 98% Prevalence), there were few of one type of trial in each Session (few target-present trials for 2% Prevalence, few target-absent trials for 98% Prevalence). As a result, there were very few of the different confidence ratings in each individual Session. Therefore, to increase the power of the analyses, and to calculate sensitivity measures on a broader number of ratings, the Sessions were grouped into four consecutive pairs. Sessions 1 and 2 were pooled to form Session 1, Sessions 3 and 4 were pooled to form Session 2, and so on.

The area under the ROC curve (A_z) was calculated for each participant in each of these Sessions and analysed using a 4 (Session: 1,2,3,4) × 5 (Prevalence: 2%, 24%,50%,76%,98%) × 3 (Search Type: Single-target metals, single-target IEDs, dual-target search) ANOVA, with Prevalence again entered as a between-subjects factor. Crucially, Prevalence did not reach significance as a main effect (F<1), or as an interaction with any of the other factors (Fs<2), indicating that the unusual behaviour exhibited by d' (i.e. increases in sensitivity coupled with decreases in the criterion when prevalence was reduced) was the result of a non-unit slope zROC curve being used to produce responses, rather than any true changes in sensitivity.

Additionally, there was a main effect of both Search Type ($F(2,20)=4.4, p<.05$), and Session ($F(3,30)=3, p<.05$), as well as an interaction between the two ($F(6,60)=3.7, p<.05$). This interaction was the result of there being significant main effects of Search Type in Session 1 ($F(2,20)=5, p<.05$), Session 2 ($F(2,20)=3.5, p<.05$), and Session 3 ($F(2,20)=6.9, p<.01$). However, there was no effect of Search Type for Session 4 (F<1), see Figure 4.3.4a, below. In fact, when Sessions 1, 2, and 3 were used together in an ANOVA examining A_z , there was no impact of Session (including main effects and interactions: F<1.5), and only an impact of Search Type ($F(2,20)=7.9, p<.01$). Further comparisons revealed that, for Sessions 1-3, when included in an ANOVA together, A_z was higher for single-target metals than single-target IEDs ($F(1,10)=10.5, p<.01$), and was higher for single-target metals than dual-target search ($F(1,10)=38.1, p<.001$). However, single-target IEDs did not differ in A_z to dual-target search (F<1). For the first three sessions, mean A_z for single-target metals was 0.85 (S.E.M.=0.02), for single-target IEDs was 0.77 (S.E.M.=0.04), and for dual-target search was 0.79 (S.E.M.=0.03).

Thus, measuring A_z gave some additional evidence that dual-target search was less sensitive than single-target search, though only for single-target metals. Still the most important point to consider here is that there were no changes in sensitivity as measured in this manner across the variations in target prevalence. Therefore it does indeed appear to be the case that, as with the previous chapter, true sensitivity was not increasing as prevalence decreased: the results that were observed were simply, as suggested by Wolfe et al. (2007) but confirmed here, the result of the fact that participants were operating using a on-unit slope zROC to base their decisions upon.

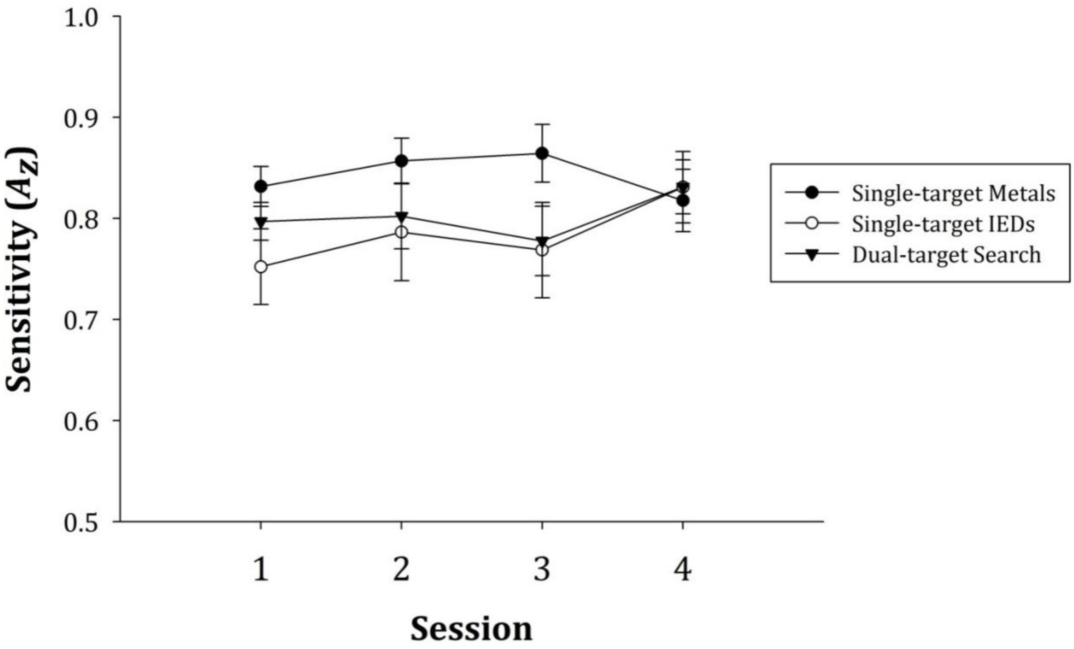


Figure 4.3.4a: Area under the ROC curve (A_z) as a function of Search Type and Session.

4.3.5 Matching False Alarms and Misses

In the previous chapter, there was mixed evidence to support Wolfe et al.'s (2007) claim that observers attempt to match the raw number of false alarm errors with the raw number of miss errors in visual search. Unfortunately, there were too few participants involved in the present experiment to be able to test such claims directly.

4.4 Discussion

The goal of the present study was to further explore the interaction between the prevalence effect and dual-target cost of a wide range of prevalence levels, and has produced a number of useful results in terms of exploring the criterion shift account and prevalence effect.

4.4.1 The Dual-target Cost

Unfortunately, the dual-target cost was not detected for response accuracy here. Still, despite this, the dual-target cost was detected for sensitivity (in fact, for both using both *d'* and A_z), and also for RTs. There are a number of likely reasons for the dual-target cost not being detected for response accuracy in the present study. One is a reduction in power, by using only three participants in each prevalence condition. This may have been compounded by the participant in the 50% prevalence condition who showed rather poor performance compared to the other participants. Additionally, there was some evidence to suggest that the participants in the present study adopted a more liberal criterion in dual-target search than they did in single-target search. This may have had the effect of reducing error rates in dual-target search, such that the dual-target cost could not be detected simply be examining response accuracy.

4.4.2 The Prevalence Effect

Despite the problems with the dual-target cost, in line with the previous chapter, the present results showed that, as prevalence decreased, the miss rate increased. Additionally, as prevalence decreased, the false alarm rate also decreased. This was embodied within a criterion shift, in signal detection terms. In the previous chapter, it was reported that there was an imbalance in responses between miss rates and false alarm rates, with false alarm rates being much higher in high prevalence than miss rates in low prevalence. The same pattern of results was observed here.

In the previous chapter, it was reported that the signal detection parameter *d'* increased as prevalence decreased, for single-target IEDs and dual-target search, but not for single-target metals, in partial agreement with the experiments conducted by Wolfe et al. (2007). Here, *d'* only dropped for the high prevalence condition (98% prevalence), and did so for all targets. In some senses, this result provides further support for Wolfe et al. (2007), yet in others, it does not: Wolfe et al. (2007) found that *d'* decreased between 2% prevalence and 50% prevalence. That was not the case here, not between 2% prevalence and 50% prevalence, or even between 2% prevalence and 76% prevalence.

Was the reported drop in *d'* indicative of a true drop in sensitivity in the 98% prevalence condition? Using an alternative index of sensitivity, one which is not contaminated by unequal variances, it was revealed that sensitivity did, in fact, not vary at all with variations in target prevalence. Thus, it seems likely that the apparent reduction in *d'* coupled with increases in criterion for low prevalence that were reported in the previous chapter and by Wolfe et al. (2007) were the result of unequal Signal and Noise distributions, and were not the result of true changes in sensitivity.

The present results can also aid in the resolution of the argument put forward by a group of researchers who have claimed that the prevalence effect does not really exist (Gur, Rockette, Armfield, et al., 2003; Gur, Rockette, Warfel, et al., 2003). The previous research used A_z and detected no differences across variations in prevalence: the same results were found here. However, these previous studies failed to examine the criterion (Wolfe et al., 2007), and so may have missed the underlying shift in behaviour that occurs when prevalence is varied. Using both the criterion and A_z, the present study has detected evidence to suggest that sensitivity does not change across variations in prevalence, yet the criterion does.

Finally, as with the previous chapter, and previous examinations of the prevalence effect (Wolfe, et al., 2005; Wolfe, et al., 2007), participants responded 'absent' more rapidly when prevalence was low. One could argue that the elevation in hit rates as prevalence increased was due to motor errors for responding 'present' (this would be the inverse account of the arguments made by Fleck & Mitroff, 2007). However, in the 76% prevalence condition, target-absent RTs were very long indeed, which could imply that participants were actively searching the display, expecting to see a target. On trials where a target was present, this would have the obvious advantage of increasing the chance that the target would be detected.

4.4.3 General Discussion and Relevance to Airport X-ray Security Screening

The present study has been useful in further exploring the imbalanced nature of the prevalence effect across a wide range of prevalence levels, and has also been useful in resolving whether or not sensitivity truly increases when prevalence is low. This is an important result: if sensitivity was increasing, then it would be rather difficult to explain logically. How could observers essentially perform *better* in a task when biased towards responding 'absent'? Fortunately, sensitivity did not vary across prevalence, when using a non-parametric measure of sensitivity.

Overall, the experiments conducted in the present thesis so far, as well as a number of previous experiments, have now shown that there is a dual-target cost when searching for two targets simultaneously (Menneer, et al., 2004, 2007), and that observers tend to miss targets that are presented infrequently (Wolfe, et al., 2005; Wolfe, et al., 2007). One important question lingers, however: to what extent do the dual-target cost and prevalence effect extend to the performance of actual airport security screening personnel? The next chapter seeks to explore this question, in order to test the ecological validity of the extant visual search literature on a group of experts.

5

Are Airport X-ray Security Screeners impacted by the Prevalence Effect and the Dual-target Cost?

Applied Tests of Core Issues

5.1 Introduction

The previous chapters have made a strong case for the notion that low target prevalence (or TIP prevalence at least) and the dual-target cost may impair the target detection performance of airport X-ray security screeners. The key question that follows from the investigation of the prevalence effect and dual-target cost is, therefore: do the factors that affect participants with limited training also affect security screening personnel? Of course, if the answer to this question is 'yes', then not only can the value of the present work be established, but, more importantly, suggestions can me made in terms of how to improve visual search performance in airport security screening. The present chapter therefore presents the results from an experiment where the dual-target cost and prevalence effect were tested on a group of airport security screeners.

5.1.1 Previous Research: Visual Search and Expertise

A number of studies examined in the earlier Literature Review focused upon the examination of the search task carried out by radiographers. The prevailing view of the radiographic literature is that radiographers first scan an overall view of a given X-ray, after which their perceptual and search systems 'flag' areas of the X-ray which require further inspection. Areas which are flagged tend to be atypical areas of the image which would not normally be present in an X-ray of an individual with no tumour (Nodine & Kundel, 1987). Recently, Gale et al. (2000) have successfully applied the radiographic model to the performance of security screeners searching baggage for IEDs (for more detail, these the preceding Literature Review in Chapter 1).

McCarley, Kramer, Wickens, Vidoni and Boot (2004) have recently examined skill acquisition in the X-ray screening task in detail. They examined the performance of a set of novice participants, using RT, accuracy and eye-movement data, as the participants searched for threats in X-ray images of baggage. Targets were images of knives, inserted digitally into baggage, appearing on 10% of the trials. McCarley et al. (2004) reported a number of useful results: first of all, sensitivity (measured by A_z) improved with practice. However, when a new set of target images (the targets were still knives) were used, sensitivity decreased significantly. As a result, McCarley et al. (2004) argued that screener performance would best be served by presenting screeners with a large set of target images, in order to facilitate learning. Furthermore, they suggest that improving screeners' knowledge of targets will aid the disentanglement of occlusion problems in passenger baggage. It is often the case that X-ray images of baggage present a large number of occluding objects, and thus greater knowledge of potential target shapes, in a variety of orientations, will likely increase the probability that threat items will be detected amongst the clutter of actual baggage. A similar account has emerged from studies conducted using an adaptive training regime: Schwaninger and colleagues (Schwaninger, 2004; Schwaninger & Hofer, 2004) have reported increases in sensitivity following the presentation of threat items from a variety of different orientations, with varying levels of overlap and complexity.

5.1.2 Core Concerns: the Dual-target Cost and Prevalence Effect

Whilst the studies described in the previous section have investigated the performance of individuals searching X-ray screening images, so far, the impact of the dual-target cost and prevalence effect have not been examined using actual screening personnel. How might actual screener performance be impacted by the dual-target cost and prevalence effect?

The Dual-target Cost and Expertise: An early set of seminal experiments conducted by Shiffrin and Schneider (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977) reported that, with extensive practice, participants were able to search effectively for several target letters simultaneously; similarly, Neisser, Novick and Lazar (1963) found that, after practice, participants were able to search for at least ten letters in a list with high efficiency. Does this imply that the dual-target cost can be eliminated with practice? A number of recent experiments have been conducted in order to answer just such a question: the simple outcome is that the dual-target cost can not be eliminated, even with extensive practice. Menneer, Cave and Donnelly (under review) trained participants in the search for X-ray threat items for eleven sessions of around one hour each, and reported that the dual-target cost was present, even in the final session.

Why is it the case that the experiments conducted by Menneer et al. (under review) found the dual-target cost to not be eliminated by practice, whilst earlier work indicated that practice enabled observers to search efficiently for a number of targets (Neisser, et al., 1963; Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977)? It seems likely that, in fact, the early studies were able to train participants to become efficient in the detection of multiple targets because the targets and non-targets were simple letters, rather than complex X-ray images from airport screening. Thus, it seems sensible to suggest that airport screeners, despite their considerable experience, will still be impacted by the dual-target cost.

Prevalence and Expertise: Although it appears to not be possible to eliminate the dual-target cost for X-ray screening images, is it at all possible to eliminate the prevalence effect with practice? In an early study, Fortune (1979) engaged a small group of participants in the search for abnormal (diseased) tissue samples under a microscope. The participants were all trained toxicologists, and the results showed a clear effect of prevalence: low-prevalence targets were more likely to be missed than high-prevalence targets.

Expertise versus Motivation: It has long been argued that, from a Signal Detection perspective, the optimal route towards educing a criterion shift within participants is to either vary the probability that a signal will occur (i.e. the prevalence effect), or to vary the payoff scheme given to participants (Macmillan & Creelman, 2005). If participants are rewarded for hits and correct rejections, and penalised for misses or false alarms, then one can see that the participants will quickly attempt to maximise their hits and correct rejections, whilst minimising their misses and false alarms. It is thus rather surprising that attempts to motivate participants using a point-based scoring system have been unable to eliminate the prevalence effect (Wolfe, 2007; Wolfe, et al., 2005), as have been attempts at giving participants 'speeding tickets' for responding 'absent' too rapidly (Wolfe, 2007).
Is the prevalence effect immune to motivational factors? Fleck and Mitroff (2008) recently engaged a set of Video Game Players (VGPs) in a visual search task, and,

rather surprisingly, reported that the VGPS showed no evidence of a prevalence effect, whereas a control group of non-VGPs did show prevalence effects. Fleck and Mitroff (2008) argue that there are two possible causes for their findings. First of all, it may have been the case that video game experience alters a set of perceptual and visual skills that somehow renders VGPs immune to the effects of prevalence. This possibility seems plausible given the plethora of studies that have reported improved visual skills in VGPs (Castel, Pratt, & Drummond, 2005; C. S. Green & Bavelier, 2003, 2007; Riesenhuber, 2004). A second possibility is that VGPs saw the search task more of as a game rather than a dry, empirical study, and were highly motivated to succeed. If that is true, then the previous studies which have produced null effects of incentives simply did not motivate the participants sufficiently as to eliminate the prevalence effect (i.e. they failed to manipulate the motivation factor sufficiently, lacking power). If Fleck and Mitroff (2008) are correct in their suggestion that VGPs may be immune to the effects of prevalence as a result of intrinsic motivation to succeed, then airport screeners may also be immune to the prevalence effect for exactly the same reason.

5.1.3 The Present Study

There are a number of salient points that must be considered with regards to the present study. First of all, no novice participants were recruited: the only participants were security screeners with at least one year of experience in screening. Recruiting novices would not be informative with regards to the research question (i.e. whether or not screeners affected by the dual-target cost and prevalence effect); there is a sufficient body of previous research which has demonstrated that novices are indeed subject to the effects of prevalence and the dual-target cost.

Additionally, the task given here is somewhat different to the visual search task that screeners typically carry out. Although the stimuli were X-rays of threat and non-threat items (the exact same stimuli as used in the previous experiments reported in the present thesis), no images overlapped with one another, as would occur in real baggage. There is every reason to suspect that such a procedure will have made the task considerably easier to complete than the actual X-ray screening task. Indeed, McCarley et al. (2004) have noted that occlusion and overlap between objects and threat items in passenger baggage is likely to impair threat detection rates considerably (see also Schwaninger, et al., 2008). Still, no overlap was used here because the underlying impact of occlusion and overlap are, at present, poorly understood. There is no reason to suspect that using overlapping images will change the presence of a prevalence effect or dual-target cost. Indeed, Wolfe et al. (2007) did use overlapping, and found results that (generally speaking) were replicated by the experiments using non-overlapping images reported in Chapters 2 and 3.

Finally, airport screeners, when conducting their screening task, have at their disposal a wide range of image-manipulation algorithms that are built into the software of the X-ray scanners. Those algorithms were not available here, and screeners were made aware of that fact. Typically, the algorithms can be used to adjust the colour scheme of the display, as well as remove certain colours and highlight others. However, it has been reported that screeners may, in fact, not always benefit from using these enhancement algorithms, and, in some instances, their performance actually suffers when the images are 'enhanced' in this manner (Michel, Koller, Ruh, & Schwaninger, 2007). In real screening, the screeners can also request that baggage be re-scanned. For example, if an object in an X-ray image is somewhat difficult to recognise, a screener can reverse the bag in its progress along the search conveyor belt, and have the bag's orientation changed so that the resultant X-ray can be examined more easily. Again, this procedure was not available here. It should be noted that, although the screeners who participated in the present study all claimed to regularly use the image manipulation algorithms available to them, and to also regularly require baggage to be re-oriented and scanned for a second time, it is unclear as to whether or not their claims were accurate.

5.2 Method

5.2.1 Participants

Participants were eighteen X-ray security screeners, consisting of ten males and eight females recruited from an airport. All screeners had a minimum of 12 months' experience with the screening task. Mean age was 51.8 years (SD=8.6 years), mean months of experience was 32.5 months (SD=16 months). The participants took part in the experiment when they were available to be freed from their normal working duties (i.e. during 'downtime' or quiet periods); thus, they took part at a point in each day when they would often normally be conducting their screening task.

In line with the UK Department for Transport's policy, no participants were paid for their participation in the experiment. All participants were made fully aware of the fact that their employment and work would be unaffected by the outcome of the study, as well as the fact that their participation would be entirely anonymous, with no possible route towards linking any screener to their actual data.

5.2.2 Apparatus and Stimuli

The apparatus and stimuli were the same as in Chapters 2 and 3.

5.2.3. Design and Procedure

Participants took part in three blocks of trials, lasting around 30-45mins each. Due to the fact that the screeners were also present at the airport to work when they participated, the blocks of trials were, in some cases, conducted consecutively, and, in other cases, were broken up over the course of a single day, or across several weeks (data collection occurred from around 11am-3pm weekly on a Wednesday, which was the quietest period in the week for the screeners). Every effort was made to keep the time between blocks of trials as short as possible.

Before the trials began, twenty samples of each target class were presented. Key points were made salient (for example, screeners would normally be expected to search for 'incomplete' IEDs, essentially broken up into separate explosive and detonator components; here, all of the IEDs were 'complete', and had the explosive and detonator components connected).

The three blocks consisted of three different sets of trials: single-target search for metals, single-target search for IEDs, and dual-target search for metals and IEDs. The blocks began with five practice trials, followed by 160 experimental trials. Participants were given the opportunity to take a break every 50 trials. Only one target could appear on any trial. Block order was counterbalanced across the participants. The trial design was identical to those used in Experiment 2 of Chapter 3.

The study used a mixed design, with three independent variables, consisting of: Target Type (metals, IEDs, absent), Search Type (single-target search, dual-target search), and target Prevalence (5%, 20%, 50%). Due to time constraints, a high prevalence condition (i.e. >50% prevalence) was not employed; similarly, fewer trials per block were employed than previous experiments, and each participant was involved in only one session. The 5% prevalence condition is intended as a close approximation to the 1% prevalence level of TIP images. Target Prevalence was a between-subjects variable. Dependent variables were response accuracy and response time.

5.3 Results

As with the previous chapters, the results will initially be examined in terms of error rates (miss errors and false alarm errors), after which the Signal Detection parameters *c* and *d'* will be used to examine the results in more detail. This will enable further tests of Wolfe et al.'s (2007) criterion shift account, as well as enabling an assessment of the dual-target cost and prevalence effect upon actual airport security screening personnel. All *t*-tests show Bonferroni-corrected values, and Greenhouse-Geisser degrees of freedom and *p* values are used whenever sphericity is violated. In all figures, error bars represent ±S.E.M.

5.3.1 The Prevalence Effect and Dual-target Cost

Miss and False Alarm Rates. Miss and false alarm rates were initially examined in two separate 2 (Target Presence: present, absent) × 3 (Search Type: single-target metals, single-target IEDs, dual-target search) × 3 (Prevalence: 5%, 20%, 50%) repeated-measures ANOVAs. Prevalence was entered as a between-subjects factor in both cases. Figure 5.3.1 shows the impact of target prevalence and dual-target search upon error rates, with the left panel presenting miss errors, and the right panel showing false alarm errors. Unlike the previous experiments, there was no main effect of Prevalence ($F(2,15)=2.2$, $p=.14$), yet still there was a main effect of Search Type ($F(1.3,20)=10.5$, $p<.01$), as well as a main effect of Presence ($F(2,30)=21.1$, $p<.001$). Additionally, there were interactions between

Search Type and Presence ($F(2,30)=4.8$, $p<.05$), and a slight trend between Prevalence and Presence ($F(2,15)=2.8$, $p=.092$).

Two further ANOVAs were used to explore the interaction between Search Type and Presence: one ANOVA examined the miss rates, and the other examined the false alarm rates. These were identical to the initial ANOVA that was conducted, except that the Presence factor was removed. For the false alarm rates, there was no impact of Search Type ($F<2.7, p>.05$). However, for the miss rates, there was a significant main effect of Search Type ($F(2,30)=8, p<.01$). Three further ANOVAs were conducted to examine the differences between the levels of the Search Type factor. These revealed that miss rates were higher in dual-target search than in single-target metals ($F(1,15)=20, p<.001$), and that miss rates were higher for single-target IEDs than single-target metals ($F(1,15)=8.9, p<.001$). However, single-target IED miss rates were no different to error rates in dual-target search ($F<2.7, p<.05$). The mean miss rate for single-target metals was [redacted] (S.E.M.=[redacted]), for single-target IEDs was [redacted] (S.E.M.=[redacted]), and for dual-target search was [redacted] (S.E.M.=[redacted]). The mean false alarm rate for single-target metals was [redacted] (S.E.M.=[redacted]), for single-target IEDs was [redacted] (S.E.M.=[redacted]), and for dual-target search was [redacted] (S.E.M.=[redacted]).

Thus, although there was no impact of Prevalence upon the error rates, there was evidence to suggest that screeners were able to detect metals more regularly when only searching for metals, than when searching for IEDs as well as metals (i.e. there was a dual-target cost, but only for metals when analysed in this manner).

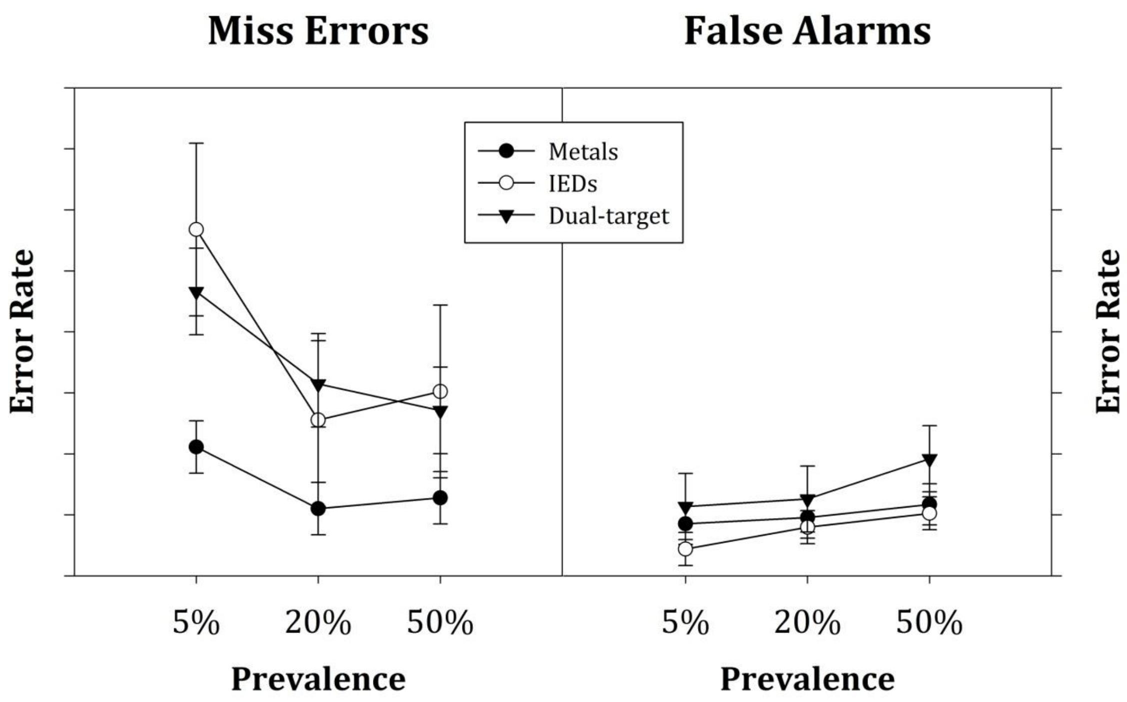


Figure 5.3.1: Miss errors and false alarms for the different Search Types as a function of Target Prevalence.

5.3.2. Impact of Prevalence on Each Target

In order to examine whether or not the targets were differentially impacted by prevalence in dual-target search, a 2 (Target Type: metals, IEDs) × 2 (Search Type: single-target search, dual-target search) × 3 (Prevalence: 5%, 20%, 50%) repeated-measures ANOVA was carried out. Prevalence was entered as a between-subjects factor. The results of this ANOVA are depicted in Figure 5.3.2, below, in which the dual-target cost was replicated. There was no effect of Prevalence, and Prevalence did not interact with any of the other factors (all *F*s<2.8). However, there was a dual-target cost ($F(1,15)=10.6$, $p<.01$), as well as an overall difference in performance between search for metals and search for IEDs ($F(1,15)=10.5$, $p<.01$: mean error rate for metals was [redacted], S.E.M.=[redacted]; mean error rate for IEDs was [redacted], S.E.M.=[redacted]), and an interaction between the dual-target cost and Target Type ($F(1,15)=6.8$,$p<.05$).

The interaction was then examined using two further ANOVAs, conducted upon the error rates for metals and IEDs separately. These were of identical design to the initial ANOVA, except that the Target Type factor was removed. For metals, there was a clear dual-target cost ($F(1,15)=42.8$, $p<.001$), yet for IEDs, there was no evidence of a dual-target cost ($F<1$). The mean error rate for single-target metals was [redacted] (S.E.M.=[redacted]), for dual-target metals was [redacted] (S.E.M.=[redacted]), for single-target IEDs was [redacted] (S.E.M.=[redacted]), and for dual-target IEDs was [redacted] (S.E.M.=[redacted]).

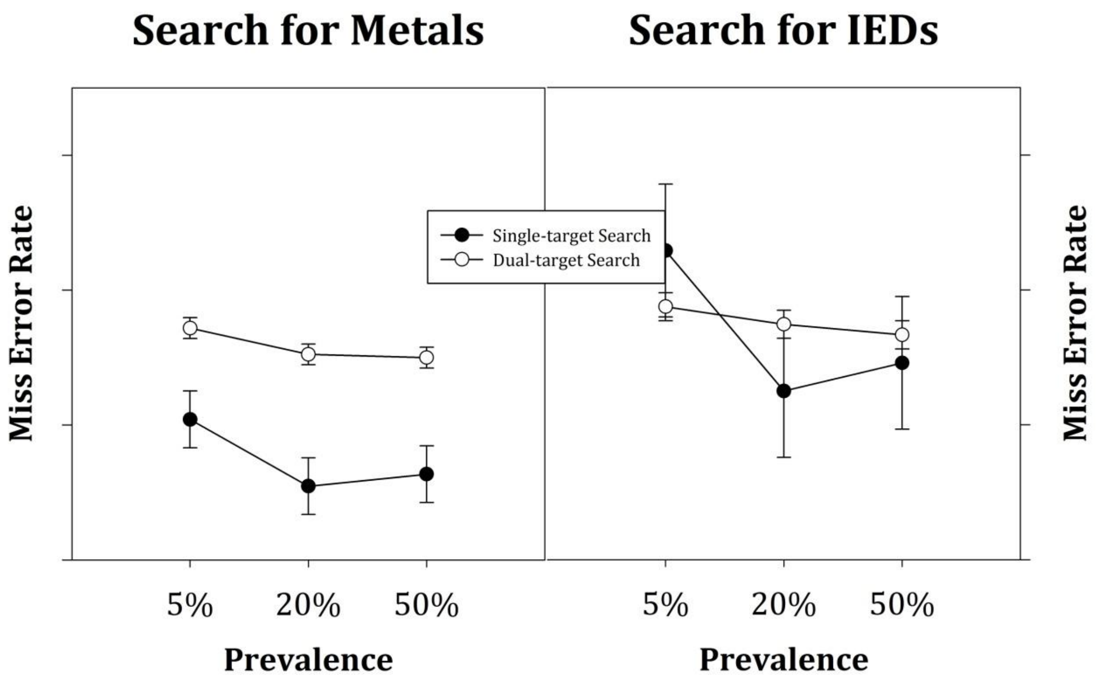


Figure 5.3.2: Miss error rates for Metal or IED targets as a function of Prevalence and Single- or Dual-target Search.

5.3.3 Signal Detection Theory Parameters

Sensitivity and Criterion: Signal Detection parameters *d'* and *c* were computed for the screening personnel. Overall results of these parameters are depicted in Figure 5.3.3a. Each parameter was examined using a 3 (Prevalence: 5%, 20%, 50%) × 3 (Search Type: Single-target Metals, Single-target IEDs, Dual-target Search) ANOVA, with Prevalence entered as a between-subjects factor. Examination of *d'* revealed that there was a main effect of Search Type ($F(2,30)=10$, $p<.001$). There was no main effect of Prevalence ($F<1$). Subsequent *t*-tests revealed that, overall, dual-target search was less sensitive than the search for either IEDs or metals alone. In other words, this is a clear replication of the reduction in sensitivity seen in dual-target search previously: the dual-target cost affects airport security screeners, just as it affects naïve undergraduate participants. Mean *d'* for single-target metals was X.X (S.E.M.=[redacted]), for single-target IEDs was [redacted] (S.E.M.=[redacted]), and for dual-target search was [redacted] (S.E.M.=[redacted]).

A series of *t*-tests compared sensitivity across these search conditions. Single-target metals versus single-target IEDs: $t(17)=1$, $p>.05$; single-target metals versus dual-target search: $t(17)=6.1$, $p<.01$; single-target IEDs versus dual-target search: $t(17)=3.1$, $p<.01$.

The criterion showed a main effect of prevalence ($F(2,15)=3.8$, $p<.05$) with the criterion becoming more conservative as prevalence was reduced. The criterion was no more conservative in 5% prevalence (mean=[redacted], S.E.M=[redacted]) than in 20% prevalence (mean=[redacted], S.E.M=[redacted]): $t(34)=2$, $p>.05$. Similarly, the criterion was also no more conservative in 20% prevalence than in 50% prevalence (mean=[redacted], S.E.M=[redacted]): $t(34)=0.6$, $p>.05$. However, the criterion was more conservative when comparing 5% prevalence with 50% prevalence: $t(34)=2.7$, $p>.05$. Thus, the prevalence effect has been replicated in security screening personnel, albeit at a reduced extent than seen in previous studies.
The criterion parameter also showed evidence of a strong trend for Search Type ($F(1.4,21)=3.6$, $p=.057$), yet failed to reach significance. Part of this is likely due to the reduced power (i.e. limited number of trials, and number of participants) in the present study.

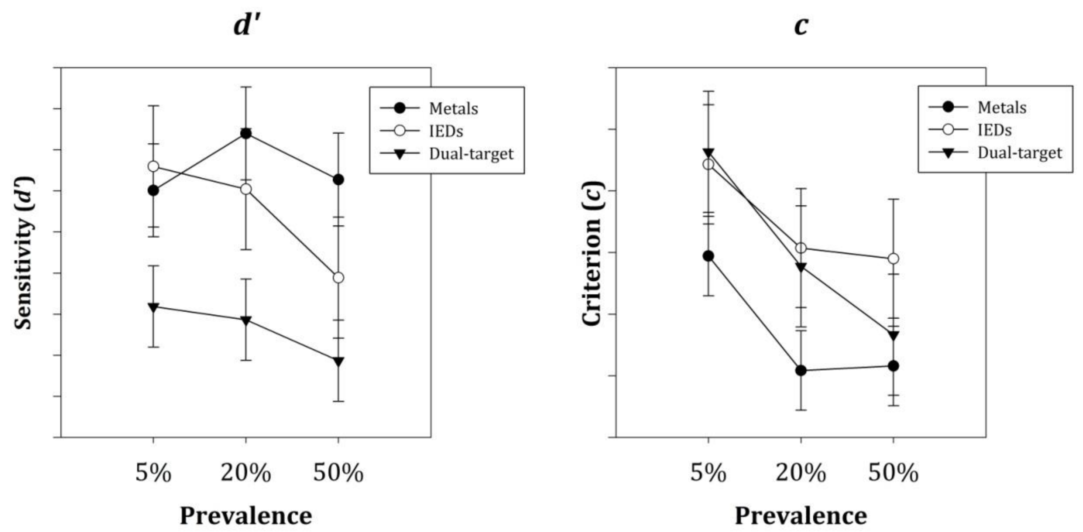


Figure 5.3.3a: Sensitivity (*d'*) and Criterion (*c*) parameters as a function of Target Prevalence and Search Type.

zROC Curves and Slopes: As with the Chapter 3, zROC curves were plotted for the participants across the prevalence levels. These produced slopes of 0.36 for

Single-target IEDs, 1.6 for Single-target Metals, and 0.6 for Dual-target search. These are all markedly different from the slopes seen in the previous experiments, so it is apparent that the screeners are, as may be expected, making decisions in the threat detection task in a somewhat different fashion to the participants who were involved in the previous experiments that have been reported here.

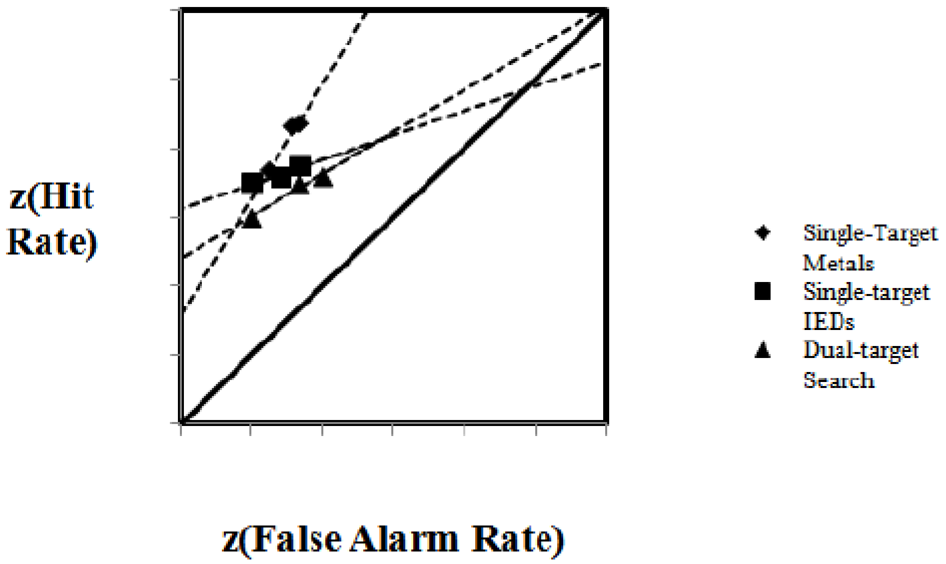


Figure 5.3.3b: Fitted slopes for zROC curves across different prevalence levels for the different search types.

5.3.4 Reaction Times

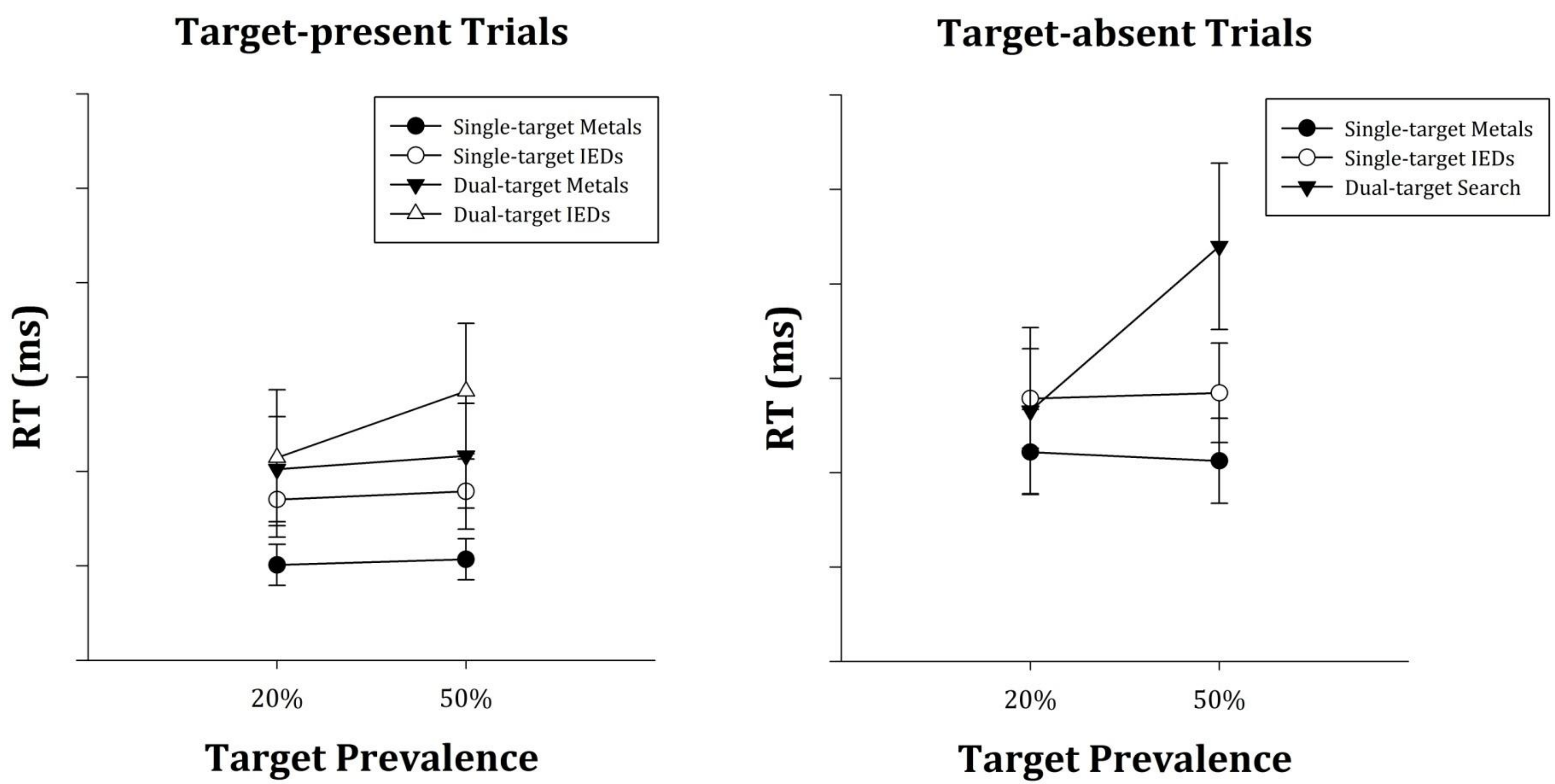
As the number of target-present trials in the 5% prevalence condition was very low indeed, and the error rates were rather high, RT data were not analysed for the 5% prevalence group. Examinations of Reaction times could be made more lucid by averaging across the single-target absent trials. Doing so would then enable a simple examination of Single-versus-Dual-target search, combined with examinations of Target-present versus target-absent trials. In order to test whether or not this was permissible, an initial 2 (Search Type: Single-target Metals, Single-target IEDs) × 2 (Prevalence: 20%, 50%) ANOVA was conducted. Unfortunately, this indicated significant differences between the single-target Metals and IEDs target-absent trials in terms of RT ($F(1,10)=7.4$, $p<.05$).

Therefore, a somewhat different approach was needed to examine the RTs, beginning with a 7 (Search Type: Single-target Metals Hit, Single-target IEDs Hit, Dual-target Metals Hit, Dual-target IEDs Hit, Single-target Metals Correct Rejection, Single-target IEDs Correct Rejection, Dual-target Search Correct Rejection) × 2

(Target Prevalence: 20%, 50%) ANOVA. Overall results for the ANOVA are depicted in Figure 5.3.4a. This ANOVA revealed that there was no main effect of Prevalence ($F<1$), yet there was a main effect of Search Type ($F(2,23)=11.8$, $p<.001$). The main effect of the Search Type factor revealed that, for target-present trials, there was a dual-target cost ($F(1,10)=12.2, p<.01$), as well as differences in RTs between metals and IEDs ($F(1,10)=11.7, p<.01$), with metals being detected more rapidly than IEDs.

The Prevalence Effect: The presence of a prevalence effect is normally seen in terms of a decrease in target-absent RTs, such that they are equally as rapid as target-present RTs (Wolfe, et al., 2005; Wolfe, et al., 2007). As described above, Prevalence had no impact upon the RTs, making it seem unlikely that this is the case. However, were the screeners involved in the present study responding 'present' more rapidly than they were responding 'absent'? Target-absent and target-present RTs for each search block were compared. For all search types, target-present RTs were more rapid than target-absent RTs. Target-present responses were faster for single-target metals than target-absent responses (metals present mean RT=[redacted], S.E.M.=[redacted], metals absent mean RT=[redacted], S.E.M.=[redacted]): $F(1,10)=34.6, p<.001$. Similarly, target-present responses were faster for single-target IEDs than target-absent responses (IEDs present mean RT=[redacted], S.E.M.=[redacted], IEDs absent mean RT=[redacted], S.E.M.=[redacted]): $F(1,10)=15.7, p<.01$. Finally, target-present responses for metals in dual-target search were faster than target-absent responses in dual-target search (metals present mean RT=[redacted], S.E.M.=[redacted], dual-target absent mean RT=[redacted], S.E.M.=[redacted]): $F(1,10)=18.9, p<.01$. Likewise, target-present responses for IEDs in dual-target search were faster than target-absent responses in dual-target search (IEDs present mean RT=[redacted], S.E.M.=[redacted]): $F(1,10)=12.9, p<.001$.

Thus, it appears to be the case that prevalence had no impact on RTs, and that target-present RTs were faster than target-absent RTs. The screeners in the present study may have been able to attenuate the prevalence effect in terms of response accuracy (i.e. the prevalence effect was only seen here in the signal detection analyses, and not in the raw accuracy rates) by preventing themselves from responding too rapidly. This will be further discussed below.


Target-present Trials
RT (ms)
Single-target Metals
Single-target IEDs
Dual-target Metals
Dual-target IEDs
20%
50%
Target Prevalence
Target-absent Trials
RT (ms)
Single-target Metals
Single-target IEDs
Dual-target Search
20%
50%
Target Prevalence

5.3.5 Matching False Alarms and Misses

Given the limited number of trials (160 per search type per participant), and the fact that only low levels of prevalence were used here for just one session, there were too few observations (indeed, for the 5% prevalence condition, there were only eight target-present trials; participants did miss some of these targets, but not all of them in most cases), and too few participants, to reliably test whether or not the participants were attempting to equate their number of false alarms with their number of misses, in line with Wolfe et al.'s (2007) claims.

5.4 Discussion

5.4.1 Airport Security Screeners are not Immune to the Dual-target Cost

In line with previous research that has found the dual-target cost to be a pervasive phenomenon in the face of extensive experience and training (Menneer, et al., under review), the present study also found that airport security screeners showed a reduction in response sensitivity in dual-target search, compared to single-target search. This is an important result, because the dual-target cost was evident regardless of the target prevalence, implying that real screeners are likely to be conducting a less sensitive search than they would be if they were searching for only one class of threat item. In a similar vein to the previous experiments reported in the present thesis, the screeners seemed also to find IEDs more difficult to detect than metals. Furthermore, RTs also showed evidence of a dual-target cost. Overall, therefore, it appears that airport security screener performance would benefit in terms of an increase in response sensitivity (increased hit rates; decreased false alarm rates), as well as a decrease in response times, if screeners were segregated and each given a separate target class to search for.

What is interesting, however, is that screeners are particularly skilled when detecting IEDs. Although IEDs exhibited higher miss rates than metals, there was, in fact, no dual-target cost the IEDs in terms of miss rates. The most likely reason for this is the training given to the screeners, throughout their careers. Due to the inherent difficulty of detecting IEDs, it is made clear to all screeners that they must be very cautious regarding the presence of IEDs in passenger baggage. Thus, it may be the case that, when told that an IED will appear at some point during a block of trials, the screeners become particularly cautious in their detection of IEDs. This may be a deliberate, conscious strategy, or, alternatively, it may have developed as a result of training and experience.

5.4.2 Airport Security Screeners are not Immune to the Prevalence Effect

The screeners here showed the presence of a criterion shift as prevalence varied; thus, there was a prevalence effect, consistent with previous work, as well as the previous chapters of the present thesis (Wolfe, et al., 2005; Wolfe, et al., 2007). It was surprising in some senses to see that the miss rates were not significantly impacted by prevalence, despite the clear trend towards increases in miss rates as prevalence decreased. The most likely reason for this is a lack of power. In the present study, the screeners were involved in fewer trials and fewer sessions than the previous experiments. Thus, it was only when the power of the examination was increased by considering the criterion parameter (which is based upon *all* of the trials in the experiment, both target-present and target-absent) that the impact of prevalence emerged. Still, the miss rates appeared to plateau around the 20% prevalence mark, whilst the false alarm rates remained exceedingly low throughout the variations in prevalence. As a result, the criterion only shifted between 5% and 50% prevalence, and not between 5% and 20% prevalence, or between 20% and 50% prevalence. Again, it is important to remember that the prevalence effect is continuous rather than categorical in nature, so there is a gradual shift of the criterion from 5% up to 50% prevalence, as would be expected.

In the previous empirical chapters, it was argued that human perception of target prevalence is intrinsically imbalanced: false alarm rates decrease by only a small amount as prevalence becomes low (<50%), whilst miss rates are still heavily impacted low prevalence. In the present study, the raw false alarm rates across the search types and prevalence conditions were exceedingly low. Low false alarm rates for screeners should not really be surprising. Although the research described in the introduction made a strong case for presenting screeners with as many varied forms of potential target images as possible (Schwaninger, 2004; Smith, Redford, Gent, et al., 2005; Smith, Redford, Washburn, et al., 2005), a large body of research has been conducted suggesting that learning and expertise can come not only through learning the target(s), but also learning the distractors.

Indeed, a number of studies have reported evidence to suggest that repeating the distractors on consecutive trials can speed search, just as repeating a target on consecutive trials (Maljkovic & Nakayama, 1994; McBride, Leonards, & Gilchrist, in press). Thus, in some senses, exceedingly low false alarm rates should be expected for actual airport screening personnel. The vast majority of the objects that they are presented with are distractors, and thus they will likely have considerable expertise in understanding the nature and variations in distractor stimuli (i.e. non-threat images within X-rays of passenger baggage).

Overall, a strong case can be made here that increasing the TIP prevalence will cause a criterion shift in actual screener performance: this will reduce miss rates, whilst increasing false alarms at a negligible rate.

5.4.3 Prevalence, Dual-target Search and Motivation

In a study of the effects of prevalence upon Video Game Players, Fleck and Mitroff (2008) reported that VGPs were immune to the prevalence effect, and suggested that either playing video games gave VGPs an enhanced visual skill so as to be able to prevent the effect from occurring, or that VGPs were highly motivated to succeed, thereby over-riding the effects of prevalence. Indeed, it is interesting to note that in Fleck and Mitroff's (2008) study, VGPs in a low-prevalence condition maintained longer target-absent RTs than target-present RTs, whilst non-VGPs rapidly shifted to the classic prevalence effect in RTs, with target-absent RTs becoming faster than target-present RTs. In the present study, a comparison of 20% prevalence and 50% RTs showed that, at both prevalence levels, participants were responding more rapidly to target-present trials than to target-absent trials, in a similar manner to the VGPs in Fleck and Mitroff's (2008) study. Indeed, there were no significant differences between target-absent and target-present RTs between the 20% and 50% prevalence conditions. The response accuracy and criterion data reflect the RT data, in the sense that there was no criterion shift between 20% and 50% prevalence, just as there were no differences in RT between 20% and 50% prevalence.

It could be the case that the high intrinsic motivation of both Fleck and Mitroff's (2008) VGPs and the screeners in the present study caused both groups to delay responding 'absent', thereby allowing themselves sufficient time to detect a target, and to be immune to the prevalence effect. Although it is unfortunate that

RTs for the 5% prevalence group could not be examined due to lack of power, it may have been the case that participants in the 5% prevalence condition began to respond 'absent' too rapidly, thereby missing the targets once they did finally appear. Still, although a strong case can be made for the impact of motivation upon performance here, it should be noted that these results do not necessarily rule out the possibility that experience and expertise can alleviate the prevalence effect to a certain extent.

If intrinsic motivation can reduce the prevalence effect somewhat, why can it not alleviate the dual-target cost? The key to answering such a question seems to lie with the fundamental differences between the prevalence effect and the dual-target cost. The prevalence effect seems to be inexorably bound to issues with search termination thresholds (Wolfe, et al., 2005; Wolfe, et al., 2007); the dual-target cost is more akin to a structural problem with the visual system. In that sense, the dual-target cost arises as a result of structural limitations in visual search and the cognitive systems that subserve the visual search process (D'Zmura, 1991; Menneer, et al., 2004, 2007; Stroud, et al., in preparation), whilst the prevalence effect arises as a result of observers essentially being too willing to respond 'absent' and not conduct a sufficient examination of a search array (Wolfe, et al., 2005; Wolfe, et al., 2007). Thus, it is possible to motivate observers to delay an 'absent' response to alleviate the prevalence effect, but it is not possible for observers to change the structure of their own visual system: not with extensive practice, and not with motivation.

5.4.4 General Discussion and Relevance to Airport X-ray Security Screening

To summarise, the present study employed a set of airport X-ray security screeners in the first full experiment that has been conducted to examine whether or not airport screeners are impacted by the dual-target cost and prevalence effect. The results were clear: screeners showed the presence of the dual-target cost, as well as a criterion shift in low prevalence. Thus, it seems that actual screener performance could be improved by a division of labour, such that several screeners search each X-ray display for the presence of just one target class (e.g. IEDs or metals, but not both). Additionally, it appears that TIP prevalence could be increased in order to cause a reduction in miss error rates. As a result of the apparently imbalanced effects of prevalence, increasing prevalence between 5%

and 20% would likely reduce miss error rates, but not increase false alarm rates, and would thus be an essentially cost-free improvement to screener performance. There exist a number of remaining factors that may have bearing upon the performance of actual airport screening personnel. In the remaining chapters, several of these factors will be examined in more detail.

6

Visual Search in the Presence of Ambient Noise

New Directions in the Study of Airport Security Screening

6.1 Introduction

During the Literature Review, a comment was quoted from Harris (2002) stating that screeners operate in a 'performance-degrading environment'. The goal of the present, and subsequent, empirical chapter, is to assess the extent to which the screening environment degrades the performance of airport security screeners. Up to this point, no previous research has attempted to examine this issue, and so the work here constitutes a series of initial, tentative steps towards understanding the interplay between a complex set of factors. To begin with, the experiment presented here will explore the impact of ambient noise upon search performance using stimuli from airport security screening.

Consider the environment in which airport X-ray security screeners carry out their screening task. Search combs are typically very busy places, with a number of potential sources of environmental distraction that could draw the attention of screeners away from the work that they are carrying out. Depending upon the structure and layout of the screening area, screeners may be able to: engage in conversation with passengers and colleagues; listen to conversations between other nearby passengers and colleagues; watch events and social interactions transpire around them (both near and far); see and hear baggage trolleys carrying their cargo around the airport (which typically emit a loud warning tone to notify those nearby of their presence); hear warnings and announcements being presented through the public announcement system, and so on. Furthermore, all of these potential sources of distraction vary in length, severity, frequency, and may overlap and interact with one another in an endless and infinite variety of combinations, making it very difficult to produce a single, simple description of the environment in which screeners operate.

6.1.1 The Impact of Auditory Distractions and Ambient Noise upon a given Task

Although common sense might lead to the assumption that the environment in which screeners operate will always cause a reduction in their search performance, a careful consideration of the factors involved is needed.
First of all, it needs to be noted that, despite a vast number of studies that have examined the impact of auditory distraction upon a task, there is little agreement with regards to *why* and *how* auditory distraction can sometimes impair performance. The influence of auditory distraction upon a primary task has been notoriously difficult to understand and define, with a plethora of conflicting evidence obtained across several decades (Taylor, Melloy, Dharwada, Gramopadhye, & Toler, 2004). Some experiments report *impairments* of performance in the presence of auditory distraction, some report *benefits* on performance in the presence of auditory distraction, and others find no impact of auditory distraction (Carter & Beh, 1987; Hygge & Knez, 2001; Jerison, 1957).
An extensive review was carried out by Koelega and Brinkman (1986), in which the impact of auditory distraction was examined upon *vigilance* tasks. Vigilance tasks require consistent monitoring for a very rare event over extended periods of time (Davies & Parasuraman, 1982) After reviewing the evidence, Koelega and Brinkman (1986) argued that very little can be concluded from the studies regarding the impact of distraction upon vigilance. A more recent review has, unfortunately, echoed similar concerns when discussing the impact of auditory distraction upon visual search tasks (Taylor, et al., 2004).

Using more tightly-specified tasks, in which participants are required to memorise a set of numbers, digits, or images for subsequent recall, it has been reported that presenting *steady-state* auditory streams (i.e. a continuous single sound or spoken word) has no impact upon memory performance, whilst *changing-state* auditory streams (in which the sound changes often radically in pitch and tone, or when different spoken words are presented successively) are detrimental to recall performance (Campbell, Beaman, & Berry, 2002; Jones, Macken, & Murray, 1993; Jones, Madden, & Miles, 1992). The examination of changing-state auditory streams upon performance has been examined in a number of recent studies, including using office noise (Banbury & Berry, 1997, 1998, 2005). For the most part, however, studies of the impact of changing-state auditory streams have used an auditory stream in conjunction with a recall task, rather than a visual search task, making it somewhat difficult to ascertain whether or not changing-state streams will have an impact upon visual search performance. However, as some have suggested that the changing-state effect is caused essentially by attentional capture, with attention being diverted to the auditory stimulus (Cowan, 1995), it may be the case that ambient noise, which is inherently changing-state in nature, will capture attention and impair search performance, or, at the very least, slow the search process.

6.1.2 Research Questions: Visual Search and Ambient Noise

As a first step towards examining the impact of environmental factors upon the performance of security screening personnel, the experiment presented in the current chapter explores the impact of ambient noise upon the detection of threat items from airport security screening. Harris (2002) has argued that the screening environment is detrimental to the performance of screeners, and, furthermore, considering the results regarding distractions in office environments (Banbury & Berry, 1997, 1998, 2005) it is important that such an examination be made, especially given the constant presence of ambient noise during real X-ray screening.

6.2 Method

6.2.1 Participants

Twelve participants (three males and nine females) took part in the study, with ages ranging from 19 to 53 (mean = 22.4, SD = 8.4). All participants were undergraduates and postgraduates, and reported normal colour vision and no previous experience with the stimuli. Participants received either course credit or payment for their participation. All participants completed the study within 30 days.

6.2.2 Apparatus and Stimuli

Apparatus and stimuli were the same as in Chapter 2.

6.2.3 Design and Procedure

Participants took part in three sessions, each lasting around 45 minutes. Before the trials began, a lengthy explanation was given concerning the nature of the targets, during which participants were guided through twenty examples of each type of threat item that they were to search for. Each session was blocked into three different sets of trials: single-target search for metals, single-target search for IEDs, and dual-target search for metals and IEDs. Each block began with five practice trials, followed by 160 experimental trials (giving rise to 480 trials overall per session). Participants were able to take a break every 50 trials. In terms of the visual search task, all sessions were identical, with the exception of the training given in the first session. The order of the blocks was counter-balanced across participants. A target was presented on 50% of all trials, and only one target could appear on any trial.

In each session, participants wore a set of headphones connected to a CD-player containing an ambient noise recording. The CD recording playback was initiated as the participants began the experiment. The ambient noise was recorded in a busy café during a lunch period, and lasted around one hour fifteen minutes, giving sufficient time for the participants to complete each session. Although it would have been desirable to use noise recorded in a real airport, there were security and time constraints in conducting this experiment, and it was not possible or practical to record and process the noise from an actual airport.

There were three noise conditions in total, one for each session: *Silent, Consistent* and *Inconsistent*. No ambient noise was played during the Silent session; however, participants still wore headphones. In order to examine whether or not changing-state aspects of the ambient noise captured attention and subsequently impaired search performance (Cowan, 1995), two ambient noise sessions were utilised. The ambient noise itself contained a great deal of changing-state material: as it was recorded in a café during a lunch period, the recording included a body of background noise, as well as clearly audible conversations that were taking place nearby, and, additionally, 'incidental' events, such as plates and cups being dropped, mobile telephones ringing, and so on. Given the variation inherent within the ambient noise, the noise itself was repeated in two noise sessions. The noise was analysed using *Wavelab 6* software, and normalised throughout at the original recording's mean volume (11.62dB). This was used in the *Consistent* noise session.

The *Inconsistent* noise session was employed to further be certain whether or not changing-state auditory material impacted search performance. For the Inconsistent session, the noise volume was varied in a pseudo-random fashion by dividing the recording up into segments of different volume. *Wavelab 6* was once again employed to control the volume of the ambient noise, and was thus modified to be either *High, Medium,* or *Low* in volume. The High volume was set to 200% of the Consistent session's volume (i.e. 23.24dB); the Medium volume was set to the same as the Consistent session's volume; and, finally, the Low volume was set to 50% of the Consistent session's volume (i.e. 5.81dB). To prevent the changes in volume from being temporally predictable, three different durations for the segments were utilised: 4, 6, or 8 seconds. Thus, there were 3 × 3 = 9 possible variations in the noise. After each noise segment, there was a 2-second transition period in which the volume was set to the half-way point between the concurrent segments' volume levels. A random-number sequence was generated (using the numbers 1-9), and each variation of noise type was assigned a number. Each random-number sequence lasted 72 seconds in total. The random-number sequences were generated successively from the beginning until the end of the recording. The same sequence was used for all participants.

Each trial began with the appearance of a small fixation cross at the centre of the display, followed by the presentation of the search field. There were two possible responses from the participants in any trial: "present" or "absent". The search field remained visible until the participant made a response, which ended the current trial and began the next. When a participant gave an incorrect response, an audible tone was produced by the computer.

6.3 Results

As with the previous empirical chapters, all *t*-tests have had their *p* values Bonferroni-corrected before being reported; additionally, the Greenhouse-Geisser *F* values, degrees of freedom, and *p* values are reported for repeated-measures ANOVA results wherever tests of sphericity are violated (i.e. Mauchly's test of sphericity shows a *p* value of less than .05). In all figures, error bars represent ±S.E.M.

In order to examine the impact of ambient noise upon search performance, a mixed-design ANOVA was required. In the present experiment, the different noise condition sessions (Silent / Consistent / Inconsistent) were counterbalanced across the participants (with three participants in each condition). This gave rise to a total of six different orders for the ambient noise presentations. Labelling Silent as 'S', Consistent as 'C' and Inconsistent as 'I', this gives: SCI, CSI, CIS, ICS, ISC, and SIC. Embedded within the different groups, there may be a learning effect between the consecutive sessions. Indeed, in the previous experiments reported in the present thesis, error rates often dropped between the early session(s) as participants learned the nature of both the targets and the distractors, which were highly unfamiliar to them.

Thus, when conducting the ANOVAs described below, search performance was examined using Session as a repeated-measures factor, whilst ordering of ambient noise was used as a between-subjects factor. This is useful for a number of reasons. It enables an examination of learning effects between the sessions, and additionally enables a detailed test of whether or not ambient noise causes a decrement in search performance. In other words, practice effects can be disentangled from bona fide effects of noise condition. If, in one session, one group shows increased error rates or RTs, then it can be argued that the noise presented in that session was likely to have impacted search performance.

6.3.1 The Impact of Ambient Noise upon Detection Performance

Examinations of the error rates detected the presence of the dual-target cost, yet there was no effect of ambient noise upon search performance. Error rates were examined using a 6 (Ambient Noise Order: SCI, CSI, CIS, ICS, ISC, SIC) × 2 (Search Type: Single-target Metals, Single-target IEDs, Dual-target Search) × 3 (Trial Type: Present, Absent) × 4 (Setsize: 4, 8, 12, 16) × 3 (Session: 1, 2, 3) ANOVA. Ambient Noise Order was entered as a between-subjects factor. The Ambient Noise Order factor failed to reach significance as either a main effect, or as an interaction with the other factors (all Fs<2.2). However, there were effects of Session ($F(2,24)=29.6$, $p<.001$), of Trial Type ($F(1,12)=43.8$, $p<.001$), of Search Type ($F(1.4,16.7)=8.6$, $p<.01$), of Setsize ($F(1.8,21.4)=13.9$, $p<.001$), as well as interactions between Session and Search Type ($F(2.5,30)=6.9$, $p<.01$), between

Trial Type and Setsize ($F(3,15)=6$, $p<.01$), and finally between Trial Type, Search Type and Setsize ($F(6,72)=3.7$, $p<.01$).

The Session × Search Type interaction (see Figure 6.3.1a, below) was caused by the fact that error rates for single-target metals did not vary across the three sessions. whilst error rates for both IEDs and dual-target search reached a plateau in the second session. Additionally, single-target metals exhibited reduced error rates compared to dual-target search in the first and second sessions, but not in the final session. A set of *t*-tests examined the Session × Search Type interaction in detail: the results of these *t*-tests are presented in detail in Table 6.3.1b.

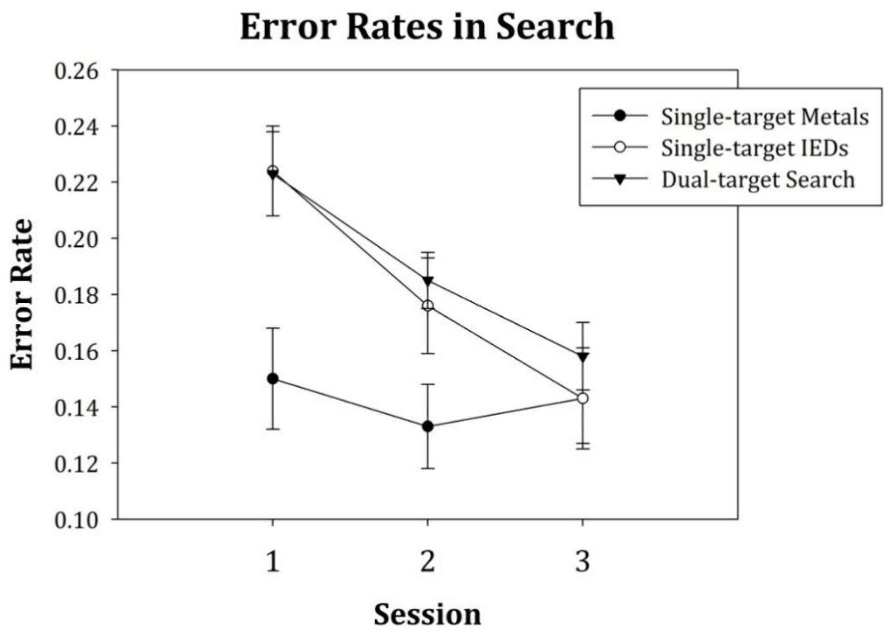


Figure 6.3.1a: Error rates in the Different Search Types, displayed as a function of Session number.

Table 6.3.1b
Results of t*-tests examining the Search Type × Session Interaction*

Comparison		*t*	*df*	*p*
S1 Metals	S2 Metals	1.47	17	ns
S2 Metals	S3 Metals	1.01	17	ns
S1 IEDs	S2 IEDs	3.86	17	<.05
S2 IEDs	S3 IEDs	2.60	17	ns
S1 Dual-target	S2 Dual-target	3.22	17	0.07
S2 Dual-target	S3 Dual-target	2.54	17	ns
S1 IEDs	S1 Dual-target	0.04	17	ns
S1 Metals	S1 Dual-target	4.98	17	<.01
S2 Metals	S2 IEDs	2.22	17	ns
S2 IEDs	S2 Dual-target	0.61	17	ns
S2 Metals	S2 Dual-target	3.91	17	<.05
S3 Metals	S3 IEDs	0.01	17	ns
S3 IEDs	S3 Dual-target	1.55	17	ns
S3 Dual-target	S3 Dual-target	1.14	17	ns

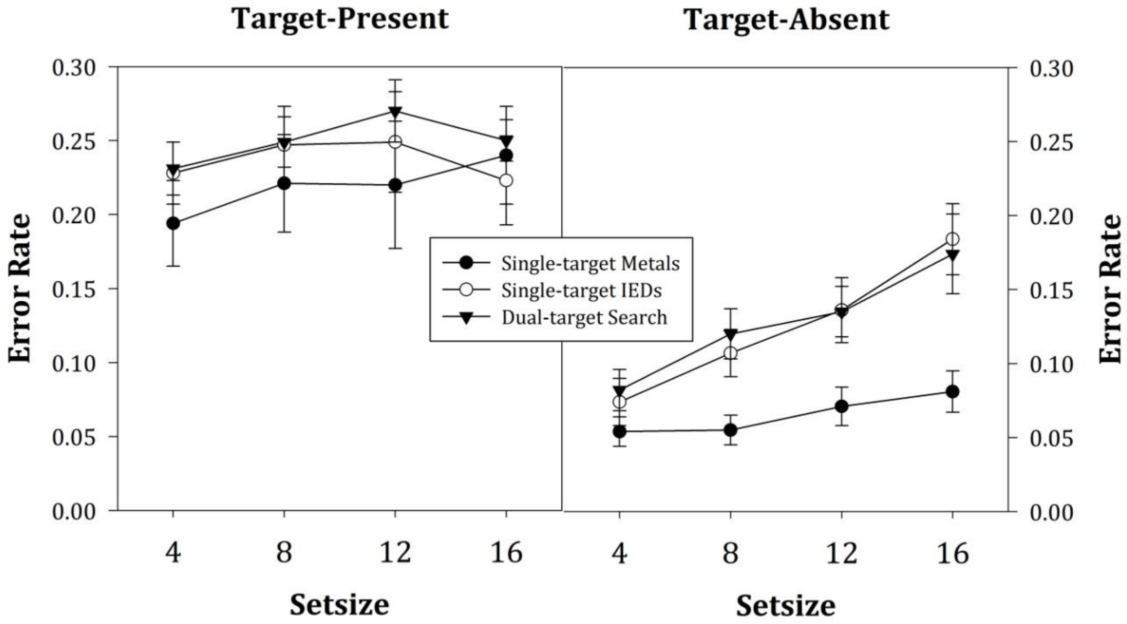


Figure 6.3.1c: Error Rates for Visual Search as function of Trial Type and Setsize.

To examine the three-way Trial Type × Search Type × Setsize interaction (displayed above in Figure 5.3.1c), Target-present and target-absent error rates were examined separately. Using additional ANOVAs, it was found that, for target-present error rates, there was no impact of Setsize, no impact of Search Type, and no interactions between Setsize and Search Type (all Fs < 2.7, all p>.05). However, for target-absent trials, there was a significant main effect of Search Type ($F(2,34)=13.2$, $p<.001$), a main effect of Setsize ($F(3,51)=18.6$, $p<.001$), as well as an interaction between Setsize and Search Type ($F(6,102)=3.9$, $p<.01$).

Further ANOVAs were used to explore the Setsize × Search Type interaction for target-absent trials in detail. For Setsize = 4, there was no difference between the Search Types ($F<1.6$). However, for the remaining Setsizes, there was a main effect of Search Type (Setsize=8: $F(2,34)=7.7$, $p<.01$; Setsize=12: $F(2,34)=10.2$, $p<.001$; Setsize =16: $F(2,34)=16.1$, $p<.001$). Detailed t-test results are presented in Table 5.3.1d. Overall, it was apparent that error rates for target-absent trials in single-target metal search were lower in setsizes of 8, 12 and 16 than the other two search types. Additionally, error rates for dual-target search in target absent trials were no different to those for single-target IEDs.

Table 6.3.1d

Results of t*-tests examining the Search Type × Setsize Interaction for Error Rates in Target-absent trials*

Comparison		Setsize	*t*	*df*	*p*
Single-target Metals	Single-target IEDs	8	3.4	17	<.05
Single-target Metals	Dual-target Search	8	3.2	17	<.05
Single-target IEDs	Dual-target Search	8	0.7	17	ns
Single-target Metals	Single-target IEDs	12	3.6	17	<.05
Single-target Metals	Dual-target Search	12	4.3	17	<.05
Single-target IEDs	Dual-target Search	12	0	17	ns
Single-target Metals	Single-target IEDs	16	4.9	17	<.05
Single-target Metals	Dual-target Search	16	5.2	17	<.05
Single-target IEDs	Dual-target Search	16	0.5	17	ns

6.3.2 The Dual-target Cost for Target-Present Trials

Error rates for each of the targets (metals, IEDs) in single-target search were compared with error rates for each of the targets in dual-target search, in order to determine whether or not the dual-target cost had a greater impact on one of the targets than the other. A 2 (Search Type: Single-target Search, Dual-target Search) × 3 (Target Type: Metals, IEDs) × 4 (Set size: 4, 8, 12, 16) × 3 (Session: 1,2,3) ANOVA was used. This revealed a main effect of Session (F(2,34)=6.67,p<.01), a main effect of Setsize (F(3,51)=2.99,p<.05), and an interaction between Session and Target Type (F(2,34)=10.65,p<.001), which was subsumed by a further interaction between Session, Target Type and Search Type ($F(1.5,25.8)=5.56, p<.05$).

Further comparisons conducted on metals and IEDs separately revealed the presence of a dual-target cost for metals ($F(2,34)=4.97, p<.05$) and for IEDs ($F(1,11)=5.48, p<.05$). Additionally, error rates for metals did not decrease as the sessions progressed ($F<2.8$). This was not the case for IEDs, which showed a main effect of Session ($F(2,34)=10.9, p<.001$), and *t*-tests examining the IED error rates revealed that error rates decreased as the sessions progressed, reaching a minimum in Session 2 (comparing Session 1 and 2: $t(35)=3.2, p<.05$; comparing Session 2 and 3: $t(35)=1.9, p>.05$).

6.3.3 The Impact of Ambient Noise upon Reaction Times

Examination of the RT data replicated RT data from previous experiments, detecting the presence of a dual-target cost for RTs; however, as might be expected from the preceding examinations of the error rate data, RTs were unaffected by the presence of ambient noise. In the only previous chapter where setsize was manipulated, Chapter 3, RTs were examined in terms of the actual time taken to respond. This approach was adopted in order to examine the role that target prevalence had upon RTs, as has been done in previous studies where target-absent RTs tend to become as fast as target-present RTs (see Wolfe, et al., 2007). In the present study, however, an index of overall search performance in the different conditions was needed, so, rather than examine RTs, a set of *slopes* were computed based upon a linear regression through the RTs at each setsize. This is a more traditional approach to examining search performance, dating back to the early days of visual search (Treisman & Gelade, 1980). The modern use of slopes is to assess search efficiency: when the slope is lower, the observer is able to search through a larger number of distractors more rapidly (Wolfe, 1998). When the slope is higher, then search is less efficient (see also the preceding Literature Review, Chapter 1).

Target-Present Trials: Target-present and target-absent slopes were examined separately. Target-present trial slopes were examined using a 6 (Ambient Noise Order: SCI, CSI, CIS, ICS, ISC, SIC) × 2 (Search Type: Single-target Search, Dual-target Search) × 3 (Target Type: Metals, IEDs) × 4 (Set size: 4, 8, 12, 16) ANOVA. Ambient Noise Order was entered as a between-subjects factor. These revealed main effects of Search Type ($F(1,12)=20.4, p<.01$) and of Session ($F(1.2,14)=16.7, p<.01$), and an interaction between the two ($F(1.3,15)=6.4, p<.01$). No other effects or interactions, including those involving the Ambient Noise Order factor, reached significance (all Fs<2.9, all ps>.05).

The Search Type × Session interaction was examined using a series of further ANOVAs conducted upon each of the sessions separately, and is depicted below in Figure 6.3.3a. These ANOVAs were of identical design to the first ANOVA, except that the Session factor was removed. The ANOVAs revealed that there was a dual-target cost in terms of RT slopes for Session 1 ($F(1,12)=11.2, p<.001$), and for Session 2 ($F(1,12)=28.5, p<.001$), and a strong trend for a dual-target cost in Session 3 ($F(1,12)=4.3, p=.062$). Note that in the context of RT slopes, the dual-target cost takes effect when the slope is higher for dual-target search than single-target search (Menneer, et al., 2007).

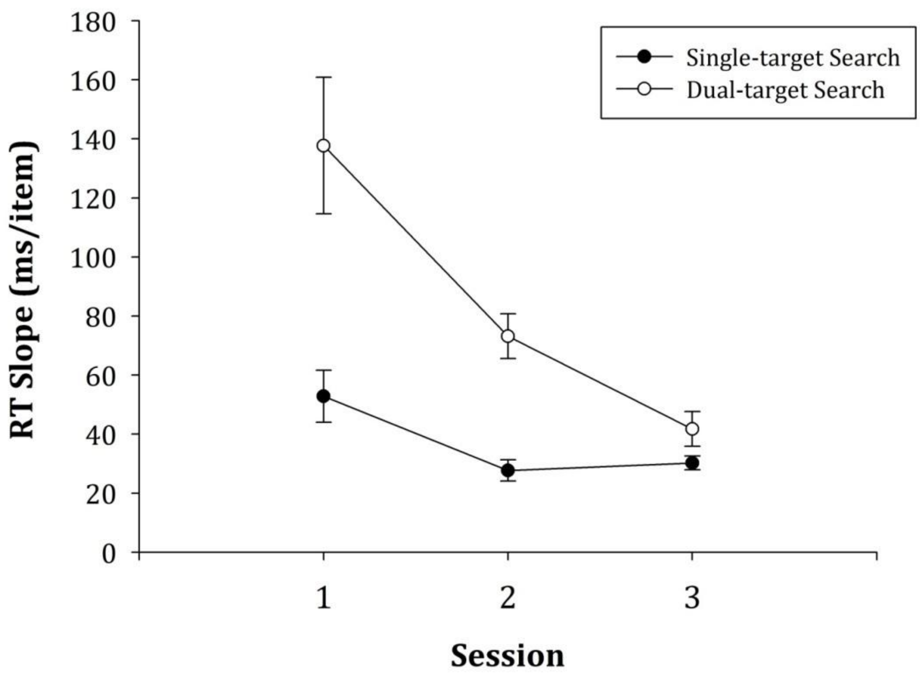


Figure 6.3.3a. RT Slopes (in ms/item) for Single-target and Dual-target Search as a function of Session.

Target-Absent Trials: Target-absent trial RT slopes were examined using a 3 (Session: 1,2,3) × 3 (Search Type: Single-target Metals, Single-target IEDs, Dual-target Search) × 3 (Ambient Noise Order: SCI, CSI, CIS, ICS, ISC, SIC) ANOVA. This revealed main effects of Search Type ($F(2,24)=20.9, p<.001$), and of Session

(F(2,24)=19.3,p<.001), as well as an interaction between the two (F(4,48)=3.6, p<.05).

In order to explore the Search Type × Session interaction (see Figure 6.3.3b, below) in the target-absent RT slopes, the first and final sessions were examined separately using further ANOVAs of identical design, except with the Session factor removed. Only the first and final sessions were examined, in order to prevent conducting an excessively large number of comparisons. There was a main effect of Search Type for both Session 1 (F(2,24)=19.3,p<.001), and Session 3 (F(2,24)=3.6, p<.05). A series of t-tests revealed that there were no differences between slopes for single-target metals and single-target IEDs in Session 1 and Session 3 (ts<2.1,ps>.05). However, slopes were higher for dual-target search than for single-target metals in Session 1 (t(17)=2.8,p<.05), and Session 3 (t(17)=5.6,p<.05). Similarly, slopes were higher for dual-target search than for single-target IEDs in Session 1 (t(17)=2.9,p<.05), and Session 3 (t(17)=5.4,p<.05). Thus, there was a dual-target cost in terms of search efficiency which was not eliminated after practice by the third and final session.

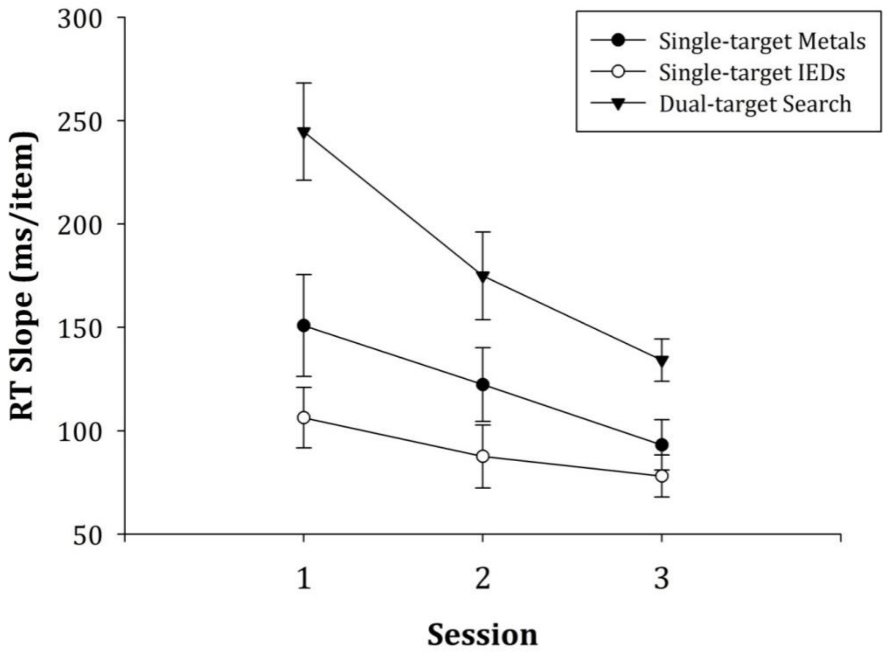


Figure 6.3.3b. RT Slopes (in ms/item) for the different levels of Search Type factor as a function of Session.

6.4 Discussion

The present study explored the impact of ambient noise upon visual search performance for X-ray threat images. Overall, the results from previous research were replicated: dual-target search exhibited decreased search efficiency and decreased detection rates.

However, it appeared that the ambient noise presented to participants had no impact upon their search performance. With relation to airport security screening, the present study actually offers some 'good news': one factor that was assumed to impair screener performance (Harris, 2002) has been shown to actually not affect performance at all. However, there are some caveats and other factors that need to be considered before it can be declared that the screening environment is not detrimental to the search performance of screening personnel. Ambient noise is just one form of potential distraction, and, a somewhat related, but different form of distraction is examined in the next empirical chapter.

7

Search Termination Thresholds and Environmental Distraction

Moving Closer to the Screening Environment

7.1 Introduction

In the previous chapter, it was reported that ambient noise did not impair search performance for X-ray images of threat items. It was noted in the discussion that, although ambient noise apparently does not impair search performance, there may be other forms of potential distractions that *do* impair search performance. The present chapter examines a previously-unexplored route to impairments of search performance from environmental distractions: namely, the involvement of search termination thresholds and how the termination thresholds are impacted by distractions.

The impact of distractions have typically been examined using dual-task methodologies, with participants being required to carry out two tasks simultaneously. The logic of the dual-task method is that if the same cognitive resources are required to perform both of the tasks, then performance should be impaired when the tasks are carried out concurrently. Indeed, such a method has been used in a number of studies, pairing visual search with a secondary task of one form or another (for example, Han & Kim, 2004 paired visual search with a mental arithmetic task where participants had to continually count down from a given number whilst searching).

7.1.1. Distractions and Search Termination Thresholds

The experiment reported in the present chapter adopts a different approach, with the goal here being to examine the extent that secondary tasks affect the termination thresholds adopted by observers engaged in visual search (Chun& Wolfe, 1996). In the previous chapters, the search termination thresholds were examined in terms of the prevalence effect, with low-prevalence search resulting in target-absent trials becoming as rapid as target-present trials: in typical visual search tasks, target-absent trials have a longer duration than target present trials (Wolfe et al., 2007).

However, previous studies of dual-task methodologies have failed to examine whether or not engaging participants in a secondary task actually has any impact upon the search terminations thresholds. Put simply, the key question being asked in the present study is this: if participants are given a secondary task to carry out alongside a primary visual search task, do the participants give themselves longer to complete the visual search task before giving a response? In other words, are the mechanisms that control the search termination thresholds sensitive to events during the search process, and are those mechanisms able to adjust the termination threshold accordingly to compensate for the time lost due to the secondary task? These are important questions to ask, because the environment in which screeners operate are potentially fraught with distractions.

Here, participants were engaged in dual- and single-target search, and in both a visual search task alone, as well as a combined-task requiring visual search in conjunction with a secondary mental arithmetic task. The secondary mental arithmetic task is, in some senses, comparable to a form of external distraction that airport screeners may have to face, such as having a conversation with another individual whilst screening. Rather than, for example, engage the participants involved in the present study in conversation, a mental arithmetic task was used in order to ensure that the secondary task was being carried out efficiently.

If it was the case that the participants here were unable to complete the visual search and arithmetic task concurrently, then one of the tasks would show very poor performance (for example, if they opted to conduct the search task and could not conduct the mental arithmetic task at the same time, then performance in the mental arithmetic task would be zero). Indeed, it should be noted that previous studies involving the impact of using a mobile telephone whilst driving (and searching for specific visual targets) have found that participants involved in a mental arithmetic task show comparable reductions in performance with participants who are involved in a conversation with another individual (Recarte & Nunes, 2003). Thus, it seems that a mental arithmetic task is a useful tool for examining search performance in an environment replete with distractions.

7.1.2 Notes on Methodological Issues

Piloting of the experiment revealed that participants were unable to perform the combined-task to any degree of competence without having first completed the two single tasks separately. In some senses, this was not really surprising: the impact of secondary tasks upon the learning of a difficult primary task has long been examined within the multi-tasking literature, and it is often the case that two complex tasks are best learned separately before being combined (Detweiler & Schneider, 1991). As a result, it was ensured that participants in the present study always performed the combined-task after having had practice upon both the visual search task, and the secondary task.

A design of this nature naturally has some implications for the predictions. It seems likely that participants will improve their performance in the visual search task between the initial practice phase, and the combined-task. Therefore, the null hypothesis is simply that, response accuracy will be improved in the combined-task, when compared to the practice phase. Similarly, search reaction times should decrease in the combined-task, compared to when search was initially conducted alone. Previous examinations of the impact of practice upon these stimuli have been presented by Menneer et al. (under review).

7.2 Method

7.2.1 Participants

Eighteen participants (three males and fifteen females) took part in this study, with an age range of 18 to 48 (mean=21.2, SD=6.8 years). Participants were undergraduates and postgraduates, and reported normal colour vision, as well as having no previous experience with X-ray stimuli. They took part either for course credit or for payment. Participation was completed in less than 30 days for all participants.

7.2.2 Apparatus and Stimuli

Apparatus and Stimuli were the same as in Chapter 4.

7.2.3 Design and Procedure

Each participant took part in three sessions, lasting around one hour each. One session involved the mental arithmetic task, the second involved the visual search task, and the third session involved the combined mental arithmetic and search task (hereafter: the 'combined-task'). Piloting of the experiment revealed that participants were unable to perform the combined-task to any degree of competence in their first session: thus, due to its inherent difficulty, the combined-task was always the final session for each participant. The first two sessions were counterbalanced, however, such that half of the participants were engaged first of all in a visual search session, followed by a mental arithmetic session, whilst the other half of the participants took part in a mental arithmetic session first of all, followed by a visual search session. Whenever an incorrect response was given for either task, an audible tone was produced by the computer.

Visual search session. The visual search session was blocked into three different sets of trials, consisting of single-target search for metals, single-target search for IEDs and dual-target search for both metals and IEDs (the order of these blocks was counterbalanced across the participants). Each block was preceded by five practice trials, after which there were 160 experimental trials. Participants were given the opportunity to take a break every 50 trials. Before this session began, participants were given a detailed explanation regarding the nature of the targets, along with 20 example images of each threat type. A target appeared on 50% of trials, and only one target could appear on any trial.

The visual search trials began with a small set of asterisks written in the same font, font size and colour as the mathematical questions situated at the centre of the display. This was followed after 1s by the presentation of the search field, during which participants had an unlimited time period to respond "present" or "absent". After their response had been given, the display was cleared, and replaced with the instruction "Please Press Enter" (this was used to bring response complexity in the visual search task closer to that of the combined-task). Once participants had pressed the "Enter" key on the keyboard, the next trial began. The asterisks, "Please Press Enter" instruction, and mathematical questions were all presented in a Font Size of 12 using the Arial font, presented in the centre of the display.

Mental arithmetic session. The mental arithmetic session was also blocked into three sets of trials, with 160 trials in each and five practice trials beforehand. Each trial began with the presentation of a mathematical question. This could involve the addition or subtraction of two two-digit numbers. The first of the numbers could be between 40 and 70, whilst the second could be between 10 and 30. Both the calculation required (addition or subtraction), and the numbers involved were randomly generated. The number ranges were chosen specifically such that the mental arithmetic required would be of easy to medium difficulty, and the answer would never be a negative number.

The mathematical question was visible for one second, after which the display was cleared and a neutral display (consisting of a series of asterisks in the centre of the display) was presented for 1s. Participants were then presented with the instruction “Please Type Your Answer Followed By Enter” in the centre of the screen. As they typed, the number was visible upon the screen in front of them. They could also use the backspace key to correct any mistakes, and were made aware of this fact. After pressing the “Enter” key, the trial ended and the next one began.

Combined session. The combined session involved the interleaving of the visual search and mental arithmetic tasks. Each trial began with the presentation of the mathematical question, followed in turn by the visual search task. Once participants had responded “present” or “absent” to the visual search task, the search display was cleared, and participants were presented with the “Please Type Your Answer Followed By Enter” instruction from the mathematical task, at which point they simply had to type their answer to the mental arithmetic question. Once they had done so, the trial ended and the next began. When an incorrect response was given to either components of the task, an audible tone was produced by the computer.

7.3 Results

In the following results, all *t*-tests have had their *p* values Bonferroni-corrected before being reported; additionally, the Greenhouse-Geisser *F* values, degrees of freedom, and *p* values are reported for repeated-measures ANOVA results wherever tests of sphericity are violated (i.e. Mauchly's test of sphericity shows a *p* value of less than .05). In all figures, error bars represent ±S.E.M. An initial examination of the data flagged one participant as an outlier, and they were thus removed from the analysis (their target detection levels in the visual search task were below chance). Due to a program error, data from one block of mathematical trials for one participant were lost, so the median response accuracy and response time for the same block of mathematical trials was calculated for this participant, based upon the results obtained from the other participants.

7.3.1 Overall Performance in the Mathematical Task

Examining performance in the mathematical task revealed that response accuracy for the mathematical task was no different between the mathematical task and combined-task, and that RTs were faster in the combined-task than in the mathematical task.

The three blocks of trials in the mathematical task were labelled as *Block M*, *Block I* and *Block D*. In the combined-task, Block M involved the search for Metals only in conjunction with the mathematical task, Block I the search for IEDs and Block D involved Dual-target search alongside the mathematical task. In order to mitigate against order effects, the blocks of trials for the session involving just the mathematical task were also labelled as Blocks M, I, and D. The labels were based upon the counterbalancing for each participant: so, for example, if that participant, in their visual search session, and in the visual search component of the combined-task, was asked to search for metals in their first block of trials, then the first block of mathematical trials was labelled Block M.

Analysis of the results from the mathematical task proceeded using two 3 (Block Type: M, I, D) × 2 (Task Type: mathematical task, combined-task) repeated-measures ANOVAs, one for response accuracy, and one for response time. The ANOVA conducted upon the response accuracy data revealed no significant effects or interactions (all Fs<2), see Figure 7.3.1a, below (left panel). However, the ANOVA conducted upon the response time data indicated that response times were generally faster in the combined-task ($F(1,16)=22.4$, $p<0.05$): there were no other significant effects or interactions (all Fs<2). Mean time to respond in the mathematical task was 2813ms (SD=316ms), and in the combined task was

1388ms (SD=136ms). The effect of the task upon RTs for the mathematical task are displayed graphically in Figure 7.3.1a, below (right panel).

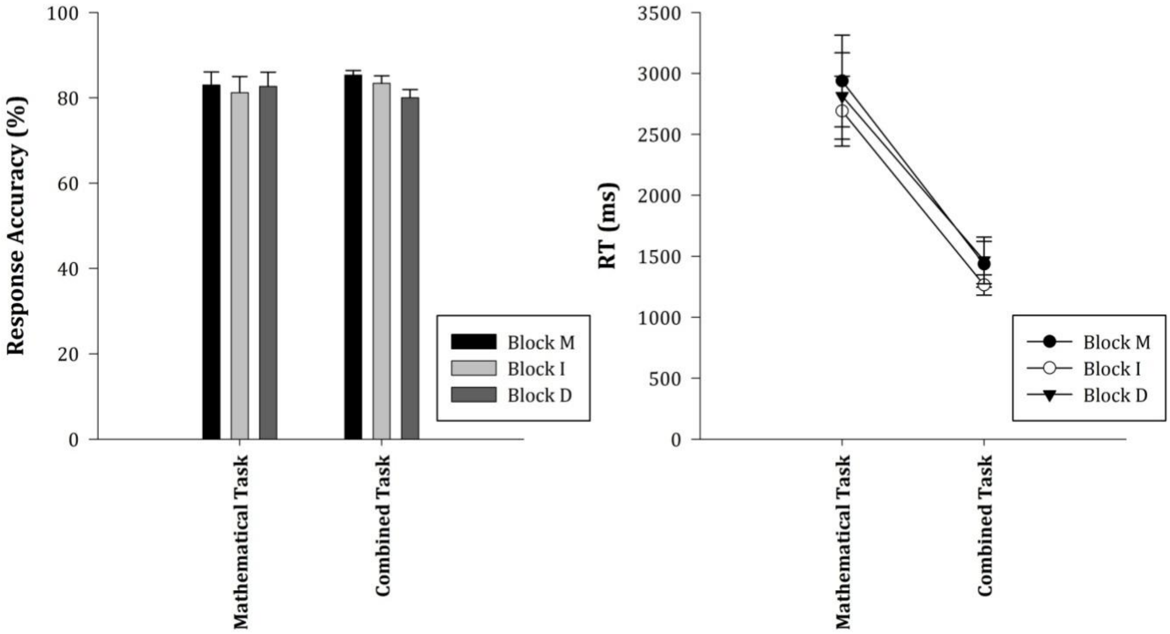


Figure 7.3.1a. Left panel: Response Accuracy data for the mental arithmetic task when conducting the task alone ('Mathematical Task') or in conjunction with the visual search task ('Combined Task'); Right Panel: Reaction Times for the mathematical task for the different Block Types when the mathematical task ('Mathematical Task') was conducted alone, or when paired with a visual search task ('Combined Task').

The reduction in response times for the combined-task is not surprising. In the combined-task, participants were presented with the mathematical question, required to carry out a visual search, and then asked to provide an answer to the mathematical question. Thus, when they were asked to provide their mathematical answer, they had been given ample time to calculate that answer. The reduction in response times here for the mathematical trial therefore suggests that participants were calculating the answer to the mental arithmetic task after being presented with the mental arithmetic question, then proceeding to carry out their visual search.

7.3.2 Performance in the Visual Search Task

Examining performance in the visual search task revealed some evidence to suggest that search performance was impaired by conducting a mental arithmetic task alongside a visual search task.

Participants, on each trial, could produce a correct or incorrect response to both the search component of the task, and the mathematical component of the task. As a result, there are four different outcomes in each combined-task trial. To begin with, let us denote the four outcomes as follows, using two letters (overall, the classes of outcomes are coded as in Table 7.3.2a as: *XX*, *SX*, *XM* and *SM*). The first letter denotes performance on the search component of a given trial; the second letter denotes performance on the mathematical component of a given trial. If a participant incorrectly answers one component, the component which was answered incorrectly is labelled as 'X'. Thus 'XX' denotes that a participant responded incorrectly to *both* the search *and* the mathematical task. If the participant correctly responds to the search component, but not to the maths component, then a 'SX' outcome is produced. Likewise, a correct answer to the maths, but not to the search gives an 'XM' outcome, whilst correctly answering both components of the combined-task produces an 'SM' category. These are illustrated below, in Table 7.3.2a:

Table 7.3.2a

Coding for the Different Outcome Categories in the Combined-Task Session

Search	Maths	Coding
Incorrect	Incorrect	XX
Correct	Incorrect	SX
Incorrect	Correct	XM
Correct	Correct	SM

The four possible outcome categories that were actually observed in the combined-task session can then be compared to a fifth category representing search performance in the single-task search session. As the single-task search session involves only the search task, this will be denoted as 'S', for the proportion of correct responses in search for a given target type or search type (e.g. single-target search or dual-target search).

The response accuracy results were therefore examined in a 3 (Search Type: Single-target Metals, Single-target IEDs, Dual-target Search) × 2 (Presence: Present, Absent) × 4 (Outcome Category: S, XX, SX, XM, SM) ANOVA. This revealed that search alone (S) was higher in response accuracy than the other outcome categories (see Figures 7.3.2b, 7.3.2c, and 7.3.2d, below). There were main effects of Search Type ($F(2,32)=9.52, p<.01$), of Outcome Category ($F(4,64)=320, p<.001$), and interactions between Presence and Outcome Category ($F(1.5,24.6)=4.5, p<.05$), between Search Type and Outcome Category ($F(3.3,52.7)=4.1, p<.01$), and a strong trend towards an interaction between Presence, Search Type and Outcome Category ($F(4.5,72)=2.3, p=.064$).

There is little sense in examining the interaction between Presence, Search Type and Outcome Category to its fullest extent: instead of examining and comparing all of the three Search Types and five Outcome Category factors in both target-present and target-absent trials in extensive detail, a more precise set of questions and analyses can be tested. However, including all of the five Outcome Category factors has been useful because the proportion of XX, XM, and SX trials indicates that participants were not giving up on the task completely as a result of it being too difficult (or else there would be a large number of XX trials), and that, generally speaking, participants were not trading off one task for the other (i.e. giving up on the search task and focusing on the mathematical questions, or vice versa), because of the low proportion of both SX and XM trials.

The remainder of the analyses here will focus on comparing S and SM trial proportions, to examine whether or not there is a real cost in search performance when successfully completing the secondary task. A simplified ANOVA, identical in design to that which was initially conducted upon all of the Outcome Category factors, was focused upon a comparison between the S and SM categories. This revealed that response accuracy was reduced in SM compared to S: there was a main effect of Outcome Category ($F(1,16)=24.9, p<.001$), which did not interact with any of the other factors ($Fs<1$). Overall response accuracy dropped from a mean of 80% correct responses (S.E.M.=0.014) in search alone to a mean of 69% correct responses (S.E.M.=0.028) when responding correctly to both the mathematical component of the combined-task and the search component of the combined-task.

Furthermore, a dual-target cost was detected, with a main effect of Search Type ($F(2,32)=10.1, p<.001$). The dual-target cost did not interact with Outcome Category ($F<1$, $p>.05$) or with Presence ($F<1$, $p>.05$). Subsequent *t*-tests revealed no differences between single-target metals (mean=0.77, S.E.M.=0.01) and single-target IEDs (mean=0.74, S.E.M.=0.02) in terms of response accuracy ($t(33)=1.61, p>.05$). However, response accuracy for dual-target search (mean=0.71, S.E.M.=0.02) was significantly lower than response accuracy for single-target metals ($t(33)=4.88, p<.01$), and single-target IEDs ($t(33)=2.79, p<.05$).

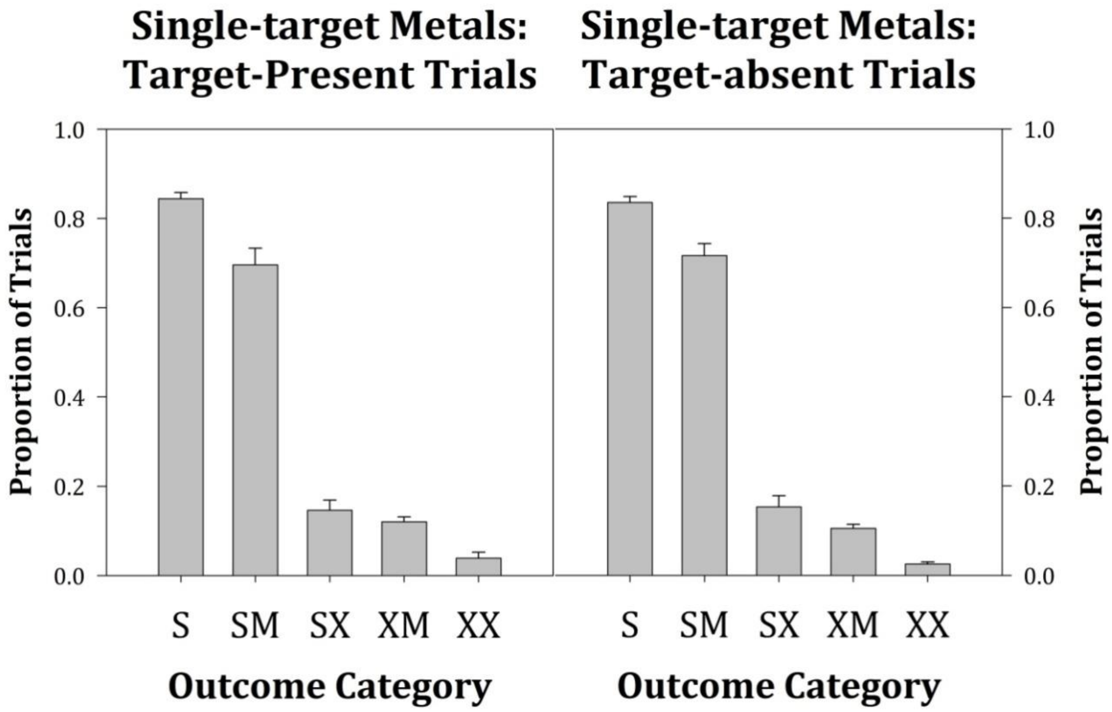


Figure 7.3.2b: Proportion of responses in the five Outcome Categories for single-target metal search.

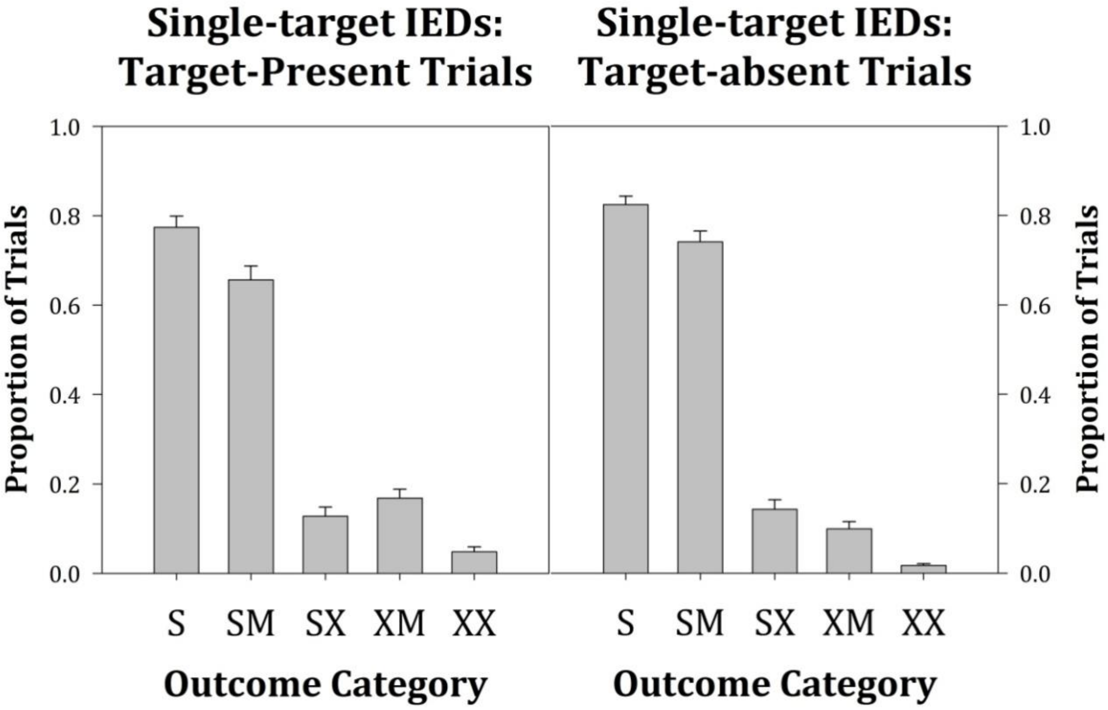


Figure 7.3.2c: Proportion of Responses in the five Outcome Categories for single-target IED search.

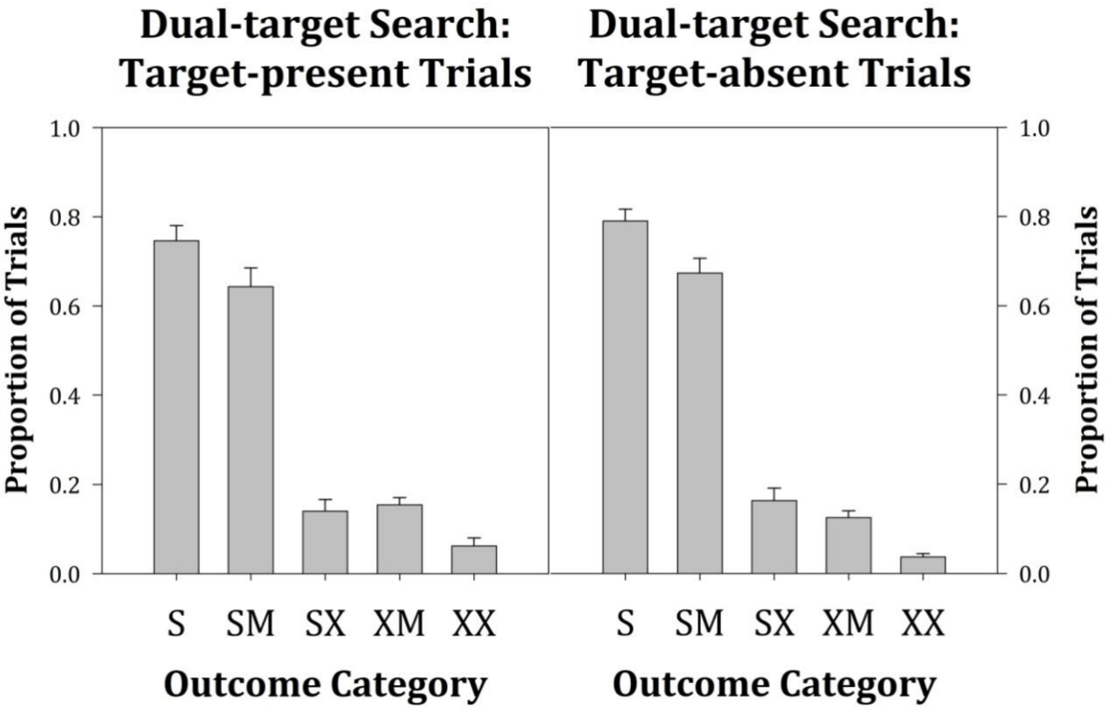


Figure 7.3.2d: Proportion of Responses in the five Outcome Categories for target-absent trials.

7.3.3 Impact of the Combined-task upon the Dual-target Cost

The dual-target cost was not amplified in the combined-task session. Response accuracy for S and SM trials were compared in a 2 (Search Type: Single-target, Dual-target) × 2 (Search Target: Metals, IEDs) × 2 (Outcome Category: S, SM) ANOVA. Only S and SM trials were selected because the focus here is upon response accuracy in search, and not on the proportion of other categories (i.e. SX, XM, and XX). There was a main effect of the dual-target cost ($F(1,16)=10.38, p<.01$), with dual-target search exhibiting lower response accuracy than single-target search (single-target search mean accuracy=0.74, S.E.M.=0.022; dual-target search mean accuracy=0.69, S.E.M.=0.033). The dual-target cost did not interact with any other factors (Fs<1.1). Additionally, there were main effects of Search Target ($F(1,16)=9.5, p<.01$) and Outcome Category ($F(1,16)=17.4, p<.01$), and an interaction between the two ($F(1,16)=6.7, p<.01$). This interaction is depicted graphically below in Figure 7.3.3a.

Examination of the Search Target × Outcome Category interaction revealed that response accuracy was lower in the combined-task than the search task alone. This was the case for metal targets ($t(33)=6.3, p<.01$), and IEDs ($t(33)=3.5, p<.01$). The interaction was caused by the fact that response accuracy for metals was higher in the S category than IEDs ($t(33)=4.4, p<.01$), but not in the SM category ($t(33)=1.8, p>.05$).

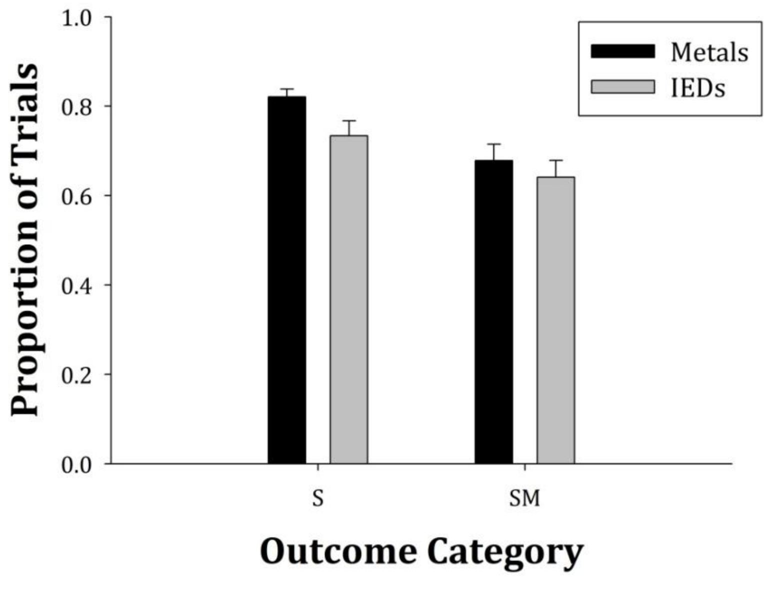


Figure 7.3.3a: Proportion of S and SM trials for Metals and IEDs.

7.3.4 Overall Search Response Accuracy in the Combined-task Session

The preceding analyses of the response accuracy data focused on the overall balance between the manner in which participants responded to the two tasks. Participants did not give up on the task entirely, and did not trade one task off for the other (shown by the low XX, SX and XM rates). However, the division of search trials into the SM and SX categories may be hiding another effect: whether or not *overall* search performance was impaired in the combined-task, when compared to the single-task session.

In order to examine for such a possibility, a somewhat different set of comparisons were made. Response accuracy in single-task search was compared with overall response accuracy in the combined-task, regardless of whether or not, on any given trial, the mathematical task was responded to correctly. The data were entered into a 2 (Task: Single-task, Combined-task) × 3 (Search Type: Single-target Metals, Single-target IEDs, Dual-target Search) × 2 (Presence: Target-present, Target-absent) ANOVA. This revealed no main effect of Task, and no interactions between Task and the remaining factors (all Fs<1.2). Thus, it is apparent that overall search performance was not impaired, when one does not consider whether or not participants were correctly answering the mathematical component of the combined-task.

7.3.5 Response Times in the Visual Search Task

As can be seen from Figures 7.3.2b, 7.3.2c and 7.3.2d, the frequency at which SX, XM, and XX trials occurred was very low indeed. For some participants, one or more of these cells were empty. Indeed, for most participants, the number of occurrences of each of these three categorical outcomes was less than 10. As a result, RT analyses were focused upon SM trials, in which participants correctly responded to both the search and mathematical components of the combined-task. These could then be compared directly to the RTs for search alone (S trials), as well as the RTs for maths alone.

Examination of the RT data revealed that combined-task search trials (SM trials) exhibited markedly higher RTs than search-alone trials (S trials): see Figure 7.3.5a. An initial ANOVA examined the impact of the combined-task upon search RTs across all of the search trial types. To simplify the ANOVA comparisons, RTs for single-target absent trials were mean averaged for each of the search sessions.

This was shown to be permissible through the use of a 2 (Task: Single, Combined) × 2 (Single-target Block: Metals, IEDs) ANOVA conducted upon the target-absent RTs, which indicated no main effect of Single-target Block (F<1), and no interaction between Single-target Block and Task (F<3.7,p>.05).

Thus, RTs were examined using a 2 (Task: Single, Combined) × 2 (Search Type: Single-target, Dual-target) × 3 (Trial Type: Metals, IEDs, Absent) ANOVA. This revealed main effects of Task (F(1,16)=12.4,p<.01), of Search Type (F(1,16)=4.8,p<.05), and of Trial Type (F(2,32)=24.3,p<.001). The main effects were embedded within a series of interactions, between Task and Search Type (F(1,16)=7.4,p<.05), and between Task, Search Type, and Trial Type (F(1.4,21.7)=6.9,p<.05). Additionally, the Search Type × Trial Type interaction narrowly missed significance (F(1.2, 18.6)=4,p=.054).

In order to examine the Task × Search Type × Trial Type interaction, each of the Trial Types were examined separately using a series of further ANOVAs. These ANOVAs revealed that the Combined-Task RTs were longer than for Search Alone, for Metals (F(1,16)=12.6,p<.01), IEDs (F(1,16)=9.9,p<.01), and Absent trials (F(1,16)=11.9,p<.01). There was a main effect of the dual-target cost for IEDs (F(1,16)=4.7,p<.05), yet not for Absent trials (F<1). Metals showed a strong trend towards a dual-target cost (F(1,16)=3.8,p=.07). For IEDs only, the dual-target cost interacted with Task (F(1,16)=10.6,p<.01). Subsequent t-tests showed that this was due to the presence of a dual-target cost for the search alone session (t(16)=3.3,p<.01), but not for the combined-task session (t(16)=0.9,p>.05).
The interaction between Task × Search Type × Trial Type, depicted in Figure 7.3.5a, below, is interesting because it revealed that there was no dual-target cost for RT in the combined-task condition. In some senses this is surprising, because there was a dual-target cost for response accuracy detected for metals, IEDs and target-absent trials. In some senses, this can be described as a ceiling effect in the search termination thresholds: all else being equal, there should have been a dual-target cost for RT in the combined-session. Thus, it appears that participants failed to increase the amount of time that they were willing to search for on each trial during the combined-task condition, and this may be the cause of the reduced response accuracy rates in the combined-task that were noted above.

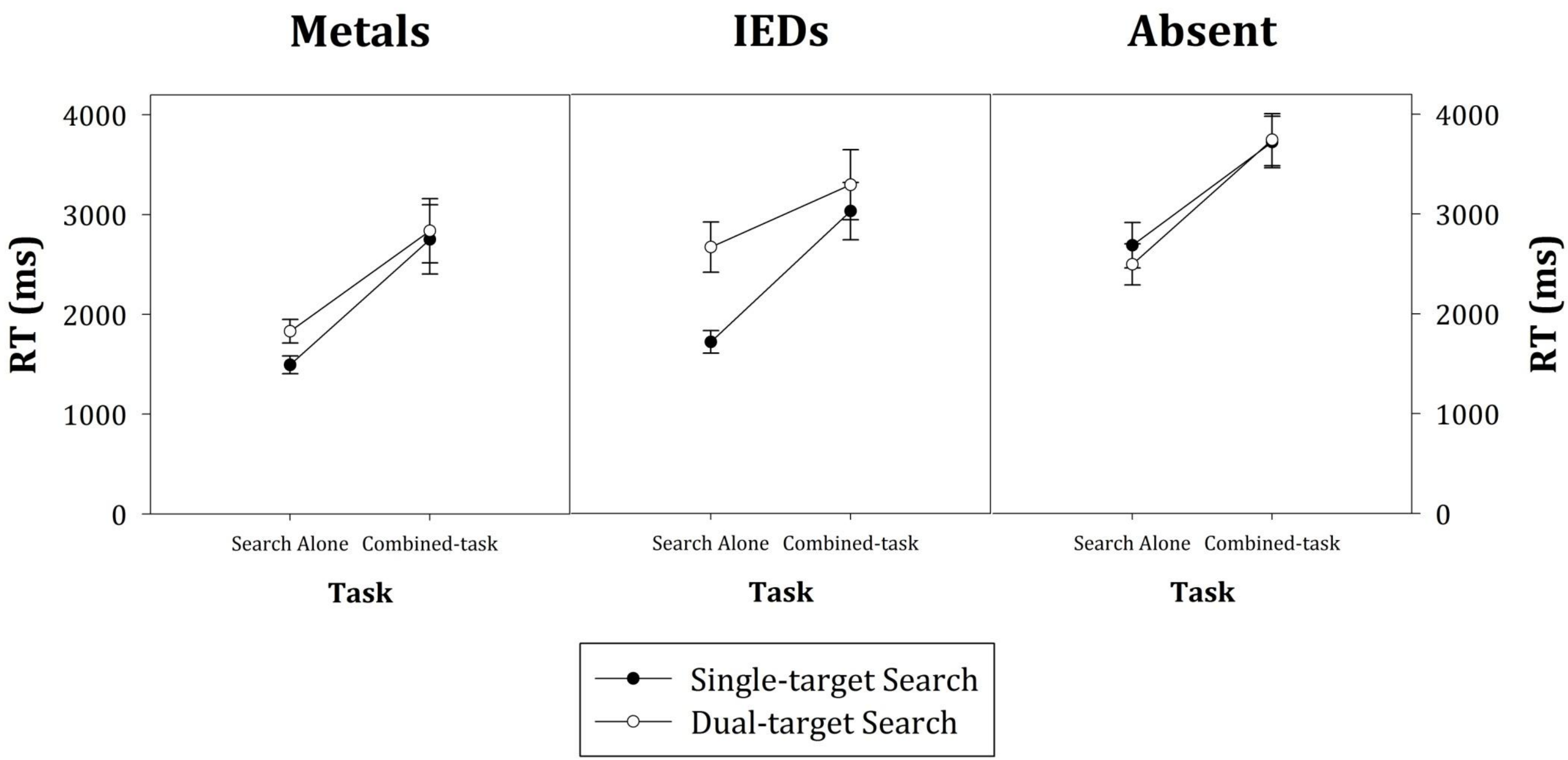


Figure 7.3.5a: Reaction Times (RTs) for the different Trial Types, Search Types, and Task Types

7.4 Discussion

The goal of the present study was to assess whether or not a secondary task results in an extension of search termination thresholds to accommodate the additional time required to conduct the secondary task. Overall, response times were heavily affected by the addition of a secondary task. Participants required more time to detect targets, and more time to respond 'absent' in the combined-task than when conducting visual search alone. It is likely that the impact upon the response accuracy was driven by participants not giving themselves sufficient time to detect targets in the combined-task session, as shown by the ceiling effects for the reaction times in the search component of the combined-task.

Previous experiments that have explored the impact of a secondary task upon a primary visual search task (e.g. Han & Kim, 2004) have typically explored the impact that a continuous secondary task has upon search performance, and thus have effectively ignored the role that distractions play upon the search termination threshold. In airport screening, the search termination threshold is of vital importance, as highlighted by the previous chapters: target-absent trials should typically exhibit longer reaction times than target-present trials. Previous studies of the prevalence effect have tied the effect to an acceleration of target-absent trials (Wolfe at al., 2007).

As noted in Chapter 5, airport screeners may be able to limit the impact of target prevalence by delaying their 'absent' responses. The present study indicates that this strategy may not be sufficient to prevent screeners from limiting the impact of external distractions. For example, consider a scenario where a supervisor asks a screener a simple question. After the screener replies to the question, they will return to their X-ray display and continue searching. The present study has revealed that, in fact, during this period of replying to the supervisor, the screener's search mechanisms may have continued to move closer to the search termination threshold, regardless of the fact that the screener was not fully engaged in carrying out their visual search task. Of course, this would need to be tested in actual screening personnel, as their expertise and motivation to succeed may still over-ride the impact that distractions can have.

Indeed, participants here showed some evidence of increasing their RTs in single-target search when involved in the combined-task session, yet did not increase their RTs by an equal amount for dual-target search. This implies that they were either unwilling or unable to spend additional time conducting dual-target search in the combined-task condition. It also implies that the dual-target cost may also be related to search termination thresholds. Perhaps if participants were to search for *even* longer in dual-target search, then the dual-target cost for response accuracy may be attenuated.

Further examinations are needed to determine whether or not real-world visual search tasks may benefit from the present results, and to develop a full taxonomy of the impact of the various and complex forms of distraction that real-world visual searches may face. It is important to note that the target prevalence used in the present study was 50%, which is a markedly higher level than in real security screening. If a secondary task requiring the central executive slows search response times, how will performance be impacted in a low prevalence condition, where observers tend to respond very rapidly, and before carrying out a full inspection of the display? In low prevalence, will distractions actually *amplify* the prevalence effect? Such questions need to be addressed in future research, in order to develop both current theories, and bring the research closer to the real security screening environment.

8

Synthesis and Critical Review

Discussing the Empirical Results with Relevance to Theoretical and Applied Interests

8.1 Introduction

The central purpose of the empirical studies conducted during the course of the present thesis has been to explore a number of factors that have been thought to impair the threat detection performance of airport X-ray security screeners. Previous research has focused on screener performance in a number of diverse paradigms and frameworks (Gale, et al., 2000; Ghylin, et al., 2006; Schwaninger, 2004). Recently, the previous research has been extended via concerns raised regarding the prevalence effect (Wolfe, et al., 2005; Wolfe, et al., 2007), and the dual-target cost (Menneer, et al., 2004, 2007). Both of these phenomena were examined in detail during the course of the present thesis, and will now be discussed in turn.

8.2 The Prevalence Effect: Development and Novel Insights

Although the impact of target or stimulus frequency has long been examined (Estes, et al., 1957; Simpson & Voss, 1961), extending the role of target prevalence into current models of visual search is a relatively new endeavour. Thus far, only three papers have been published that have dealt with the prevalence effect with the goal of understanding the causes of the effect (Fleck & Mitroff, 2007; Wolfe, et al., 2005; Wolfe, et al., 2007). As a result, there is much work that needs to be done to understand the nature of the prevalence effect, and what it means in terms of the performance of airport screening personnel.

8.2.1 Relevance to Existing Models of Search

In the first empirical chapter (Chapter 2), it was noted that there are some marked similarities between accounts of the prevalence effect, specifically, the *criterion-shift* account of the prevalence effect (Wolfe, et al., 2007), and the *shifting-criterion* account of the stimulus probability effect (Miller & Bauer, 1981). Together, these accounts posit that, when a target or stimulus is presented less frequently, observers require more evidence to actually *perceive* and *report* that target or stimulus as being present. Although it has been argued that the prevalence effect could be caused entirely by motor priming (Fleck & Mitroff, 2007), there are a number of reasons to believe that motor priming is only a small part of the prevalence effect.

Evidence against the motor priming account comes from Wolfe et al. (2007), where, in one experiment, it was found that observers consistently missed the same targets in low prevalence. If participants were randomly guessing and responding 'absent' without due care and attention, then there should have been no element of consistency in the targets that they missed (Wolfe, et al., 2007). Furthermore, this thesis presented some strong evidence against motor priming accounts of the prevalence effect. In a paradigm that echoed the design of an earlier experiment (LaBerge & Tweedy, 1964), it was found that, when motor priming was held constant (by capping target prevalence at 50%), participants still missed targets in dual-target search when relative prevalence was varied (i.e. when one target was presented nine times more regularly than the other). If the prevalence effect was due to motor priming, then there should have been no reduction in hit rates for the lower-prevalence target in dual-target search.

If the prevalence effect is the result of a shift in response criterion, as would be predicted by Signal Detection Theory (D. M. Green & Swets, 1966), an important issue has been raised as to whether or not there is also a change in sensitivity (*d'*) when prevalence varies. Wolfe et al. (2007) reported that sensitivity was elevated when prevalence was low: this was a surprising result, especially considering that participants were missing a large number of the targets that were presented in low prevalence. Broadly speaking, sensitivity is a measure of *how well* observers can perform a task. Sensitivity increases when either hit rates increase or miss rates decrease, and it decreases when either hit rates decrease or miss rates increase.

Thus, as noted by Wolfe et al. (2007), sensitivity in low prevalence should not really be higher than in high prevalence, because the stimuli were the same (i.e. the task is equally difficult across all prevalence conditions). In two experiments, this rather surprising result was tested. There was a partial replication of Wolfe et al.'s (2007) results in Chapter 3, with *d'* increasing for low prevalence for both IEDs and dual-target search, but not for metals. However, in Chapter 4, *d'* did not increase in low prevalence. Fortunately, however, there are alternative measures of sensitivity that can be employed when *d'* is displaying aberrant behaviour. Chapter 4 made use of ROC curves and the sensitivity metric A_z, and detected no changes in sensitivity across a broad range of prevalence levels (replicating previous experiments which have shown the area under the ROC curve to be invariant with changes in prevalence: Gur, Rockette, Armfield, et al., 2003; Gur, Rockette, Warfel, et al., 2003). One could argue that the lack of any changes in A_z were not surprising, given that, in Chapter 4, no changes in *d'* were detected for the low prevalence condition: however, *d'* was substantially reduced for *high* prevalence, so, in fact, this was a valid method to use. Even when *d'* was reduced for high prevalence, A_z was not.

The experiments reported in the present thesis have therefore shown that the prevalence effect is not just due to motor priming, and that the prevalence effect is also the result of a genuine shift in the response criterion, with no change in sensitivity across prevalence levels. Additionally, the experiments reported here also examined *high* prevalence levels (>50% prevalence), which have been neglected in previous work. These revealed some rather surprising results: participants responded to variations in target prevalence in an *imbalanced* manner, responding 'present' more readily in high prevalence than they responded 'absent' in low prevalence. One very real and likely possibility is that 'present' and 'absent' responses are treated in a categorically different manner by the perceptual and response systems. The underlying reason for this could be that, even though there were no real consequences for missing targets in the experiments reported here, participants (especially the airport screeners) implicitly perceived that missing a target was more dangerous than producing a false alarm (as missing a target could lead to a threat item being used to damaging effect on an aircraft).

An alternative possibility is that, in fact, the imbalance between high and low levels of event occurrence are somehow an innate part of the cognitive architecture. Indeed, in a number of previous studies, it has been observed that participants tend to *overshoot* their estimation of how often an event has occurred when the probability of that event occurring is high. Erickson (1966) presented participants with a task in which they had to estimate how regularly a set of four lights flashed during a block of trials, and found that, when the lights flashed very regularly, participants consistently over-estimated how often those lights flashed. Using a similar paradigm involving estimating the regularity of occurrence of a set of flashing lights, Simpson and Voss (1961), and Myers and Cruse (1968) have reported similar results.

To highlight the imbalance in responding across a range of prevalence levels, consider the plots presented in Figure 8.2.1a. These plots were produced by calculating the criterion for each participant in Chapters 3 and 4. The average criterion for each participant was calculated across all setsizes and sessions, and is useful as a measure of response bias because it is somewhat removed from overall performance (i.e. it is based upon the z-transformed hit and false alarm rates, rather than the *actual* false alarm rates, which may vary between the two experiments because of different numbers of participants, and different numbers of sessions, etc.). To plot the regression measures in Figure 8.2.1a, a cubic function was used. What is perhaps most remarkable about these functions is how well they fit the data, in all three search conditions.

For single-target metals, the regression revealed a very close fit to the data indeed (adjusted R^2=0.92). An overall ANOVA conducted upon the model revealed the model to be significant (R=0.96, $F(3,41)$=150,p<.001). Compared to a linear function plotted through these data, the cubic function was a marked improvement (linear function adjusted R^2=0.82). Similarly, a substantial proportion of the variance in the criterion for single-target IEDs could be accounted for by a cubic function (adjusted R^2=0.87). Again, the overall ANOVA for the model itself was significant (R=0.94, $F(3,41)$=92.3,p<.001). As with single-target metals, the linear regression could explain less of the variance in the data (adjusted R^2=0.76). Finally, for dual-target search, the cubic model was once again significant (R=0.95, $F(3,41)$=121.5,p<.001), and was able to account for a substantial proportion of the variance (adjusted R^2=0.9), compared to the linear model (adjusted R^2=0.78).

To highlight the imbalanced manner in which the criterion shifts across variations in target prevalence, compare the degree to which the criterion changes across high (>50%) and low (<50%) prevalence levels. The curve is relatively shallow for low levels of target prevalence, whilst, for high prevalence levels, the curve is quite steep. This is the imbalance in the criterion shift: at high prevalence levels, the criterion shifts to a greater degree over the same change in prevalence (i.e. 50% up to 98% prevalence, giving a net change of 48%) than low prevalence levels (i.e. 2% up to 50% prevalence, still giving a net change of 48%, but a notably smaller criterion shift).

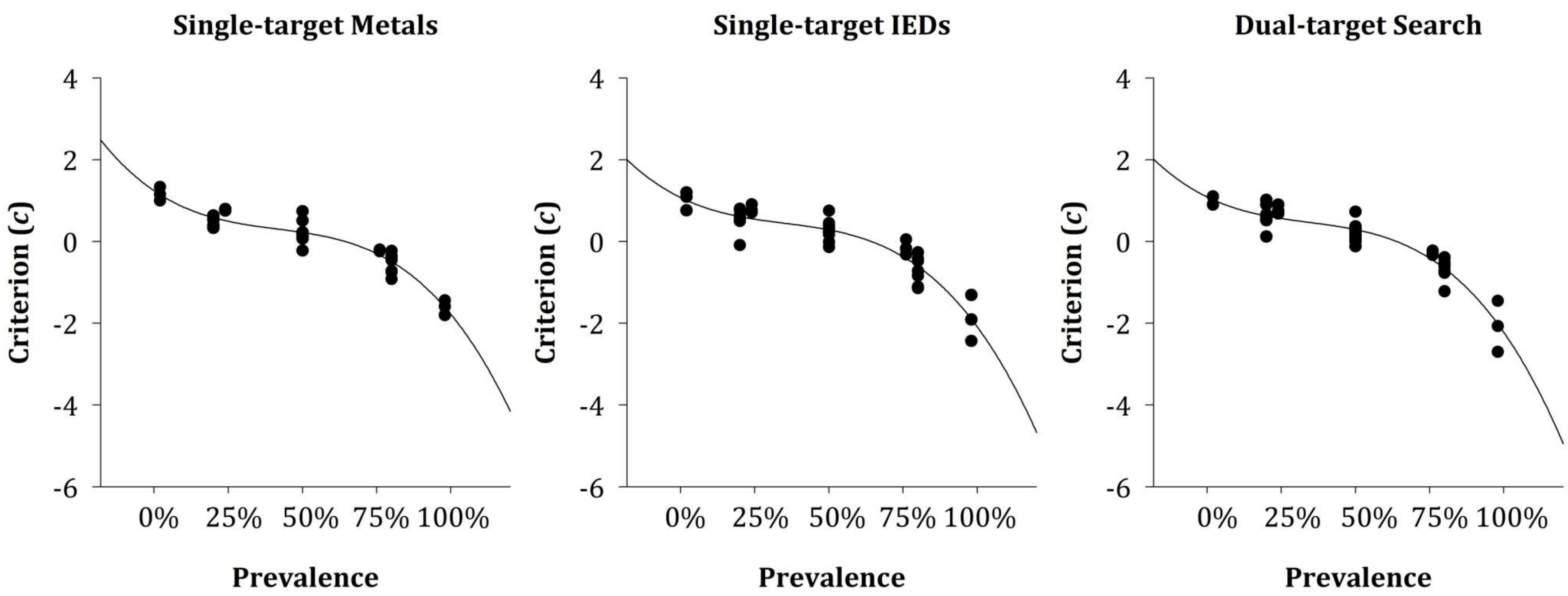


Figure 8.2.1a: Cubic regressions fitted to criterion levels across variations in target prevalence for three different search types. Criterion data for the experiments conducted in Chapters 3 and 4

8.2.2 Relevance to Airport Security Screening

Although the imbalanced responding across a range of prevalence levels is of theoretical interest, what does it mean for airport security screening? Perhaps most importantly, Chapter 5 demonstrated that airport screening personnel also exhibit a criterion shift, but only between 5% and 50% prevalence levels. Unfortunately, attempting to plot a cubic function to the criterion data for the screeners involved in Chapter 5 failed to provide any meaningful results (the ANOVAs for single-target metals, single-target IEDs and dual-target search criterion levels as prevalence varied were all non-significant: Fs<2.7, ps>.05). Part of the problem may be due to power, with only one session per search type per screener, and only six screeners in each prevalence condition overall, it is likely that more detailed analyses lack the power to provide conclusive results. Still, it is important to remember that the change in criterion across prevalence levels is *continuous* rather than *categorical* in nature, and thus, although there may be no significant differences between some of the consecutive levels of prevalence in any given experiment, the criterion is still assumed to be moving continuously as those prevalence levels change.

So, can the results provided here aid in the threat detection performance of airport screening personnel? In a set of earlier experiments, Wolfe et al. (2007) attempted to eliminate the prevalence effect by presenting observers with a series of high-prevalence 'burst' trials. They reported that high-prevalence bursts were able to shift the criteria set by participants, so as to eliminate the prevalence effect once the burst had ended. However, it was not clear how often the bursts would be needed in an operational setting, or how long the shift in criterion following a burst will actually last.

The results provided here suggest an alternative solution to 'curing' the prevalence effect, in a much more subtle, and, although it is irrelevant to theorists, a much more cost-effective manner than the proposals put forth by Wolfe et al. (2007). The key to understanding this solution is that the prevalence effect is *imbalanced and non-linear*, even at low levels of prevalence. Furthermore, the prevalence effect seems to be, at low levels of prevalence, centred on changes in miss rates, rather than changes in false alarm rates. For example, consider Chapter 3: miss rates significantly *increased* between 20% and 50% prevalence; however, false alarm rates were no different to one another in 20% versus 50% prevalence.

Turning to the screener performance data in Chapter 5, an even more marked effect can be observed in terms of the false alarm rates. False alarm rates across all prevalence levels for screeners were very low indeed. This is likely to be the result of the screeners' experience with non-targets: in other words, screeners may become experts at learning what is *not* a target, following many thousands of occasions on which they inspect a bag which does not contain a target. Results from a number of experiments have suggested that, when distractors are repeated, observers benefit considerably and show increased search efficiency (Maljkovic & Nakayama, 1994; McBride, et al., in press). Thus, screeners are experts in *non-targets* as well as targets.

The small changes in false alarm rates between low levels of prevalence can be described as a *compression* of false alarm rates, and the compression effect certainly appears to be amplified by experience, at least in the case of the screening personnel. Why might this be important? Consider for a moment the use of 'dummy' or TIP images in airport baggage displays. At present, TIP images are presented with a prevalence of around 2% (Schwaninger, et al., 2005), in order to monitor screener performance. Based upon the predictions of Signal Detection Theory, increasing the TIP rate would, of course, increase the hit rate, but, would have the unfortunate effect of increasing the false alarm rate. False alarms in screening are highly undesirable. They require detailed hand-searching of passenger baggage, and this can, of course, be especially problematic when there are long queues in the screening area.

The experiments conducted in the present thesis show that, in fact, the TIP rate could be safely increased. Doing so would cause a criterion shift in screening personnel, the benefit of which would be an increase in hit rates, with a *non-significant change in false alarm rates*. This is a cost-free solution: there would be a demonstrable increase in hit rates, with no cost in terms of an increase in false alarm rates. Essentially, this solution involves taking advantage of the imbalance in how observers respond to low levels of prevalence. Additionally, the implementation of increasing TIP rates would involve the simple tweaking of current TIP rates to increase hit rates to a desirable level. Wolfe et al. (2007) suggested that high prevalence bursts could reduce the prevalence effect. However, as already noted, the criterion shift following high-prevalence bursts may or may not be long-lasting in duration, and would require screeners to be retrained with high bursts of prevalence periodically. The solution suggested here, involving a baseline increase in TIP rates, would require no additional training, and would simply increase hit rates overall across all screeners. The duration of the benefit would last as long as the TIP levels were elevated to a desirable point.

8.3 The Dual-target Cost

8.3.1 Replication of the Dual-target Cost and Overall Relevance to Airport Screening

Airport screening personnel are tasked with searching for a large number of threat and prohibited items simultaneously. In the present thesis, the focus has been upon the core threat items: guns, knives, and IEDs.

For the most part, the dual-target cost has been detected, either in terms of reduced response accuracy in dual-target search, increased time taken to detect targets, or reduced response sensitivity (*d'*) in dual-target search. In some cases, the dual-target cost was not detected: most notably, there was no dual-target cost for response accuracy in Chapter 4 (although there was for RTs). The most likely cause of this was a lack of power, as there were only three participants in each condition.

Unfortunately, unlike the prevalence effect, there is no 'easy' solution to alleviating the performance impairments proffered by the dual-target cost. The only viable option appears to be a rather *cost-ineffective* solution of dividing the labour of screeners, such that each screener searches for a single target. This has been suggested previously (Menneer, et al., 2004, 2007), and, given that the screening personnel in Chapter 5 also showed reduced sensitivity in dual-target search, it seems that even expert screening personnel are not immune to the dual-target cost. Such a result is not really surprising, as previous attempts to train observers sufficiently so as to eliminate the dual-target cost have not been successful (Menneer, et al., under review).

Dividing the labour of screeners would be logistically very difficult to achieve. The strictest implementation would have each X-ray image being examined by *n* screeners, with *n* being equal to the number of different threat or prohibited items that screeners need to search for. One route to making this less problematic would be to group threat and prohibited items into categories of targets which are similar in appearance. Indeed, it is already convention to group guns and knives, which are both blue in colour, under the category of 'metals'. In a related study, Menneer et al. (under review) found that grouping same-coloured targets showed no dual-target cost, even when the targets were a different shape. Thus, metal threats could be grouped together; organic threats, including IEDs and liquids could also be grouped together. Further work is underway to assess whether or not individual screeners can be trained to carry out a *sequential dual-target search* strategy, whereby, upon being presented with an X-ray image to examine, the screener searches for one target at a time.

8.3.2 The Dual-target Cost and the Prevalence Effect

From an applied perspective, it was rather fortuitous that the experiments presented in this thesis suggested that the dual-target cost and prevalence effects did not interact with regards to response accuracy. One notable issue with current airport regulations is the fact that liquids are prohibited from being carried through the X-ray scanner. Passengers are required to remove any liquids from their baggage before having it placed in the scanner. This has met with considerable resistance, both from passengers, and from the media. However, the complaints against preventing liquids being carried onto aircraft may have been made prematurely. Given the relationship between the prevalence effect and stimulus probability effect that was examined in the present thesis (see Chapter 2), it may, in fact, be the case that requiring screeners to search for liquid threats actually *improves* their ability to search for IEDs.

On the surface, this may appear to be a rather bizarre suggestion, but consider once again that targets which are similar in appearance can be searched for with no dual-target cost (Menneer, et al., under review). Furthermore, recall that, when searching for one target that is presented more than a second target, observers focus on searching for the more-frequent target (Chapter 2; LaBerge & Tweedy, 1964): this is not the case, however, when the targets are similar in appearance (Dykes & Pascal, 1981; Miller & Bauer, 1981). How does this relate to the banning of liquids from baggage? As was noted by the screeners who participated in the experiment reported in Chapter 5, liquids currently have a very high prevalence level indeed. Passengers regularly leave liquids within their baggage, and, consequently, the screeners are regularly detecting those liquids and removing them. Now, given that IEDs have a primary explosive component that is orange is colour, and liquids are also primarily orange in colour, it can be therefore be suggested that having a high-prevalence orange target (liquids) will also improve screeners' ability to detect liquids. Although this remains to be tested empirically, there is already precedent for such a suggestion in two experiments within the stimulus probability literature (Dykes & Pascal, 1981; Miller & Bauer, 1981). Future research may benefit from testing this directly, of course.

8.4 External Distraction and Search Performance

Previous claims have been made that airport X-ray security screeners work in a 'performance-degrading environment' (Harris, 2002), although, to date, there have been few actual tests of such claims. The present thesis began what will eventually be a substantial body of work by investigating the impact of ambient noise upon search performance, as well as the impact of a simple mental arithmetic task upon search performance.

To begin with, it was found that ambient noise had no impact upon search performance (Chapter 6): this is perhaps one of the first pieces of 'good news' in terms of the experiments that have been conducted into the potentially detrimental effects that real-world factors have upon screener performance. Simply having ambient noise played whilst conducting visual search does not increase RTs, increase the dual-target cost, or decrease response accuracy.

However, with respect to the results presented in Chapter 7, it is clear that conducting a simple arithmetic task in conjunction with a visual search task produced a cost in response accuracy, and a notable impact upon response time. Considering that Chapters 6 and 7 were intended as 'first steps' towards understanding the impact of the screening environment upon search performance, much more work is needed to understand the phenomena involved in more detail. Chapter 7 used a secondary task which participants could easily ignore if necessary. In the screening environment, this may not be possible when another individual approaches a screener and requires a question to be answered immediately. Furthermore, it is not clear how the effects of environmental distraction may interact with other factors, such as the prevalence effect. Again, further work will be needed to understand these issues in more detail.

One further factor to consider with regards to the screening environment is that of *deliberate self-distraction* whilst screening. Three of the screeners during debriefing for the experiment conducted in Chapter 5 reported that they regularly watched the passengers queuing to have their baggage X-rayed. The screeners claimed that they examined the passengers for any signs of suspicious activity, such as 'unusual' behaviour, or moving backwards in the queue, as if to avoid having their bags scanned. It is unlikely that any activities involving watching the passengers are actually beneficial to screener performance, and indeed, any time spent watching passengers is time spent *not searching* the X-ray display. As a result, it may be very beneficial to prevent screeners from observing the area around them. Given that ambient noise appears to not detract from search performance, a small partition could be placed around screeners, to reduce the visual distraction that the busy airport environment offers. This would be a relatively cost-effective method as well.

8.5 Summary: Key Results and Future Questions

8.5.1 Key Results

In brief, the key results reported in the present thesis will now be summarised:

1. **The prevalence effect is imbalanced across variations in prevalence**

   The result of this is that airport screeners may benefit from having increased TIP rates, with an increase in hit rates, but no significant increase in false alarms.

2. **The dual-target cost can not be eliminated, even for screeners**

   Screeners showed a dual-target cost, and, as a result, real screener performance will likely benefit from a division of labour, with a number of screeners searching each X-ray for different categories of target.

3. **Ambient noise does not impair performance, but external distraction does**

   Although there was no negative impact of ambient noise upon searching whilst in the presence of ambient noise, performance was somewhat impaired by a simple mathematical task.

8.5.2 Future Questions and Further Factors to Consider

As the goal of the present thesis was to integrate a number of existing notions regarding visual search, it is important to note that there are still a number of further factors that require consideration before a more complete picture can be formed. These factors are all outside of the scope of the present work, so will only be described briefly here.

Overlap of Images: One notable difference between the displays used in the experiments reported here, and the actual screening displays, is that X-rays of passenger baggage tend to have a considerable degree of overlap between different objects. In other experiments, Wolfe et al. (2007) actually used overlap in displays that more closely resemble X-ray images from screening, and find qualitatively similar results to those reported here. The only difference that having overlap will make is that, in fact, having overlap in displays will make those displays more difficult. A number of examinations of the impact of overlap or 'superposition' have shown that both novices and real airport screeners show decreased detection rates when the target overlaps with other objects in a bag (Koller, Hardmeier, Michel, & Schwaninger, 2007; Schwaninger, et al., 2008).

Ageing: The participant screeners who were involved in the experiment reported in Chapter 5 were notably older than those who were involved in the other experiments reported in the present thesis. This is not a problem for the results, because the research questions posed in Chapter 5 did not resolve around comparing novices to experts (i.e. comparing the screeners to a set of undergraduate students who were much younger). The focus was on whether or not screeners were affected by target prevalence and the dual-target cost. The role of ageing is interesting for a number of reasons, however. A number of studies have demonstrated that older individuals show poorer search performance than younger individuals (for example, see Hommel, Li, & Li, 2004). Additionally, studies of ageing have shown that, with age comes an increased susceptibility to distraction (Stevens, Hasher, Chiew, & Grady, 2008). It would be interesting, therefore, to repeat the experiments conducted in the present thesis on a more elderly group of participants, to assess their susceptibility to the dual-target cost, prevalence effect, and also the effects of ambient noise and distraction.

Video Games: A number of studies have now demonstrated that individuals who spend time playing computer games show improved performance in visual search, and a number of other perceptual tasks (Castel, et al., 2005; C. S. Green & Bavelier, 2003, 2007; Riesenhuber, 2004). Indeed, Fleck and Mitroff (2008) reported that Video Game Players (VGPs) showed no prevalence effect. VGPs withheld their 'absent' responses, and gave themselves more time to detect the target in low prevalence, thereby enabling them to eliminate the prevalence effect. It is interesting to note that the screeners involved in the experiment conducted in Chapter 5 also showed increased target-absent RTs. At present, it is unclear whether the changes in expert behaviour (i.e. for screeners or VGPs) in the face of low target prevalence is the result of motivational of perceptual factors. Further work may also benefit from examining whether or not the dual-target cost can be alleviated or eliminated by experience playing video games. Considering the rather costly methods that would be needed to eliminate the dual-target cost (the division of labour approach, described above), then it would certainly be worth examining such a possibility.

Time Pressure: One rather notable difference between the screening task and the experiments conducting during the course of the present thesis is that there may be some elements of time pressure when screeners search X-ray images of passenger baggage. Here, no time pressure methods were employed, in order to gain a basic hold on the phenomena that were under examination. Previous studies have reported that, when under time pressure, individuals can often make irrational, or 'risky' decisions. For example, Dror, Busemeyer, and Basola (Dror, Busemeyer, & Basola, 1999) engaged participants in a simplified version of the game 'blackjack'. Participants were presented with two cards and controlled the degree to which the score on the cards was close to the target score (21). Somewhat surprisingly, when the participants were under a high degree of time pressure, they were *more likely* to make a risky decision (i.e. ask for another card when the score on their given cards was already high) than they were when the time pressure was low. When faced with lengthy queues during busy periods, screeners may thus opt to behave in a more risky fashion, responding 'absent' more rapidly to reduce the queues and increase passenger throughput. This may, as a result, enhance the prevalence effect, where observers tend to respond rapidly anyway. Although a game of blackjack is clearly different from the task given to screening personnel, further work is underway to examine the relationship between visual search, time pressure, and target prevalence in more detail.

8.5.3 Closing Remarks

The present thesis took a set of real-world factors and applied them to current models of visual search. A number of developments have been made, both in terms of progress towards resolving theoretical issues, and towards improving the search performance of airport screening personnel. Still, as the previous section has highlighted, there are a large number of factors that need to be examined in order to gain a more complete account of the airport screening task, as well as, in turn, visual search when conducted in the 'real world'. Primarily, future work is needed to explore the manner in which the factors that have been explored here interact with other factors that are yet to be examined. Ultimately, the goal, and indeed the challenge, will be to develop the task given to airport screening personnel, so that they are performing at optimal rates, without any hindrance to the actual security of the airport, and without imposing an unreasonable cost upon the operators.

Appendix A

Sample images from Airport Screening used as Stimuli in the Present Thesis

The contents of Appendix A have been removed for security reasons.

B

Appendix B

Signal Detection Theory: An Overview and Description

B.1 Introduction

The detection and recognition of a stimulus in the outside world (a *distal* stimulus) involves a decision. At a neural level, the brain must decide whether or not the incoming information represents a *target* or a *non-target*. This process is made problematic by the fact that targets and non-targets are often quite similar to one another, and because the perceptual process itself is quite noisy, and can often give a less-than-perfect account of the distal stimulus. The goal of Signal Detection Theory (D. M. Green & Swets, 1966; Macmillan & Creelman, 2005) is to explore the processes involved in how an observer decides that a target is present or not, or in the language of Signal Detection Theory, whether or not a given trial contains either a *Signal* or *Noise*. Much of the language of Signal Detection Theory (including terms such as Signal and Noise) are taken from its early use, where the focus was upon examining the performance of radar operators in the Second World War: since that time, Signal Detection Theory has met with considerable success, and has been used in a wide variety of different applications (Macmillan & Creelman, 2005). The purpose of the present Appendix is to describe Signal Detection Theory, and give details of how Signal Detection Theory parameters are produced, and how those parameters are typically interpreted. The account given here is based upon a number of key sources (D. M. Green & Swets, 1966; Macmillan & Creelman, 2005; Verde, Macmillan, & Rotello, 2006; Wickens, 2002), and further sources are cited where appropriate.

B.2 Basic Concepts

Exploring the manner in which observers make decisions regarding the complex and often noisy perceptual processing of distal stimuli may, at first, appear to be an insurmountable problem. However, Signal Detection Theory makes a set of assumptions that allow us to build up an account of the decision processes being used in a task. Signal Detection Theory is founded upon the robust General Linear Model, and begins by assuming that the internal response generated by a distal stimulus can be represented as a number. Furthermore, Signal Detection Theory assumes that, in any trial in an experiment, an observer proceeds by sampling information from the display (although many different types of trial can be used for Signal Detection Theory experiments, the focus here is on a visual search task, so observers in these examples are examining a display and searching for a target amongst a set of non-targets).

The information that is gathered from the display is represented on a numerical continuum. That information is then used to make a decision about whether or not a target is present. In most cases, it is assumed that the Signal and Noise stimuli can be plotted on the continuum as two normal distributions, which often overlap with one another. This formulation is useful for a number of reasons. First of all, the overlap allows us to gain a measure of the intrinsic uncertainty in perception: as already noted, perception can often be quite noisy and imperfect, and, furthermore, targets and non-targets can often be quite similar (especially in the case of visual search). It therefore makes sense that an observer should be able to confuse a target with a non-target, because real observers often mistakenly report a target to be present when there is no target present, and also often mistakenly report that no target is present, when, in fact, there is a target present in a display. Perhaps more importantly, though, is that the modelling of the Signal and Noise distributions allow for the development of measures of observer performance in a task, via *criterion* and *sensitivity* parameters.

B.3 Measures of Criterion and Sensitivity

In a situation where an observer must classify a stimulus into one of two categories (i.e. report ‘present’ or ‘absent’, or, analogously, report ‘signal’ or ‘noise’), the underlying processes involved in the observer’s decision classifications are examined by Signal Detection Theory upon the basis of several response variables. A *hit* occurs when a Signal is correctly detected, a *miss* occurs when a Signal is present and not detected, a *false alarm* occurs when a Signal is reported, but is in fact not present, and a *correct rejection* occurs when a Signal is absent and reported to be so. An overall account of performance can therefore be given based upon the number of hits, misses, false alarms and correct rejections.

For the most part, the focus is upon the hit rate (h) and the false alarm rate (f). Both h and f must be considered when attempting to describe the performance of any given observer. Returning to the central issue of how an observer decides whether or not a target is present in a given display, it is assumed under Signal Detection Theory that a *criterion* is invoked to decide when to respond 'present' and when to respond 'absent'. Recall that the evidence gathered from the distal stimulus is represented as a number. The simple manner in which the criterion is used is as follows: the observer sets a decision criterion such that, when the information that is gathered is higher than the criterion, a 'present' response is given, and, when the information that is gathered is lower than the criterion, an 'absent' response is given. This process is depicted in B.3a, below.

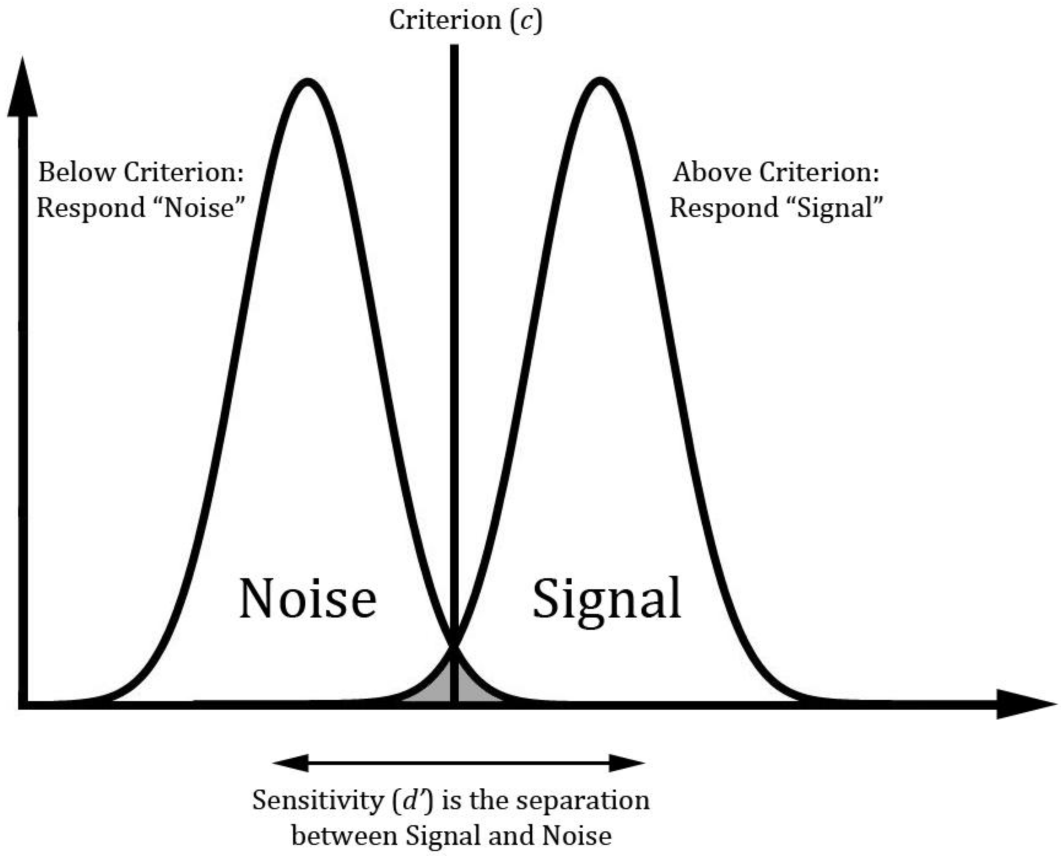


Figure B.3a. Illustration of basic signal detection concepts, including the Noise, Signal, Criterion and Sensitivity. The shaded area between the Noise and Signal are areas of overlap where stimuli are confusable, and could be identified as Signal *or* Noise. The Criterion divides the shaded area, so only the shaded area to the right of the Criterion is reported to be a Signal. Of course, this then means that the shaded area to the left of the Criterion that falls under the Signal distribution is reported as 'Noise'.

The criterion can be set to different levels, which causes changes in both the hit rate and the false alarm rate. The notion of the decision criterion (denoted by c) is invoked to account for how the Noise and Signal distributions are divided and for how responses are produced. The criterion can be set to different levels, which causes changes in both the hit rate and the false alarm rate. It is important to note that Signal Detection Theory assumes h and f to be yoked together: one can not really improve the probability that a target will be detected without also increasing the probability that a false alarm will be made. Consideration of the response criterion is important, because, for example, it can be used to give an account of situations where an observer reports a target to be present very regularly, achieving a high value of *h*, but, at the same time, also achieving a high value of *f*. In such a situation, the observer can be said to be *biased* towards reporting the presence of a Signal. Had analyses of the observer's performance focused solely on *h* whilst ignoring *f*, then it would have been said that the observer was performing the task very well: however, with a high value of *f*, one would be inclined to assume that the observer was simply responding 'present' too readily.

A useful mnemonic for the criterion is to label it as a measure of how *conservative* an observer is in making their responses. When observers are more conservative, they require more evidence before responding 'present'; when observers are less conservative, they respond 'present' liberally. This process is depicted in Figure B.3b below, with the criterion being shifted to a more liberal position (in the figure, the criterion has been moved to the left). This now includes more of the Signal than in Figure B.3b, which thus increases the hit rate. However, as the shaded area indicates, more of the Noise is now included, thus increasing the chance that a false alarm will be made. Adopting a more conservative criterion has the opposite effect: as shown in Figure B.3c, the moving criterion now decreases the hit rate, but also decreases the false alarm rate (as shown in the shaded area).

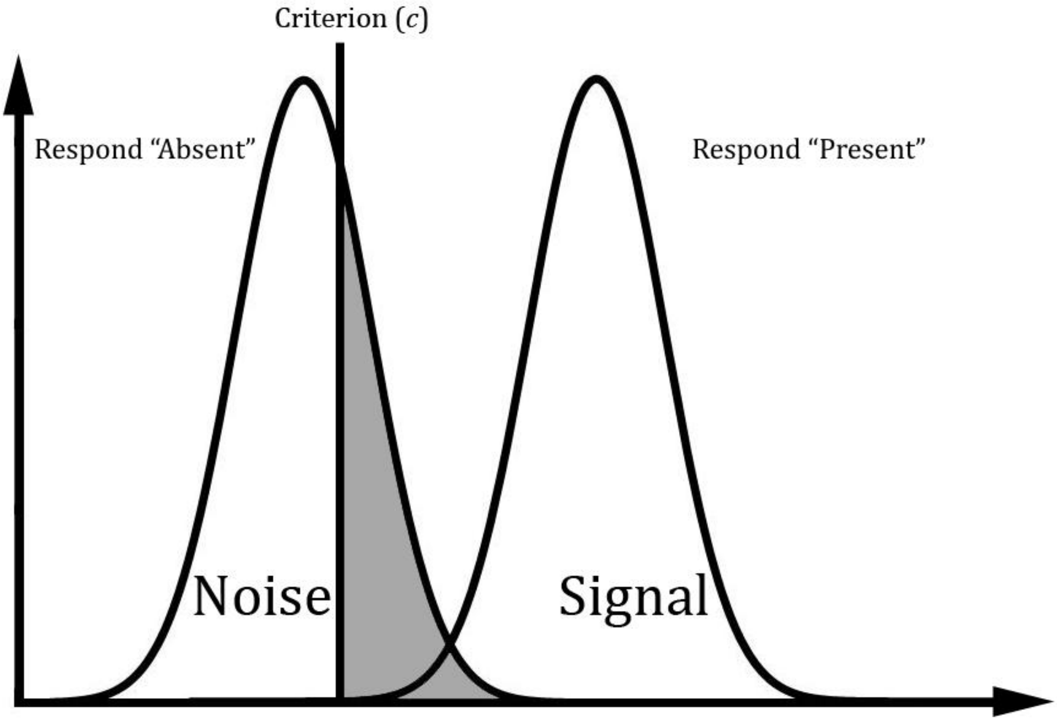


Figure B.3b. Illustration of a liberal criterion from a signal detection perspective. The shaded area falling under the Noise distribution will be reported to be a Signal, generating false alarms.

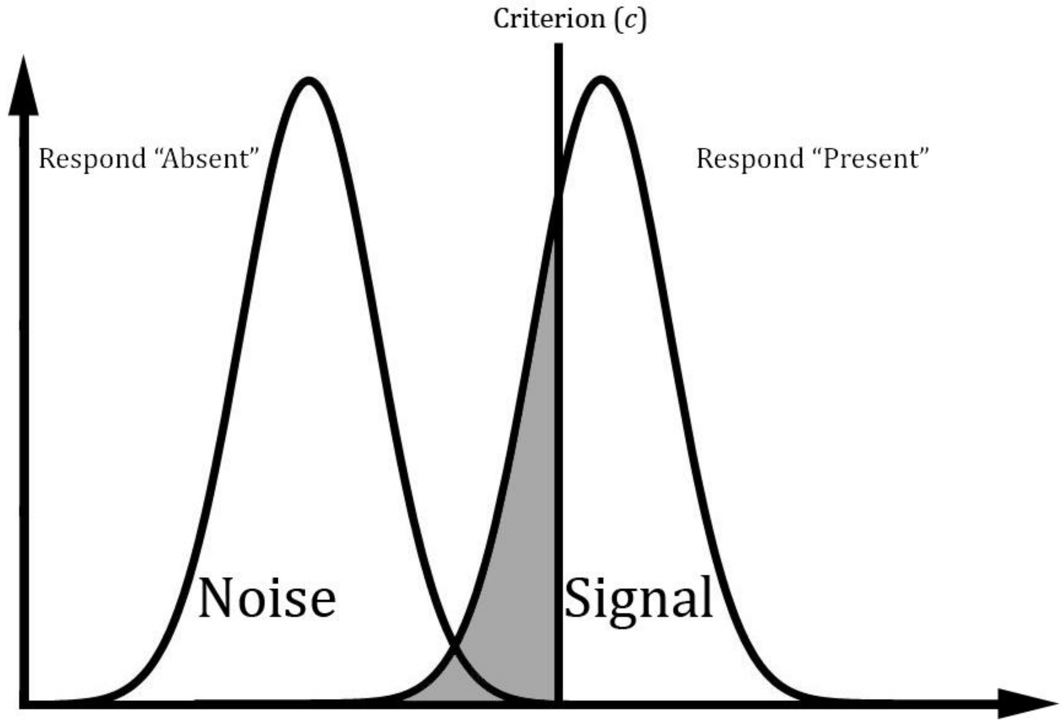


Figure B.3c. Illustration of a conservative criterion from a signal detection perspective. The shaded area falling under the Signal distribution will be reported to be Noise, generating misses.

As already noted, the decision regarding the presence or absence of a target is made considerably more difficult when the Signal and Noise distributions overlap with one another. In such instances, the stimulus being observed could be either a Signal or Noise: the true answer is ambiguous to the observer. The ambiguity can often be resolved by the setting of the criterion, but Signal Detection Theory also offers a measure of the overlap of the distributions, in the index of *sensitivity*.

Put simply, response sensitivity is a measure of how easily the Signal stimuli can be discriminated from the Noise stimuli. For this reason, sensitivity is also often referred to as *discriminability*. In a task where the Signal is easily discriminated from the Noise, the Signal and Noise distributions are distant from one another, resulting in high sensitivity. Returning to Figure B.3a, above, the sensitivity can be sketched as the separation between the Signal and Noise stimuli. When the task is made more difficult, and the targets are more similar to the non-targets, the Signal and Noise distributions move closer to one another, resulting in lower levels of sensitivity. For the most part, sensitivity or discriminability are indexed by the parameter *d'*. The depiction of sensitivity in Figure B.3a represents a rather high level of sensitivity, as the Signal and Noise distributions are relatively distant from one another. Compare this to Figure B.3d, below, where the distributions show a considerable degree of overlap. In such a scenario, there will be a large number of false alarms, and a large number of misses, as participants find it difficult to distinguish between Signal and Noise trials.

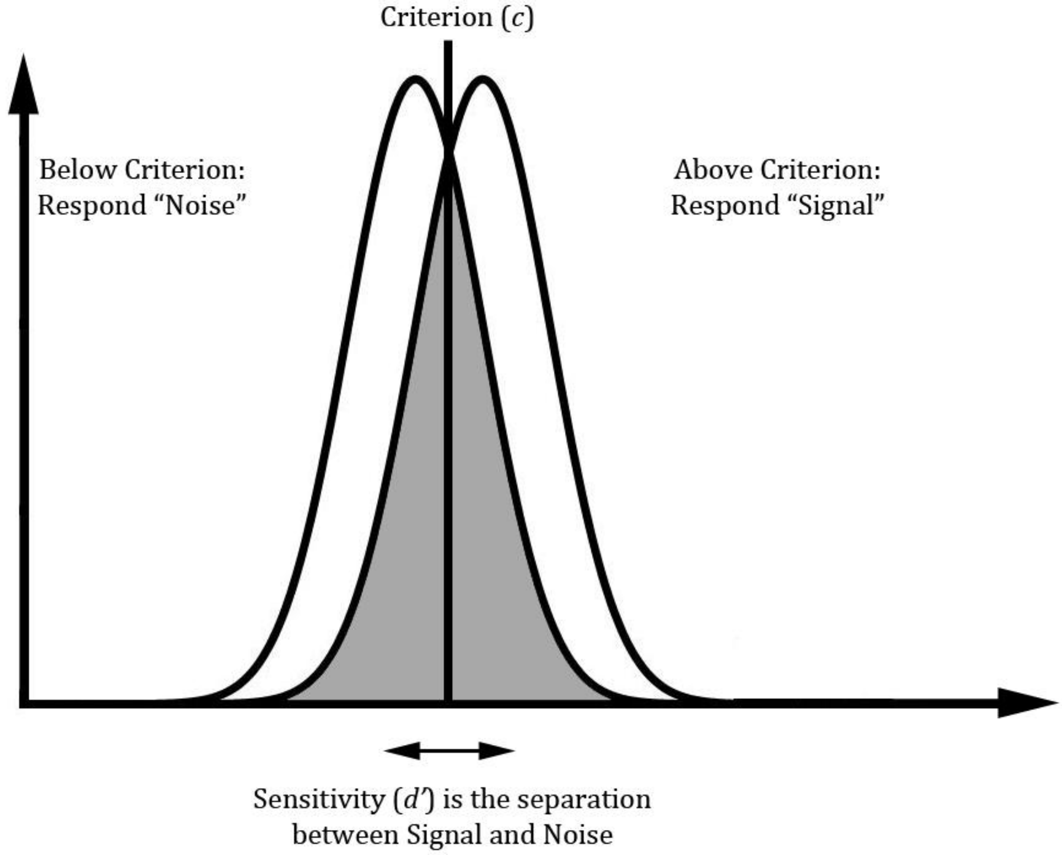


Figure B.3d. Low sensitivity: the Noise and Signal Distributions have considerable overlap (the shaded area), so the hit rate decreases and the false alarm rate increases, compared to Figure B.3a.

B.4 Calculating Criterion and Sensitivity: the Mathematics of Signal Detection Theory

The distributions of both Signal and Noise are typically assumed to be *normal* or *Gaussian* in nature. With this assumption in mind, the values of c and d' can be calculated via a set of mathematical functions which will not be discussed in detail here. Throughout the present thesis, the Signal Detection Theory parameter for the criterion will be calculated as follows:

$$c = -0.5[z(h) + z(f)]$$

The criterion is calculated using the z-transformed hit and false alarm rates. A c value of 0 represents neutral bias, i.e. no response bias, whilst a value of greater than 0 represents a bias towards responding *absent*, and a value of less than 0 represents a bias towards responding *present*.

The Signal Detection Theory parameter for sensitivity will be calculated using the formula:

$$d' = z(h) - z(f)$$

As with the criterion, d' is also calculated using the z-transformed hit and false alarm rates. Higher values of d' are produced when observers achieve higher hit rates, and lower false alarm rates. It should be noted that c and d' are not the only Signal Detection Theory parameters available for examining performance: for example, the response criterion can be computed in a somewhat different manner, and there alternatives to d', yet, for various reasons, the alternatives have been criticised (the details of these issues are beyond the scope of the present discussion, although see Verde, et al., 2006 for more information). A further reason to focus upon using d' and c is that a key study of a number of the issues that are under examination in the present thesis, Wolfe et al. (2007), also used d' and c. Thus, to maximise the effectiveness of replicating and extending previous work, d' and c will be primarily used here.

B.5 The Receiver Operating Characteristic Curve

In most cases, for a given task, variations in the response criterion can occur, and, if the actual task itself remains unchanged, response sensitivity should remain unchanged as well. Referring back to Figures B.3b and B.3c, it is the case that the criterion can move, but, unless the stimuli are changed (i.e. the locations of the Signal or Noise distributions move), the sensitivity should be unaffected by changes in the criterion. With this in mind, an *isosensitivity* curve can be produced (with *iso* meaning 'same'), plotting the full range of variations in the criterion, as sensitivity is held constant. Another name for such a curve is a *Receiver Operating Characteristic* curve, or ROC curve. The ROC plots the false alarm rate against the hit rate: an example is given below in Figure B.5a.

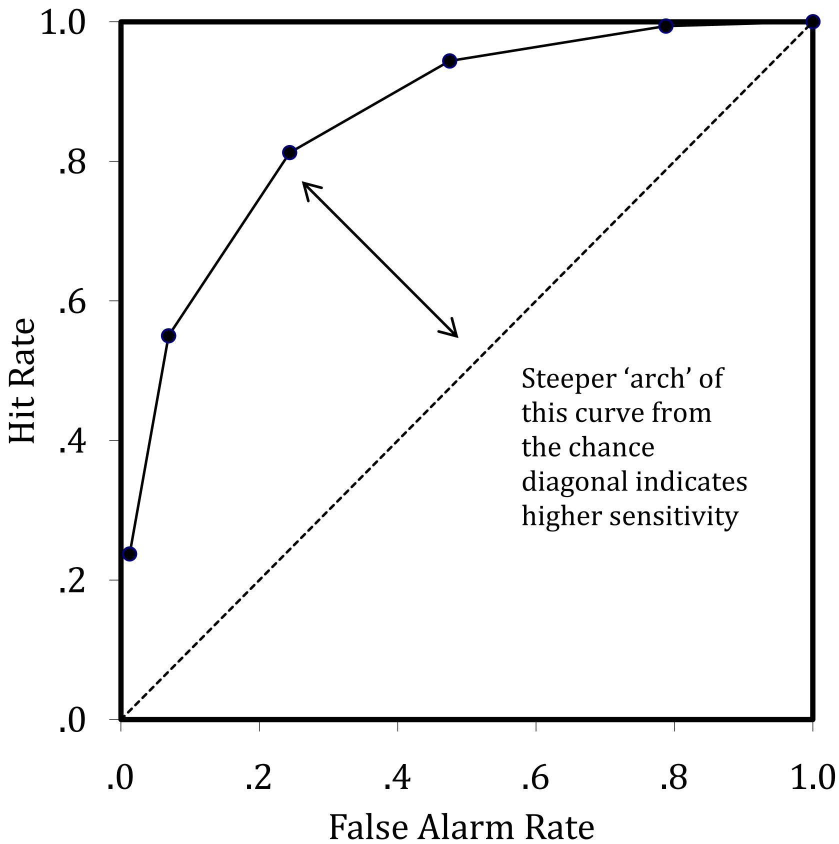


Figure B.5a: Example ROC curve, with hit rate plotted against the false alarm rate. The chance performance line is indicated as a dotted line.

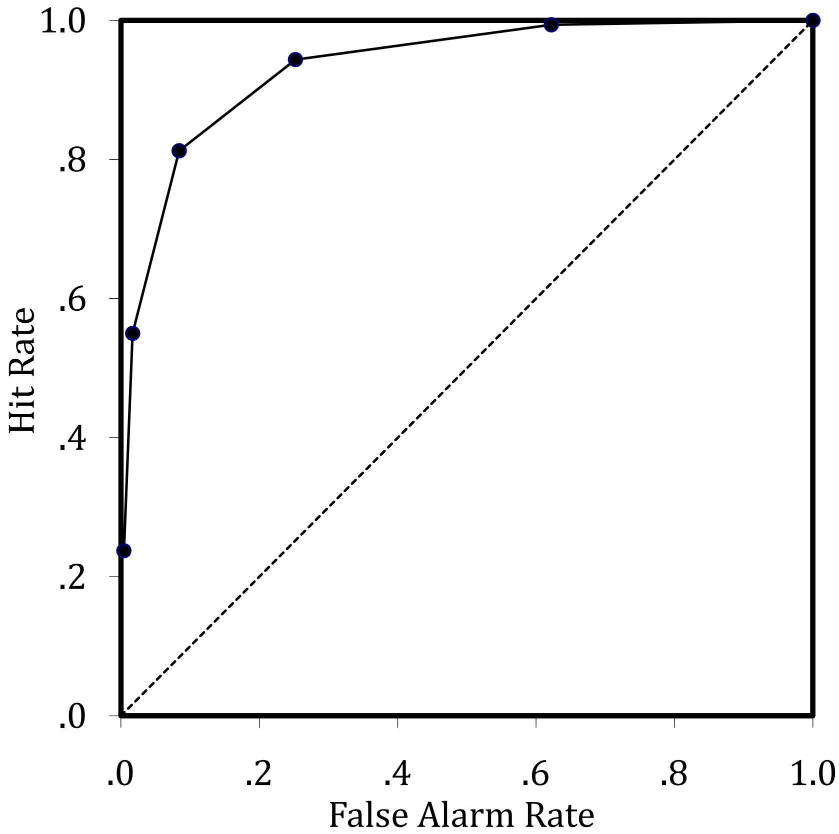


Figure B.5b: Example ROC curve, with hit rate plotted against the false alarm rate. This shows a higher level of sensitivity than in Figure B.5a. The chance performance line is indicated as a dotted line.

The overall 'arch' of the ROC curve gives a measure of sensitivity: the greater the arch, the greater the sensitivity (Figure B.5a shows a higher level of sensitivity in Figure B.5b, above). The importance of the arch of the ROC curve can be understood simply by examining how the hit and false alarm rates are altered to produce a movement of the arch itself: when the arch of the ROC curve moves towards the upper-left corner of the Figures presented above, then the hit rate is higher and the false alarm rate is lower. This will have the effect of increasing sensitivity (*d'*). Perfect performance, which would occur with a 100% hit rate and 0% false alarm rate, would be placed within the top-left corner. Chance performance, plotted in the above figures as a dotted line, occurs when the hit rate equals the false alarm rate.

One important feature of the ROC or isosensitivity curve is that sensitivity (i.e. *d'*) should not change as the criterion varies. Conventional wisdom holds that there are two possible routes which can be followed to result in a shift in the criterion:

first of all, a change in Signal probability (i.e. when a Signal occurs infrequently, then the criterion shifts towards becoming more conservative, when a Signal occurs frequently, the criterion shifts towards becoming more liberal); second, through the induction of a payoff matrix (i.e. providing incentives to increase the hit rate via, for example, payment for hits). In many experiments, the second route is adopted, primarily because, having low Signal probability rates requires participants to be engaged in many hundreds of trials.

An alternative, and more practical, method that is used is to ask participants to provide a *confidence rating* on each trial in an experiment. The rating reflects how confident they are that a signal is present. At one end of the scale, participants are certain that a Signal is *present*; at the other end, they are certain that a Signal is *absent* (i.e. they are reporting a Noise or "target-absent" trial). Confidence ratings are collated and the distribution of the ratings is used to produce an ROC curve. Thus, in the example ROC curves above (Figures B.5a and B.5b), the points on the curve could, in one experiment, be different conditions for different payoff regimes, or, the points could be, in a separate experiment, different conditions with different Signal frequencies, or, finally, could be different points on a rating scale in a single experiment. This is displayed in more detail below in Figure B.5c.

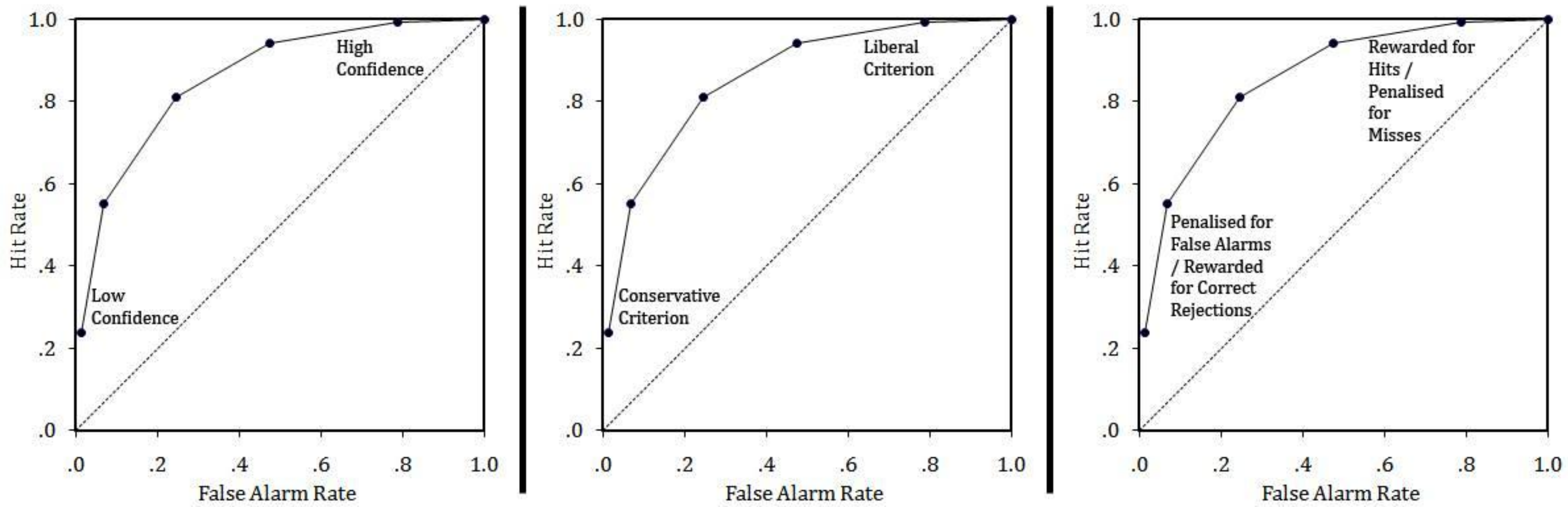


Figure B.5c: ROC curves generated using different (example) methods. The left-hand graph depicts an ROC curve generated by asking participants to give confidence ratings; the central graph depicts an ROC curve generated by variations in the criterion (caused by changes in signal frequency); the right-most graph depicts an ROC curve generated by different payoff regimes.

The ROC curve allows for a number of useful analyses; in particular, it gives an opportunity to use an alternative measure to *d'*. Based upon the ROC curve, it is possible to compute the *area under the ROC curve* or A_z. Given that, as already noted, a curve with a higher 'arch' (i.e. closer to perfect performance with 100% hit rate and 0% false alarm rate) will then show higher levels of sensitivity, this means that A_z will be higher when sensitivity is higher. One important issue that will be explored within the present thesis is whether or not sensitivity (measured in terms of A_z) changes when signal frequency is varied. In order to explore this issue, confidence ratings are taken under several different levels of signal frequency, and then compared (see Chapter 4).

B.6 The Normalised Receiver Operating Characteristic

Although the interpretation of an experiment's results in terms of the criterion and sensitivity measures is often a straightforward process, there can sometimes be problems with understanding the results in terms of Signal Detection Theory. The core problem that will be examined within the context of the present thesis is a situation whereby *d'* increases along with increases in *c*. As already noted, Signal Detection Theory assumes that, for the most part, sensitivity will remain unchanged (unless the given task itself changes) whilst the criterion varies. This assumption is based upon the modelling of the Signal as Noise distributions, which are treated as being normally distributed. When both Signal and Noise are normally distributed, changes in the criterion do not affect the resultant calculation of *d'*. However, when the variance of the Noise is greater than that of the Signal (as depicted below in Figure B.6a), then *d'* is no longer invariant with changes in the criterion.

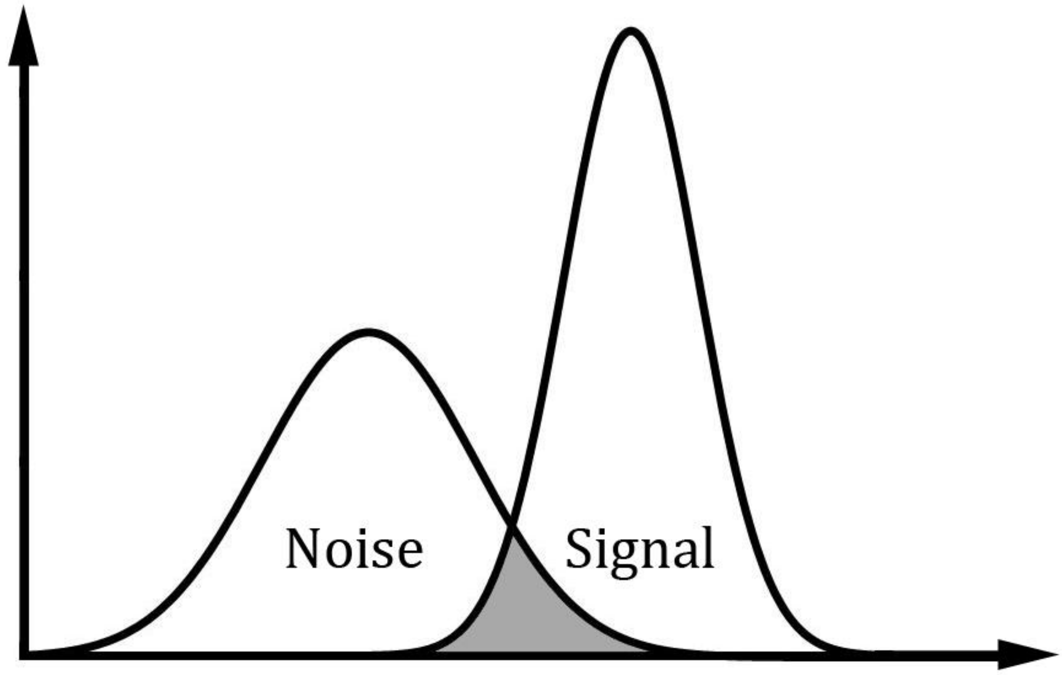


Figure B.6a. Unequal variance of Signal and Noise. Note how the shaded area is now skewed: *d'* will provide an inaccurate estimate of sensitivity in such a situation.

A useful way of examining such a situation is to plot the *normalised ROC* or *zROC* curve. Recall that calculations of *d'* and *c* are made using the *z*-transformed hit and false alarm rates. These *z*-transformed rates can then be used to plot the zROC curve. When plotted in such a manner, the zROC curve should have a slope of unity (i.e. a slope of 1): this occurs only when the variance of the Noise and Signal are equal to one another. When the slope is either less than or greater than 1, it is likely that a situation such as depicted above in Figure B.6a is in effect. Unfortunately, unequal-variance scenarios are somewhat open to interpretation, and not easily understood or accounted for (Macmillan & Creelman, 2005). Further details on the use of zROC curves are given in Chapter 3.

B.7 Computational Notes

Modern statistical packages (such as *SPSS* and *Microsoft Excel*) can be used to compute Signal Detection Theory parameters. For the calculation of *c* and *d'*, the NORMSINV function (i.e., the *inverse normal* function) can be used to *z*-transform the hit and false alarm rates.

Throughout the present thesis, the detection parameters *c* and *d'* were calculated using *Microsoft Excel 2007* (using formulae provided by Macmillan & Creelman, 2005). The area under the ROC curve (A_z) was calculated using the specialised ROC algorithms in *SPSS* version 16.0 for Windows.

References

Ashby, F. G. (2000). A stochastic version of General Recognition Theory. *Journal of Mathematical Psychology, 44*(2), 310-329.

Baddeley, A. D., Cocchini, G., Della Sala, S., Logie, R. H., & Spinnler, H. (1999). Working memory and vigilance: evidence from normal aging and Alzheimer's disease. *Brain and Cognition, 41*(1), 87-108.

Balota, D. A., & Rayner, K. (1983). Parafoveal visual information and semantic contextual constraints. *Journal of Experimental Psychology-Human Perception and Performance, 9*(5), 726-738.

Banbury, S. P., & Berry, D. C. (1997). Habituation and dishabituation to speech and office noise. *Journal of Experimental Psychology-Applied, 3*(3), 181-195.

Banbury, S. P., & Berry, D. C. (1998). Disruption of office-related tasks by speech and office noise. *British Journal of Psychology, 89*, 499-517.

Banbury, S. P., & Berry, D. C. (2005). Office noise and employee concentration: Identifying causes of disruption and potential improvements. *Ergonomics, 48*(1), 25-37.

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1996). Visual search for colour targets that are or are not linearly separable from distractors. *Vision Research, 36*(10), 1439-1465.

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1998). The linear separability effect in color visual search: ruling out the additive color hypothesis. *Perception and Psychophysics, 60*(6), 1083-1093.

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1999). Convex hull test of the linear separability hypothesis in visual search. *Vision Research, 39*(16), 2681-2695.

Bichot, N. P., Rossi, A. F., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science, 308*(5721), 529-534.

Biederman, I., & Stacy, E. W. (1974). Stimulus probability and stimulus set size in memory scanning. *Journal of Experimental Psychology, 102*(6), 1100-1107.

Caggiano, D. M., & Parasuraman, R. (2004). The role of memory representation in the vigilance decrement. *Psychonomic Bulletin and Review, 11*(5), 932-937.

Campbell, T., Beaman, C. P., & Berry, D. C. (2002). Auditory memory and the irrelevant sound effect: Further evidence for changing-state disruption. *Memory, 10*(3), 199-214.

Carrasco, M., Evert, D. L., Chang, I., & Katz, S. M. (1995). The Eccentricity Effect - Target Eccentricity Affects Performance on Conjunction Searches. *Perception & Psychophysics, 57*(8), 1241-1261.

Carrasco, M., & Frieder, K. S. (1997). Cortical magnification neutralizes the eccentricity effect in visual search. *Vision Research, 37*(1), 63-82.

Carrasco, M., Mclean, T. L., Katz, S. M., & Frieder, K. S. (1998). Feature asymmetries in visual search: Effects of display duration, target eccentricity, orientation and spatial frequency. *Vision Research, 38*(3), 347-374.

Carter, N. L., & Beh, H. C. (1987). The effect of intermittent noise on vigilance performance. *The Journal of the Acoustical Society of America, 82*(4), 1334-1341.

Castel, A. D., Pratt, J., & Drummond, E. (2005). The effects of action video game experience on the time course of inhibition of return and the efficiency of visual search. *Acta Psychologica, 119*(2), 217-230.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: how are visual searches terminated when there is no target present? *Cognitive Psychology, 30*(1), 39-78.

Connor, C. E., Egeth, H. E., & Yantis, S. (2004). Visual attention: Bottom-up versus top-down. *Current Biology, 14*(19), R850-R852.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201-215.

Cowan, N. (1995). *Attention and memory: An integrated framework*. London, England: Oxford University Press.

Creelman, C. D. (1965). Discriminability and scaling of linear extent. *Journal of Experimental Psychology, 70*(2), 192-200.

Creelman, C. D., & Donaldson, W. (1968). ROC curves for discrimination of linear extent. *Journal of Experimental Psychology, 77*(3), 514-516.

D'Zmura, M. (1991). Color in visual search. *Vision Research, 31*(6), 951-966.

D'Zmura, M., & Knoblauch, K. (1998). Spectral bandwidths for the detection of color. *Vision Research, 38*(20), 3117-3128.

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance.* London: Academic Press.

Deese, J. (1955). Some problems in the theory of vigilance. *Psychological Review, 62*(5), 359-368.

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual-Attention. *Annual Review of Neuroscience, 18*, 193-222.

Detweiler, M., & Schneider, W. (1991). Modeling the acquisition of dual-task skill in a connectionist/control architecture. In D. L. Damos (Ed.), *Multiple-task performance*. Basingstoke: Burgess Science.

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: the psychophysics of reentrant visual pathways. *Journal of Experimental Psychology: General, 129*, 481-507.

Di Lollo, V., Kawahara, J., Zuvic, S. M., & Visser, T. A. (2001). The preattentive emperor has no clothes: a dynamic redressing. *Journal of Experimental Psychology: General, 130*(3), 479-492.

Dror, I. E., Busemeyer, J. R., & Basola, B. (1999). Decision making under time pressure: An independent test of sequential sampling models. *Memory & Cognition, 27*(4), 713-725.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433-458.

Dykes, J. R., & Pascal, V. (1981). The Effect of Stimulus Probability on the Perceptual Processing of Letters. *Journal of Experimental Psychology-Human Perception and Performance, 7*(3), 528-537.

Erickson, J. R. (1966). On learning several simultaneous probability-learning problems. *Journal of Experimental Psychology, 72*(2), 182-189.

Estes, W. K., Burke, C. J., Atkinson, R. C., & Frankmann, J. P. (1957). Probablistic discrimination learning. *Journal of Experimental Psychology, 54*, 233-239.

Fecteau, J. H., & Munoz, D. (2005). Visual salience, biases in attention, and links to action. *Journal of Cognitive Neuroscience*, 5-5.

Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: a priority map for target selection. *Trends in Cognitive Sciences, 10*(8), 382-390.

Fific, M., Townsend, J. T., & Eidels, A. (2008). Studying visual search using systems factorial methodology with target-distractor similarity as the factor. *Perception & Psychophysics, 70*(4), 583-603.

Findlay, J. M. (1997). Saccade target selection during visual search. *Vision Research, 37*(5), 617-631.

Findlay, J. M., Brown, V., & Glichrist, I. D. (2001). Saccade target selection in visual search: the effect of information from the previous fixation. *Vision Research, 41*(1), 87-95.

Fitts, P. M., Peterson, J. R., & Wolpe, G. (1963). Cognitive aspects of information processing: II. Adjustments to stimulus redundancy. *Journal of Experimental Psychology, 65*, 423-432.

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets rarely missed in correctable search. *Psychological Science, 18*, 943-947.

Fleck, M. S., & Mitroff, S. R. (2008). *Videogamers excel at finding rare targets.* Paper presented at the annual meeting of the Vision Sciences Society.

Fortune, B. D. (1979). The effects of signal probability on inspection accuracy in a microscopic inspection task: an experimental investigation. *The Academy of Management Journal, 22*(1), 118-128.

Gale, A. G., Mugglestone, M. D., Purdy, K. J., & McClumpha, A. (2000). Is airport baggage inspection just another medical image? In E. A. Krupinski (Ed.), *Proceedings of the SPIE - Medical Imaging 2000: Image Perception and Performance* (Vol. 3981, pp. 184-192): The International Society for Optical Engineering.

Gehring, W. J., Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Probability effects on stimulus evaluation and response processes. *Journal of Experimental Psychology: Human Perception and Performance, 18*(1), 198-216.

Ghylin, K. M., Drury, C. G., & Schwaninger, A. (2006). *Two-component model of security inspection: application and findings.* Paper presented at the 16th World Congress of Ergonomics, IEA.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*(6939), 534-537.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science, 18*(1), 88-94.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley.

Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., & Parasuraman, R. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Human Factors, 45*(3), 349-359.

Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., et al. (2003). Prevalence effect in a laboratory environment. *Radiology, 228*(1), 10-14.

Gur, D., Rockette, H. E., Warfel, T., Lacomis, J. M., & Fuhrman, C. R. (2003). From the laboratory to the clinic: The "prevalence effect"'. *Academic Radiology, 10*(11), 1324-1326.

Hamker, F. H. (2005). The reentry hypothesis: The putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cerebral Cortex, 15*(4), 431-447.

Han, S. H., & Kim, M. S. (2004). Visual search does not remain efficient when executive working memory is working. *Psychological Science, 15*(9), 623-628.

Harris, D. H. (2002). How to really improve airport security. *Ergonomics in Design*, 1 - 6.

Helton, W. S., Hollander, T. D., Warm, J. S., Matthews, G., Dember, W. N., Wallaart, M., et al. (2005). Signal regularity and the mindlessness model of vigilance. *British Journal of Psychology, 96*(Pt 2), 249-261.

Hendee, W. R. (1987). The perception of visual information. *RadioGraphics, 7*(6), 1213-1219.

Hofer, F., & Schwaninger, A. (2005). Using threat image projection data for assessing individual screener performance. *Safety and Security Engineering, 82*, 417-426.

Hoffman, J. E. (1979). A two-stage model of visual search. *Perception and Psychophysics, 25*, 319-327.

Hommel, B., Li, K. Z. H., & Li, S. C. (2004). Visual search across the life span. *Developmental Psychology, 40*(4), 545-558.

Horowitz, T. S. (2006). Revisiting the variable memory model of visual search. *Visual Cognition, 14*(4-8), 668-684.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*(6693), 575-577.

Horowitz, T. S., & Wolfe, J. M. (2001). Search for multiple targets: Remember the targets, forget the search. *Perception and Psychophysics, 63*(2), 272-285.

Hygge, S., & Knez, I. (2001). Effects of noise, heat and indoor lighting on cognitive performance and self-reported affect. *Journal of Environmental Psychology, 21*, 291-299.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10-12), 1489-1506.

Jerison, H. J. (1957). Performance on a simple vigilance task in noise and quiet. *The Journal of the Acoustical Society of America, 29*(11), 1163-1165.

Jones, D. M., Macken, W. J., & Murray, A. C. (1993). Disruption of visual short-term memory by changing-state auditory stimuli: the role of segmentation. *Memory and Cognition, 21*(3), 318-328.

Jones, D. M., Madden, C., & Miles, C. (1992). Privileged access by irrelevant speech to short-term memory: the role of changing state. *The Quarterly Journal of Experimental Psychology A, 44*(4), 645-669.

Joseph, J. S., Chun, M. M., & Nakayama, K. (1997). Attentional requirements in a 'preattentive' feature search task. *Nature, 387*(6635), 805-807.

Kaplan, I. T., & Carvellas, T. (1965). Scanning for multiple targets. *Perceptual and Motor Skills, 21*(1), 239-243.

Koelega, H. S., & Brinkman, J. A. (1986). Noise and vigilance: an evaluative review. *Human Factors, 28*(4), 465-481.

Koller, S. M., Hardmeier, D., Michel, S., & Schwaninger, A. (2007). Investigating training, transfer and viewpoint effects resulting from recurrent CBT of X-ray image interpretation. *Journal of Transport Security, 1*(2), 81-106.

Kristjansson, A. (2008). "I know what you did on the last trial" - a selective review of research on priming in visual search. *Frontiers in Bioscience, 13*, 1171-1181.

Krupinski, E. A. (2000). The importance of perception research in medical imaging. *Radiation Medicine, 18*(6), 329-334.

Kundel, H. L. (2006). History of research in medical image perception. *Journal of the American College of Radiology., 3*, 402-408.

Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology, 116*, 527-532.

LaBerge, D., & Tweedy, J. R. (1964). Presentation probability and choice time. *Journal of Experimental Psychology, 68*(5), 477-481.

Leonards, U., Sunaert, S., Van Hecke, P., & Orban, G. A. (2000). Attention mechanisms in visual search - An fMRI study. *Journal of Cognitive Neuroscience, 12*, 61-75.

Liu, T., Larsson, J., & Carrasco, M. (2007). Feature-based attention modulates orientation-selective responses in human visual cortex. *Neuron, 55*(2), 313-323.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences, 4*(1), 6-14.

Lleras, A., Rensink, R. A., & Enns, J. T. (2007). Consequences of display changes during interrupted visual search: rapid resumption is target-specific. *Perception and Psychophysics, 69*, 663-668.

Mackworth, J. F. (1968). Vigilance, arousal, and habituation. *Psychological Review, 75*(4), 308-322.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide.* (2nd ed.). Cambridge: Cambridge University Press.

Maljkovic, V., & Nakayama, K. (1994). Priming of Pop-Out .1. Role of Features. *Memory & Cognition, 22*(6), 657-672.

Maljkovic, V., & Nakayama, K. (1996). Priming of pop-out .2. The role of position. *Perception & Psychophysics, 58*(7), 977-991.

Maljkovic, V., & Nakayama, K. (2000). Priming of popout: III. A short-term implicit memory system beneficial for rapid target selection. *Visual Cognition, 7*(5), 571-595.

McBride, J., Leonards, U., & Gilchrist, I. D. (in press). Flexible target representations underlie repetition priming in visual search. *Visual Cognition*.

McCarley, J. S., Kramer, A. F., Boot, W. R., Peterson, M. S., Wang, R. F., & Irwin, D. E. (2006). Oculomotor behaviour in visual search for multiple targets. *Visual Cognition, 14*(4), 685 - 703.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science, 15*(5), 302-306.

McCarley, J. S., Wang, R. F., Kramer, A. F., Irwin, D. E., & Peterson, M. S. (2003). How much memory does oculomotor search have? *Psychological Science, 14*(5), 422-426.

McElree, B., & Carrasco, M. (1999). The temporal dynamics of visual search: evidence for parallel processing in feature and conjunction searches. *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1517-1539.

Menneer, T., Barrett, D. J. K., Phillips, L., Donnelly, N., & Cave, K. R. (2004). Search efficiency for multiple targets. *Cognitive Technology, 9*, 22-25.

Menneer, T., Barrett, D. J. K., Phillips, L., Donnelly, N., & Cave, K. R. (2007). Costs in searching for two targets: Dividing search across target types could improve airport security screening. *Applied Cognitive Psychology, 21*, 915-932.

Menneer, T., Cave, K. R., & Donnelly, N. (under review). The cost of search for multiple targets: the effects of practice and target similarity. *Journal of Experimental Psychology: Applied*.

Michel, S., Koller, S. M., Ruh, M., & Schwaninger, A. (2007). Do "image enhancement" functions really enhance x-ray image interpretation? In S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1301-1306). Austin, TX: Cognitive Science Society.

Miller, J., & Bauer, D. W. (1981). Visual similarity and discrimination demands. *Journal of Experimental Psychology-General, 110*(1), 39-55.

Miller, J., & Pachella, R. G. (1973). Locus of the stimulus probability effect. *Journal of Experimental Psychology, 101*(2), 227-231.

Motter, B. C., & Belky, E. J. (1998a). The guidance of eye movements during active visual search. *Vision Research, 38*(12), 1805-1815.

Motter, B. C., & Belky, E. J. (1998b). The zone of focal attention during active visual search. *Vision Research, 38*(7), 1007-1022.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*(3), 465-494.

Myers, J. L., & Cruse, D. (1968). Two-choice discrimination learning as a function of stimulus and event probabilities. *Journal of Experimental Psychology, 77*(3), 453-459.

Nakayama, K., & Joseph, J. S. (1998). Attention, pattern recognition and pop-out in visual search. In R. Parasuraman (Ed.), *The Attentive Brain* (pp. 279-298). Cambridge, MA: MIT Press.

Navalpakkam, V., & Itti, L. (2002). A goal oriented attention guidance model. *Biologically Motivated Computer Vision, Proceedings, 2525*, 453-461.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53*(4), 605-617.

Neisser, U., Novick, R., & Lazar, R. (1963). Searching for ten targets simultaneously. *Perceptual and Motor Skills, 17*, 955-961.

Nodine, C. F., & Kundel, H. L. (1987). Using eye movements to study visual search and to improve tumor detection. *RadioGraphics, 7*(6), 1241-1250.

Nosofsky, R. M., & Palmeri, T. J. (1997a). Comparing exemplar-retrieval and decision-bound models of speeded perceptual classification. *Perception and Psychophysics, 59*(7), 1027-1048.

Nosofsky, R. M., & Palmeri, T. J. (1997b). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*(2), 266-300.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience, 5*(11), 1203-1209.

Oh, S. H., & Kim, M. S. (2004). The role of spatial working memory in visual search efficiency. *Psychonomic Bulletin And Review, 11*(2), 275-281.

Olds, E. S., Cowan, W. B., & Jolicoeur, P. (2000). The time-course of pop-out search. *Vision Research, 40*, 891-912.

Olds, E. S., Jolicoeur, P., & Cowan, W. B. (2001). Interactions between search mechanisms in conjunction search. *Canadian Journal of Experimental Psychology, 55*(4), 285-295.

Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science, 205*(4409), 924-927.

Pashler, H. (1987). Target-distractor discriminability in visual search. *Perception and Psychophysics, 41*(4), 385-392.

Peterson, M. S., Kramer, A. F., Wang, R. F., Irwin, D. E., & McCarley, J. S. (2001). Visual search has memory. *Psychological Science, 12*(4), 287-292.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology, 53*(3), 195-237.

Rayner, K. (1978). Eye-movements in reading and information-processing. *Psychological Bulletin, 85*(3), 618-660.

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), 119-137.

Rich, A. N., Hidalgo-Sotelo, B., Kunar, M. A., van Wert, M. J., & Wolfe, J. M. (2006). *What happens during search for rare targets? Eye movements in low prevalence visual search*. Paper presented at the Vision Sciences Society.

Riesenhuber, M. (2004). An action video game modifies visual processing. *Trends in Neurosciences, 27*(2), 72-74.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I: Detection, search, and attention. *Psychological Review, 84*, 1-66.

Schwaninger, A. (2004). *Increasing efficiency in airport security screening.* Paper presented at the AVSEC World 2004, Vancouver, B.C., Canada.

Schwaninger, A. (2006). Liquid identification: Reacting to the terror threat. Analysing, controlling and adapting to meet new threats. *Airport, 5*, 30-31.

Schwaninger, A., Bolfing, A., Halbherr, T., Helman, S., Belyavin, A., & Hay, L. (2008, June 1-4). *The impact of image based factors and training on threat detection performance in X-ray screening.* Paper presented at the 3rd International Conference on Research in Air Transportation, ICRAT, Fairfax, Virginia, USA.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems, 20*(6), 29-35.

Schwaninger, A., & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In K. Morgan & M. J. Spector (Eds.), *The Internet Society 2004, Advances in Learning, Commerce and Security* (pp. 147-156). Wessex: WIT Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127-190.

Simpson, W., & Voss, J. F. (1961). Psychophysical judgements of probablistic stimulus sequences. *Journal of Experimental Psychology, 62*(4), 416-422.

Smith, J. D., Redford, J. S., Gent, L. C., & Washburn, D. A. (2005). Visual search and the collapse of categorization. *Journal of Experimental Psychology: General, 134*(4), 443-460.

Smith, J. D., Redford, J. S., Washburn, D. A., & Taglialatela, L. A. (2005). Specific-token effects in screening tasks: possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(6), 1171-1185.

Sperling, G., Budiansky, J., Spivak, J. G., & Johnson, M. C. (1971). Extremely rapid visual search: the maximum rate of scanning letters for the presence of a numeral. *Science, 174*(6), 307-311.

Sperling, G., & Melchner, M. J. (1978). The attention operating characteristic: examples from visual search. *Science, 202*(4365), 315-318.

Spitz, G., & Drury, C. G. (1978). Inspection of sheet materials – test of model predictions. *Human Factors, 20*(5), 521-528.

Stevens, W. D., Hasher, L., Chiew, K. S., & Grady, C. L. (2008). A neural mechanism underlying memory failure in older adults. *The Journal of Neuroscience, 28*(48), 12820-12824.

Stroud, M., Menneer, T., Cave, K. R., Donnelly, N., & Rayner, K. (in preparation). The breakdown of color selectivity in search for multiple targets: Evidence from eye movements. *Manuscript in preparation for publication*.

Taylor, W., Melloy, B., Dharwada, P., Gramopadhye, A., & Toler, J. (2004). The effects of static multiple sources of noise on the visual search component of human inspection. *International Journal of Industrial Ergonomics, 34*, 195-207.

Theeuwes, J., Itti, L., Fecteau, J., & Yantis, S. (2005). Top-down and bottom-up control of visual selection. *Journal of Cognitive Neuroscience*, 4-5.

Thornton, T. L., & Gilden, D. L. (2007). Parallel and serial processes in visual search. *Psychological Review, 114*(1), 71-103.

Townsend, J. T. (1972). Some Results Concerning Identifiability of Parallel and Serial Processes. *British Journal of Mathematical & Statistical Psychology, 25*(Nov), 168-199.

Townsend, J. T. (1990). Serial Vs Parallel Processing - Sometimes They Look Like Tweedledum and Tweedledee but They Can (and Should) Be Distinguished. *Psychological Science, 1*(1), 46-54.

Townsend, J. T., & Colonius, H. (1997). Parallel processing response times and experimental determination of the stopping rule. *Journal of Mathematical Psychology, 41*(4), 392-397.

Townsend, J. T., & Wenger, M. J. (2004a). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review, 11*(3), 391-418.

Townsend, J. T., & Wenger, M. J. (2004b). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111*(4), 1003-1035.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London-Series B-Biological Sciences, 353*(1373), 1295-1306.

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

Tsotsos, J. K. (1997). Limited Capacity of Any Realizable Perceptual System Is a Sufficient Reason for Attentive Behavior. *Consciousness and Cognition, 6*, 429–436.

van der Heijden, A. H. C., & Bem, S. (1997). Successive Approximations to an Adequate Model of Attention. *Consciousness and Cognition, 6*, 413-428.

van Zoest, W., Lleras, A., Kingstone, A., & Enns, J. T. (in press). In sight, out of mind: the role of eye movements in the rapid resumption of visual search. *Perception and Psychophysics*.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A. *Perception and Psychophysics, 68*(4), 643-654.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*: Oxford University Press.

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review, 1*(2), 202-238.

Wolfe, J. M. (1998). What can 1,000,000 trials tell us about visual search? *Psychological Science, 9*(1), 33-39.

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. (2007). Guided Search 4.0: Current Progress with a model of visual search. In W. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99-119). New York: Oxford.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 419-433.

Wolfe, J. M., Friedman-Hill, S. R., & Bilsky, A. B. (1994). Parallel processing of part-whole information in visual search tasks. *Perception and Psychophysics, 55*(5), 537-550.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*(6), 495-501.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*, 439-440.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General.*

Woodman, G. F., & Luck, S. J. (2004). Visual search is slowed when visuospatial working memory is occupied. *Psychonomic Bulletin and Review, 11*(2), 269-274.

Yantis, S. (2000). Goal-directed and stimulus-driven determinants of attentional control. In S. Monsell & J. Driver (Eds.), *Attention and Performance* (Vol. 18, pp. 73-103). Cambridge, MA: MIT Press.

Yantis, S. (2005). How visual salience wins the battle for awareness. *Nature Neuroscience, 8*(8), 975-977.

Zelinsky, G. J., & Sheinberg, D. L. (1997). Eye movements during parallel-serial visual search. *Journal of Experimental Psychology-Human Perception and Performance, 23*(1), 244-262.