RICE UNIVERSITY

Refining Theoretical Models of Visual Sampling in Supervisory Control Tasks: Examining the Influence of Alarm Frequency, Effort, Value, and Salience.

by

Michael D. Fleetwood

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

_____	_____
Michael D. Byrne, Assistant Professor, Chair Psychology	James R. Pomerantz, Professor Psychology
_____	_____
David M. Lane, Associate Professor Psychology, Statistics, and Management	Alex C. Kirlik, Associate Professor Engineering, Psychology, University of Illinois Urbana-Champaign
_____	
Devika Subramanian, Professor Computer Science	

HOUSTON, TEXAS

MAY 2005

ABSTRACT

Refining Theoretical Models of Visual Sampling in Supervisory Control Tasks: Examining the Influence of Alarm Frequency, Effort, Value, and Salience.

by

Michael Fleetwood

This work is concerned with examining in a formal quantitative manner what human observers look at and what the objects of their gaze tell them. Three models designed to describe and predict the allocation of human attention in supervisory control tasks were investigated. A series of three experiments examined the relative influence of five factors on the sampling patterns of participants: the information generation rate of the information signal (bandwidth), the frequency of significant, i.e., task relevant, events on an information source (alarm frequency), the payoff matrix associated with missing or detecting critical events (value), the visual salience of the events, and the cost of making an observation. The paradigm employed is similar to that developed by Senders and colleagues (1964), in which observers were asked to monitor an array of four simulated ammeters and to press a button whenever the pointer of any ammeter entered an "alarm zone." Aspects of three mathematical models, Senders's constrained random sampler, Wickens and colleagues SEEV model, and Pirolli's and Card's Information Foraging Theory Model, were combined to form seven different models predicting performance in the task. The sampling patterns predicted by each model were compared against the eye movement data of participants. Results of the three experiments indicate that participants'

sampling patterns were sensitive to the experimental manipulations. Comparisons of the model predicted patterns of attention allocation to those in the participant data indicated that different models described different participants. Participants who performed poorly at the task were best described by models incorporating bandwidth. Participants who performed well at the task were best described by models incorporating alarm frequency, and those who performed best at the task were not well-described by any of the models. Overall the models based on Information Foraging Theory were the most robust in predicting the attention allocation patterns of participants. Implications of each of the experimental manipulations and of the fit of the models to the participant data are discussed.

Acknowledgements

This work was funded in part by the National Aeronautics and Space Administration under grant number NDD2-1321. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NASA, the U.S. Government, or any other organization.

Table of Contents

List of Tables and Figures

1. Introduction

For humans, vision is a primary method for gathering information from the outside environment. However, information cannot be gathered equally well from all areas within the visual field. Due to functional and physical limitations of the eye and human attention, we are most efficient at gathering information from the center of the visual field. The visual field may be as wide as 180 degrees, but one does not see well more than two or three degrees away from the line of sight. As a result, humans must move their eyes and/or their head in order to examine whatever is to be examined in detail.

As a result of this limitation, the human eye is constantly in motion. Being a principal method for acquiring information, if we are to understand in some detail the deeper levels of human cognition responsible for manipulating any externally acquired information, we must develop an understanding of how we acquire that information. More specifically, we would like to know where the eye looks and why, how long it looks there and why, what is there to be seen, and where it looks next and why. In short, we would like to know how human visual attention is allocated. This work is concerned with examining in a formal quantitative manner what human observers look at and what the objects of their gaze tell them.

Researchers have developed several models designed to describe and predict the allocation of human visual attention. With respect to visual attention, each of these models falls into one of two categories: models of visual search or models of visual sampling. Models of visual search are concerned with how people locate an object in the visual environment (e.g., Deco & Lee, 2002; Brogan, Gale, & Carr, 1993; Wolfe, 1994;

Neisser, Novick, & Lazar, 1964; Teichner & Mocharnuk, 1979). Many of these models have centered on contrasting parallel and serial search in basic laboratory paradigms. A number of models were designed to predict the allocation of visual attention in applied contexts, such as when examining graphs (Gillian & Lewis, 1994; Lohse, 1993), maps (Wickens, Kroft, & Yeh, 2000), menus (Fisher & Tan, 1989; Norman, 1991; Hornof & Kieras, 1997; Byrne, 2001), icons (Fleetwood & Byrne, 2002), and roadway environments (Theeuwes, 1994). A key facet of such models is the emphasis on locating a single target, and an emphasis on search time as the critical dependent variable.

In contrast to search models, another class of models has focused on visual sampling or monitoring behavior in supervisory control tasks. It is these models, models of visual sampling, that I am primarily concerned with here. These models use sampling or scanning as a dependent variable (e.g., Moray, 1986; Senders, 1964, 1983; Carbonnell, Ward, & Senders, 1968; Ellis & Stark, 1986; Sheridan, 1970). Such models have typically focused on the eye as a "single server queue." Four key facets distinguish these models from models of visual search (Wickens, et al., 2001). (1) The operator is not looking for a static target, but is rather supervising a series of dynamic processes, such as temperature gauges or aircraft movements. (2) The primary focus of the models is on noticing critical events at relatively consistent locations rather than finding critical targets at uncertain locations. (3) The key dependent variable is not target detection response time, but is instead the proportion of visual attention distributed to various "areas of interest" (AOIs) as a function of the quantitative properties of those AOIs. (4) The process of defining specified AOIs means that the challenge in visual attention is not so much knowing where to look, but in knowing when to look where.

Outside of the domain of visual attention, a number of models have been developed that attempt to predict the general allocation of attention. These models are not necessarily concerned with visual attention per se, but they are often applied on a coarser grain of analysis. For example, they may predict the amount of time (as a measurement of attention) a person might devote to different tasks in an office (Pirolli & Card, 1999), or the amount of time a person may devote to different web pages when searching for a piece of information (Pirolli & Fu, 2003). Similarly, they may predict the course of action a person will be likely to take given a set of alternatives.

In general, models of visual sampling and general models of attention are attempting to do the same thing: predict attention allocation given a set of constraints and some information regarding the environment and the task. Each model brings to the table its own set of advantages and disadvantages as a predictor of human performance. It is the basic premise of the research described herein that many of the models developed to predict attention allocation in one domain (visual or general) will not only function as a valuable model in the other domain but under certain conditions will make predictions that go beyond those of the models originally developed in that domain. Specifically, I will examine the ability of general attention models to predict attention allocation in the visual sampling domain.

Despite the differing levels of task structure that these models are concerned with, there is reason to believe that either class of models may be applicable to the other class. In particular, there is evidence that the models of general attention may have application at the level of visual attention. Foremost, all of the models are attempting to address a similar issue in a similar manner. Given a set of options (whether those options are

which task to devote time to or which instrument on a control panel to attend to) and a set of constraints (such as the time needed to make a shift of attention or the time needed to complete a given task), the models attempt to determine which option will be chosen in order to achieve a given goal. In essence, the models attempt to optimize the use of a limited resource, attention. Second, other models from the decision-making literature have proven to be extremely useful in the realm of predicting visual attention. Optimal Estimation Theory (OET) and Optimal Control Theory (OCT) have been applied to the analysis of the human visual attention quite successfully (Baron, Kleinman, and Levison, 1970; Kleinman, Baron, & Levinsion, 1970; Kok & van Wijk, 1978; Stassen, 1978; Baron & Levinson, 1973, 1975, 1977; Curry, Kleinman, & Hoffman, 1977). Lastly, there is some mathematical similarity between the models in both domains (as will be discussed subsequently); although the differences may prove crucial in the transition from one domain to another.

Of course, all of the models assume that people are optimal, or near optimal, at attention allocation. There is ample evidence that people are not optimal decision makers; however, there is also evidence that suggests that people exhibit near optimal behavior for tasks in which they are highly practiced (Senders, 1964; Carbonell, 1966). Fortunately, the tasks that we are primarily concerned with are tasks in which people are generally highly practiced, such as flying an airplane or monitoring a power plant. (We will return to the question of whether people are optimal attention allocators later.)

1.1 The Current Research

There are several aims of the current research project. First, we are concerned with the comparison of a number of models. All of the models have been designed to predict attention in a particular domain and given a particular task, and we would like to see how their attention allocation predictions match up in a common task. There are two primary comparisons that will be made. One, how do the model predictions differ from each other? More specifically, we will examine how models designed to predict the allocation of attention on a coarser grain of analysis fare on a perceptual level. Do some models make predictions that others do not? For instance, some models make moment-to-moment or even eye-movement-to-eye-movement predictions where as others are only capable of making predictions on a more aggregate level.

The second objective of the current research is to make some refinements to the models where possible. One of the overarching goals of the line of research dedicated to modeling visual attention is to determine the factors that influence when and where a person will look next. Moray (1986) identified a number of factors, which under the right conditions, will influence the monitoring strategy of observers: the rate at which exogenous uncertainty is generated by the monitored process; the rate at which endogenous uncertainty is generated by forgetting; the accuracy with which the observer must read the value of a function; the threshold below which the observer is indifferent to the magnitude of the signal; the magnitude of the limit that the process must not exceed; the probability that while viewing one source another may show a critical value; the payoff matrix associated with missed critical events; the cost of making an observation; the overall strategy chosen by the observer to maximize the value of the behavior; and the

correlation between information sources. In the scenario that we will be using, we would like to determine the extent to which the above factors contribute to the attention allocation patterns of observers and incorporate such factors into the models accordingly.

Because of the number and complexity of the computational models in this domain, only a subset deemed to be representative will be examined here. Each will be described in some detail. From the visual sampling domain, the seminal model of visual sampling, developed by Senders (1964, 1983) is examined because of its influence in the domain and because of its power in predicting human performance. A more recent model, the SEEV model developed by Wickens and colleagues (Wickens, Goh, Helleberg, Horrey, & Talleur, 2003) is a relatively simple model that has been shown to be an accurate predictor of visual attention by airline pilots, and it is also examined herein. Regarding general models of attention, the Information Foraging model (Pirolli & Card, 1999) has been evaluated as a model of general attention in a wide variety of contexts, and will be examined as a model of visual attention here.

Before we introduce the four models in some detail, we will cover some of the seminal research already conducted in this field. The following section of this document, Section 2, addresses the reasons behind the importance of developing computational models of visual sampling behavior and the research of Fitts and his colleagues that ignited the interest in the domain. In Section 3, we will begin to provide some detail on the four models we will be working with, Senders' queuing model, Wickens' SEEV model (Section 4), and Pirolli's Information Foraging model (Section 5). For each model, we cover the issue or problem that each of the models was developed to deal with, the domains that the model has traditionally been applied to, the computational components

of the model, and the reasons for its application in the current work. Section 6 describes the basic paradigm and methodology we employed. It is an adapted version of the paradigm developed by Senders (1964, 1983) in his seminal work in the field. Section 7 introduces the methods by which we recorded and analyzed the data from the experiments. Three experiments and their results are presented in sections 8, 9, and 10. Finally, in Section 11, the general applications of the research are discussed.

2. Model Applications

The development of accurate predictive models of human visual sampling behavior provide a theory on which to base design decisions. For instance, it is unlikely that designers of complex interfaces, such as an airplane cockpit or the control room of a nuclear power plant, have any principled manner of determining how many of them can be accurately monitored by a single observer and similarly how many observers are needed to adequately monitor an entire system. Due to the potentially catastrophic consequences of inadequate monitoring of a system in such critical domains, the importance of never underestimating these quantities is obvious. Using models of visual sampling behavior, we can calculate the workload imposed on human beings in complex monitoring and supervisory control tasks. As the number of information sources that must be sampled exceeds some threshold of error, a user will not be able to sample the sources with sufficient frequency. We can calculate the maximum number of instruments that a single observer can adequately monitor and likewise, the minimum number of observers needed to adequately monitor a system.

Similarly, a theory of when and where people look may be employed to develop better interfaces. The most efficient placement of instruments on a display and the

presentation of the appropriate information at the appropriate time can be quantitatively examined using an adequate theory of visual sampling. This principle does not only apply to airplane cockpits and power plant control rooms, although those are two prominent examples in the literature. The same principles of interface design and the timely presentation of information apply to nearly all displays, such as the design of desktop computer software, web page design, and in-car navigation systems.

Eventually, a model of visual sampling behavior can and should be integrated into larger, more complex models of human behavior. For example, much of the research on visual sampling has been carried out in the context of airplane pilots and cockpit design. However, when and where pilots look is not solely dependent on the statistical properties of the instruments they are observing. When and where pilots look is integrated into a complex task structure. A pilot's goals and hence his or her information needs and visual sampling patterns are constantly changing over the course of the flight. Among the factors that influence them are the stage of flight (take-off, cruising, approach, landing, etc.), communication with the flight crew, passengers, and airport tower, completing standardized procedures, etc. In other words, the real-world environments and tasks of pilots are much more complex than the laboratory environment in which the visual sampling models have been developed. In order to predict human performance in these more complex environments, the models of visual sampling must be merged with a broader model of pilot performance.

Although not the focus of this manuscript, the original motivation for the research project presented here, and the ultimate goal of the project is to integrate a model of visual sampling behavior into a more complex cognitive model.

2.1 The Seminal Data

Milton, Jones, and Fitts published a series of technical reports in the late 1940s that are considered to be the seminal research on visual sampling and represent the largest collection of eye movement data collected in a visual monitoring task. A summary was published in the Aeronautical Engineering Review in 1950 (Fitts, Jones, Milton, 1950). The data encompass over 500,000 frames of movie film of over 40 pilots taken under various flight conditions. The work continued for a number of years; further studies on more modern aircraft were undertaken as late as 1966.

The data collection paradigm consisted of a pilot flying in a simulator through various flight scenarios and performing a variety of maneuvers in a variety of flight conditions. A camera was placed in cockpit such that it could record the face of pilot. The eye movements of the pilot were calibrated so that researchers could determine which instrument the pilot was observing at any point in time.

The general conclusions of the research were eloquently addressed by the study's authors (Fitts, Jones, & Milton, 1950): "It is reasonable to assume that *frequency* of eye fixations is an indication of the relative *importance* of that instrument. The *length* of fixations, on the contrary, may more properly be considered as an indication of the relative *difficulty* of checking and interpreting particular instruments. A *pattern* of eye movements, i.e., the Link Values[1] between instruments, is a direct indication of the

[1] In order to examine the patterns fixations in their experiments, Milton, Jones, and Fitts examined the probability that the participants would transition from looking at any one

goodness of different panel *arrangements*." (Emphasis by original authors.) Further, "Information about how pilots use their eyes while flying on instruments is fundamental to a basic understanding of the functions served by aircraft instruments and to simplification of the psychological processes that occur while a pilot is controlling an aircraft's altitude, location, and rates of movements in three-dimensional space. If we know where a pilot is looking, we do not necessarily know what he is thinking, but we know something of what he is thinking about."

The conclusions of their work were to be applied directly to the cockpit design. The research questions focused on how pilots monitor flight instruments. For instance, they were interested in whether pilots tended to view multiple instruments in a single glance or if they directed their attention to separate instruments sequentially, whether the frequency with which instruments were observed differed from instrument to instrument, and whether the time to check an instrument differed from instrument to instrument. The goal of the research was to design better instruments and better systems of arranging them. As stated in Fitts, Jones, and Milton (1950), "The underlying theme for the conclusions is about as follows: Those instruments at which a pilot looks often should be located in the central area of the instrument panel; Those instruments which are looked at for a long time are either very important and should be centrally located and possibly enlarged, or are difficult to read and should be redesigned; those pairs of instruments which are looked at often in succession (high link values) should be located close to one another."

---

display to any other display. These probabilities are termed the "Link Values" between any two displays.

Results of the studies were statistics regarding the duration, frequency, and sequence of fixations on the various instruments. Data for each flight condition was summarized in charts, such as that presented in Figure 2.1. The numbers below each instrument represent the frequency of fixation, duration of fixation, and percentage of total time spent on that instrument. The number on the arc connecting two instruments is the percentage of all transitions made by subjects between those two instruments. The

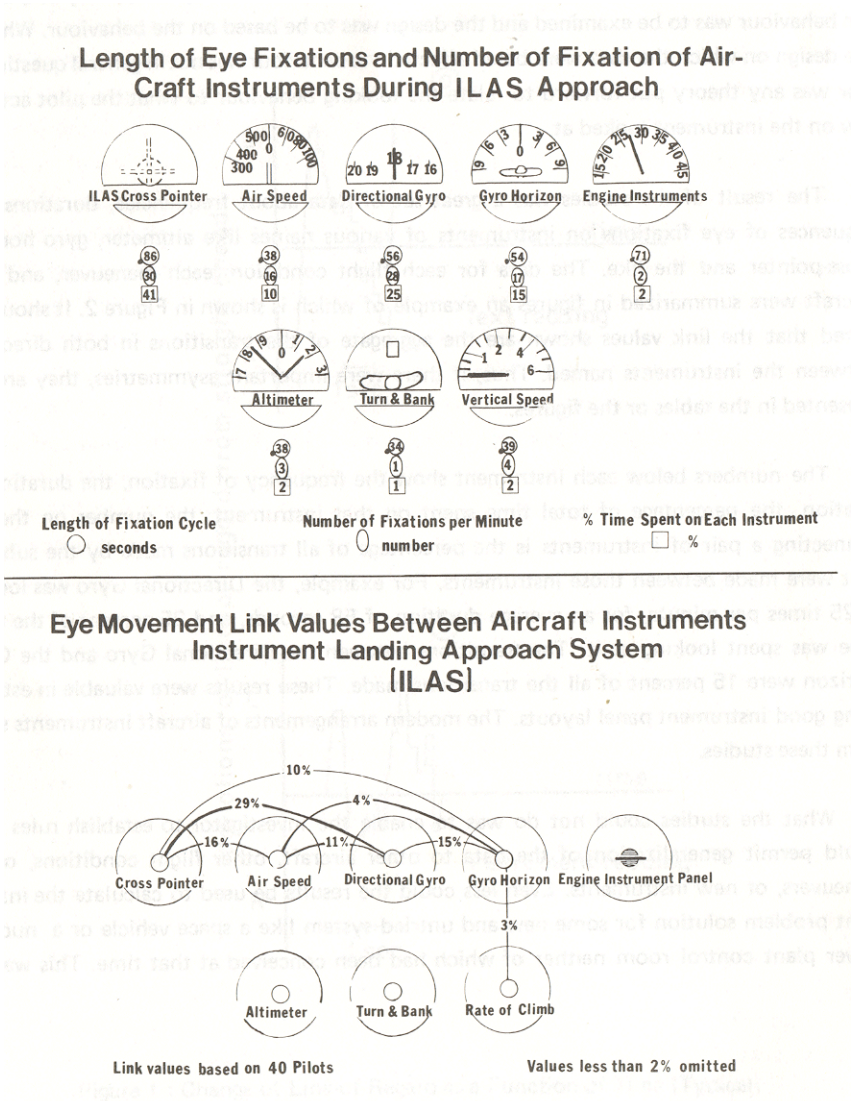


Figure 2.1. Diagram from Fitts et al. (1950) summarizing the fixation data.

link values represent the total transitions between each pair of instruments, so asymmetries, if there were any, are not presented. So for example, the air speed indicator was looked at 10% of the time. Approximately 16 fixations were directed to the instrument per minute, and each time the instrument was looked at, subjects looked at it for an average of 0.38 seconds. The transitions between the air speed indicator and the directional gyro accounted for approximately 11% of all transitions between instruments. The value of the results lay in establishing good instrument panel layouts. Indeed, modern arrangements of aircraft instruments stem from these studies (Senders, 1983).

Fitts, Jones, and Milton (1950) were able to propose a more efficient arrangement of instruments and identify those which are difficult to read, for a possible redesign of the actual instrument. However, it is important to note that no theory was put forth regarding why a pilot needed to look at any particular instrument at a particular time, nor did they consider what the pilot actually saw on the instrument. Without a theory of when and where pilots were looking, the results cannot be generalized to other aircraft, flight conditions, pilot maneuvers, or the addition of new instruments. There is even less value in extending the results to an entirely new system, such as a power plant, a space vehicle, radar displays, or submarine instrument panels. Every system developed would require a new empirical study in order to determine the arrangement of instruments. And as each new system were developed, there would be no principle to guide the initial arrangement of instruments. Further, any arrangement would be based on the assumption that a new arrangement would not substantially change the behavior of observers.

In order to make these generalizations, one must abstract the rules or laws governing the pilots' behavior. In order to develop these rules, one must know something about the information that each instrument displays. There must be a record not only where the pilot was looking, but what he or she was looking at. No data were recorded by Fitts and company regarding what information the instrument presented to the observer as he or she looked at it. A complete model of the behavior rests on developing an understanding of the link between the where the pilot is looking and what he is looking at. The goal of creating analytical models of visual sampling behavior is to determine what the statistical properties of the information presented to observers are that drive their monitoring behavior and to link them in a formal, quantitative manner. In turn, the models of visual sampling can then guide the design of complex systems.

3. Senders' Model

The first model that attempted to relate scanning behavior to the statistical properties of instruments was developed by Senders (Senders, 1955; Senders, Elkind, Grignetti, & Smallwood, 1964) and represents the earliest attempt to derive quantitative analytic models of visual sampling behavior. It was developed in light of the work by Fitts' group on the eye movements of pilots in cockpits. Senders has more recently reexamined the work, simplifying and extending the original model (Senders, 1983).

The primary assumption of the original model is that the task of the observer is to "reconstruct" the signal as best as possible from noncontiguous observations of the signal. In other words, the observer is attempting, in a minimum number of observations, to extract enough information from the signal such that the information would be

necessary and sufficient to make a copy of the displayed function on the basis of the sampled values.

Central to the mathematical formulation of the model is the Nyquist theorem, which can be elucidated by example. Let the observer monitor a random signal with limited bandwidth, W Hz. It can be shown that if the bandwidth is W Hz it is necessary and sufficient that it be sampled at $2W$ Hz, and that an ideal channel can reconstruct the signal from observations taken at intervals spaced $1/2W$ seconds apart. This is known as the Nyquist theorem, or sampling theorem, and the sampling interval is often referred to as the Nyquist interval. A human observer who behaved like such a channel would sample a signal $2W$ times per second.

The second equation that is central to the behavioral model is the calculation of the information generation rate of the instrument. This formula predicts the expected dwell time or fixation duration of a sample. For our ideal observer, which samples at a rate of $2W$ times per second, the average amount of information that the observer must assimilate at each sampling, H_i, is given by:

$$H_i = log_2 (A_i / E_i) \qquad (3.1)$$

Bits per second, where A is the root mean square (RMS) amplitude of the signal and E is the RMS error that is tolerable in the observation (i.e., the required precision). If we assume that the observer has a fixed and limited input capacity, the time required for a sample, D_i, is given by:

$$D_i = K log_2 (A_i / E_i) + C \qquad (3.2)$$

Where K has the dimension time per bit and C is a constant, which accounts for movement time and minimum fixation time (in seconds per sample).

Senders used the following experimental paradigm to test his model. Four ammeters were mounted on a board at the four corners of a square, and a movie camera was mounted in the middle of the array so that the eye movements of an operator could be filmed when the latter sat in front of the array. The meters were separated by about 6° of visual angle. The observer was told to monitor the array and to press a button whenever a pointer exceeded a specified limit, i.e., when the pointer entered a "danger zone." The meters were driven by quasi-random forcing functions such that their movements appeared random to observers. Each meter had a different bandwidth, from less than 0.05 Hz to about 0.6 Hz. Thus each instrument was characterized by its bandwidth W_i Hz for the *i*th instrument, and each should have been sampled at a rate of $2W_i$ fixations per second. It is important to note that the total average rate of fixations was chosen so that it would not overload the observer, but would on the average almost fully occupy his or her time. The results obtained by Senders et al. (1964) are presented in Figure 3.1. Data are from five highly practiced observers. Each participant spent in excess of 30 hours practicing the task.

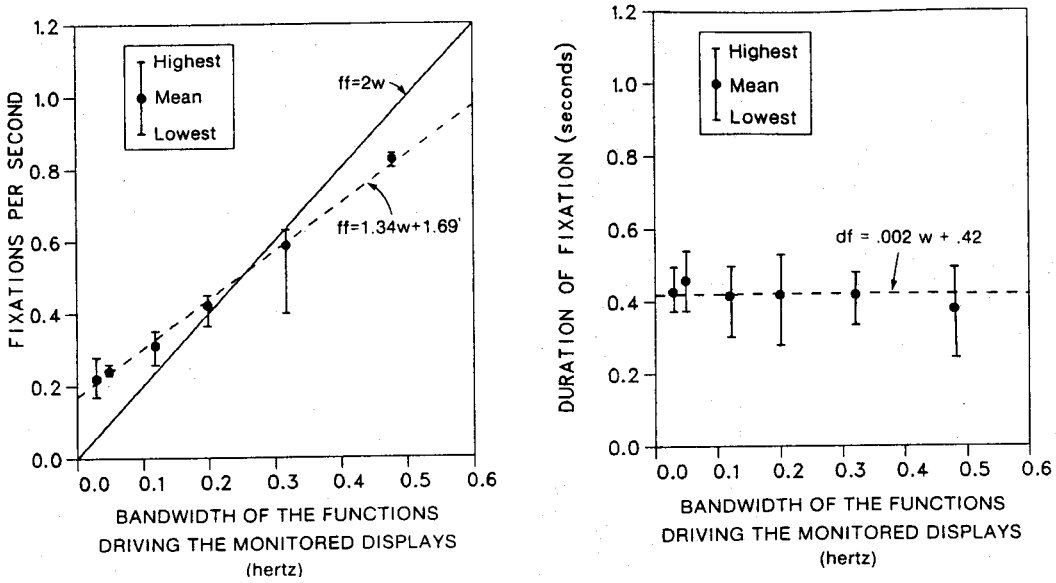


Figure 3.1. Result of Senders, et al. (1964) adapted from Moray (1986). Relations between fixation frequency and bandwidth of the observed signal (left) and between fixation duration and bandwidth of the observed signal (right) support Sender's theory. The fixation duration at each bandwidth should be approximately 0.4 seconds since the ration of the signal power to the required accuracy of observation was identical at all bandwidths.

Although the fit of the predicted values to the data is not perfect (the slope is approximately 1.34 rather than the predicted 2.0), the obtained data strongly fulfill the general predictions of the model. The function described by the data points is linear and monotonic and increases with bandwidth.

Observers had a tendency to undersample at high bandwidths. Undersampling may have been due to two causes, cases where repeated samples are taken from the same instrument without moving the eyes and cases where the position and velocity of the signal could be determined very quickly. Fogel (1956) showed that if the observer could detect both instantaneous position and instantaneous velocity when making an observation, then the appropriate sampling rate would be W Hz, rather than $2W$ Hz. The data were reanalyzed by Senders (1983) with the inclusion of these two aspects of

undersampling, and they appear to account for much of the undersampling in a systematic manner.

Observers also had a tendency to oversample at low bandwidths. As discussed by Moray (1981) and Senders (1983), oversampling may have been due to forgetting. As the Nyquist interval becomes longer and longer with decreasing bandwidths, observers may simply forget the value of the signal and need to resample not to due uncertainty produced exogenously by the varying amplitude of the signal, but by uncertainty generated endogenously by the observer. Experimental results from Moray, Richards, and Low (1980) confirm that the sampling rate predicted by the model at low bandwidths may exceed the short-term memory capacity of participants engaged in the task.

The measured fixation durations or dwell times were also in agreement with the predictions of the model. The ratio A/E was kept identical for all instruments in the experiment; hence, Equation 3.2 predicts that the fixation durations should also be identical and independent of the signal bandwidth, as was confirmed by the data (Figure 2, right).

Senders (1966) also considered the transition probabilities of fixations as the eyes moved from one instrument to another. Given that the observer is currently fixating the instrument i, what is the probability that the next instrument observed will be j? Over the long run, each instrument will receive a proportion of the total number of fixations. Suppose that following a fixation the next instrument is simply chosen with a probability equal to the proportion of the fixations it obtains overall, subject only to the necessity to satisfy the intervals defined by the sampling theorem. Since some transitions will be

made to the instrument that is already being fixated, a correction is needed, and the final equation becomes

$$P_{oab} = (2P_aP_b) / (1- \Sigma P^2_i) \qquad (3.3)$$

Where P_{oab} is the probability of an observable transition between instruments a and b, and P_a , P_b are the proportion of fixations made on instruments *a* and *b*. When the predicted probabilities of transitions were compared with the observed probabilities the model was shown to capture the general trends in the data.

The work of Senders (1964) is impressive and is rightly regarded as classic in the field of monitoring. It appears that Fitts et al. (1950) may not be entirely correct in their assumption that fixation frequency is primarily determined by the importance of the instrument, since signal bandwidth is so strong a determinant of behavior. Fitts' idea that display quality determines fixation time is supported by the results of Senders et al., to the extent a poorly designed instrument will frequently have a "noisy" display, be harder to read as a result, and have a higher information generation rate due to the signal/ noise ratio. This will require a longer fixation according to the model of Senders. Senders (1966) reanalyzed some of Fitts' data and showed a close relationship between his model and the data recorded by Fitts' group in real flights.

3.1 Optimal Models of Behavior

It should be explicitly noted that the Senders models are based on the concept that observers sample according to what is mathematically optimal. That is, the models define the patterns of fixations that human participants would show if they were sampling the environment in a mathematically optimal manner. The fact that the model predicted data

matches that of the human participants to a striking degree provides strong evidence that people are indeed sampling in a nearly optimal manner.

Interestingly, human behavior has been shown to be largely divergent of optimality in a variety of other domains, most notably decision making and reasoning regarding statistical information. There are a large number of studies documenting the inefficiency of humans as statistical decision makers (e.g Vaughn & Mavor, 1972; Kahneman & Tversky, 1973; Tversky & Kahneman, 1971, 1974; Kahneman, Slovic, and Tversky, 1982). For instance, Vaughn and Mavor (1972, p. 274) eloquently note that man "… is slow to initiate action and conservative in his estimates of highly probable situations, and when he does act or accept a diagnosis, he is reluctant to change an established plan or a situation estimate when the available data indicate that he should. He is generally a poor diagnostician and does not learn by mere exposure to complex tasks but only when specific relationships that make up the complexity are explained."

Studies from the decision-making domain would predict that humans would not be nearly optimal samplers of the visual environment, yet Senders's data indicate that they are. Much of this discrepancy can be traced to degree of conscious involvement the task requires of the human participant. As noted by Moray (1986), "it is as a conscious decision maker, reflecting and working in a knowledge-based mode of operation, that the human fails. As a highly practiced, automatic operator, using skills that require no conscious reflection, he or she becomes optimal. Skill-based behavior is far more optimal than problem-solving behavior." Rasmussen (1983) makes a similar distinction between "skill-based" behavior (automatic) and "knowledge-based" behavior (conscious problem

solving). In order to see optimal behavior, prolonged practice to the point where behavior is automatic is required. As Senders (1983) put it, “No novices need apply.”

The observers in the experiments of Senders (1964) were far from novices at the task, as they received 20-30 hours of practice before data were collected. They were not told the relative bandwidths of the signals, but it seems likely that after such extensive practice, that they were able to develop an accurate model of the statistical properties of each dial, as confirmed by their stabilizing sampling patterns. According to Moray (1986), in order to see optimal behavior, it is not sufficient to tell operators about the statistical properties of the visual environment. Rather, they must “live” the experience of the dynamics of the process in order to embody the data rather than just know it.

There are two clear implications for the research this research regarding optimality in visual behavior. First, on the practical end, participants in all of the experiments needed to be extensively trained on the task. Likely due to the simplicity of the task, Senders (1983) notes that the sampling patterns of participants “settled” on a nearly optimal sequence after approximately 3-4 hours of training. Five hours of training at the task were employed for each of the experiments presented herein. The second implication for the current project is theoretical in nature. We examined a number of factors that are likely to influence sampling behavior. Careful examination of the data from each experiment indicated whether the participants were influenced by the independent measures under manipulation. Comparing the data to mathematically defined optimal models allowed us to determine if humans were capable of incorporating the manipulations into their sampling patterns in an optimal, or near optimal manner. This

allowed us to determine the relative influence of the various factors in achieving the benchmark of mathematically optimal sampling patterns.

4. Expected Value Models and the SEEV Model

Just as Fitts' did not account for bandwidth in developing his qualitative model of eye movements, Senders model does not explicitly account for the relative value of the instruments. Following Senders' original work, others elaborated on his model to include a notion of the relative value of an instrument (Sheridan, 1970; Sheridan & Rouse, 1971; Carbonell, 1966; Carbonell, Ward, & Senders, 1968). In order to account for value, in addition to bandwidth, and to dictate optimal scanning strategies, the models use the notion of the expected value of perceiving information at different instruments (more generally referred to as areas of interest, or AOIs). More specifically, they calculate the expected cost of missing critical events on other AOIs and attempt to minimize this probability, the probability that critical events will be missed. Such models are based on normative expected value models of decision making (e.g. Edwards, 1961; Payne, Bettman, and Johnson, 1993). The expected value model may be considered a normative or optimal model. It assumes that only two factors should drive the allocation of visual attention: the information generated by the AOI and the expected value of that information. An observer with a well-calibrated mental model of those two parameters, one that captures the objective levels of the parameters in the task at hand, will be able to minimize the chance of missing important information.

Wickens et al. (2003) have simplified and extended the general expected value models. They have identified two additional factors, effort and salience, that influence the frequency of visual sampling. The effort parameter was originally investigated in this

context by Kvalseth (1977) and Sheridan (1970), who noted the inhibiting role of information access effort required to sample information. Eye movements have an inexpensive time cost, but they are not "free." When a head movement is also required to sample information, the cost of such samples can be quite high, particularly if the observer is wearing cumbersome headgear, as in many military and airplane applications. In addition to the physical effort required when shifting visual attention, there is a cognitive component. The level of cognitive effort in shifting attention is exacerbated and can inhibit the control of visual scanning when there are concurrent cognitive of perceptual tasks required of the observer (Liu & Wickens, 1992). The second factor incorporated by Wickens et al. (2003) is the salience or conspicuity of an event that occurs on an input channel or within an AOI. This property has been the focus of a large body of research on models of visual search behavior, with a focus on the concepts of "visual onsets" and "attentional capture" (Yantis, 1993; Wolfe, 1994; Folk, Remington, & Johnston, 1992; Pashler, Johnston, & Ruthruff, 2001), but it has received little attention in engineering models of visual sampling.

Based on these four parameters, Wickens et al. constructed a descriptive model of visual sampling, which characterizes the probability that a given area will be attended (or the distribution of visual attention across AOIs). The model is known as the SEEV model, where each letter in the acronym represents a parameter in the model (Salience, Effort, Expectancy, and Value), and is given by:

$$P(A) = sS - efEF + (exEX)(vV) \qquad (4.1)$$

Each term in capital letters is a characteristic of a the visual environment that is determined by (1) the physical properties of the events (Salience = S), (2) the physical distance between a previously fixated and a current AOI, or the demands of concurrent tasks (Effort = EF), (3) an information-related measure of event expectancy (e.g. bandwidth, event rate; Expectancy = EX), and (4) an objective measure of the value (=V) of processing information at the AOI in question (or the cost of failing to attend there). The coefficients, s, ef, ex, and v are scaling constants, which represent the relative influence of these four factors on human scanning.

From the SEEV model, which is descriptive in nature, they developed a simpler, normative or optimal model of sampling. Their normative model extends previous normative models, such as that of Senders (1950), in that where previous models had defined properties of each AOI or information channel purely in terms of the bandwidth and value of events along that channel, it explicitly incorporates the many-to-one mappings of tasks to channels and channels to tasks in complex environments. This required the integration of models of information seeking with those of task management in an aviation context (Chou, Madhavan, & Funk, 1996; Dismukes, 2001; Raby & Wickens, 1994).

This model predicts that the probability of attending to an AOI is related to the sum, across all tasks supported by that AOI, of the value of those tasks multiplied by the degree of relevance (importance) of the AOI for the task in question, and by the bandwidth of their informational "events":

$$P(AOI_j = \sum_{t=1}^{n} (bB_t)(rR_t)(V_t) \qquad (4.2)$$

where t represents the tasks, numbered 1-N; B represents the bandwidth of the AOI; R represents the relevance of the task; and V represents the value of the task.

In relation to the original SEEV model, Expectancy is incorporated by the bandwidth of the channel, while Value is incorporated by the product of relevance and the value (ordinal rank) of the task in the priority hierarchy. Bandwidth (B) may sit inside or outside of the summation depending on whether it is assumed to be a general property of an AOI or is assumed to be a specific property of each task serviced by the AOI. Additionally, Equation 4 differs from the SEEV model (Equation 3) in that the former does not contain the non-optimal salience and effort factors.

In providing numerical predictions for such a model, matrices were created to accommodate the different parameter estimates of bandwidth, relevance and value that will be imposed in the varying conditions whose visual scanning measures can be employed to validate the model. Numerical coefficients were then assigned to the different cells of the matrices, across varying conditions, in order to compute model predictions.

Rather than attempt to estimate absolute values of the parameters, the approach of the authors (Wickens et al., 2001) in coefficient assignment was based on a "lowest ordinal algorithm." Across the range of conditions and AOIs employed, they simply ordered the values from highest to lowest based on an ordinal ranking system. For example, aviating, or keeping the plane in the air, was generally considered a higher priority task than navigating, or controlling the direction of the plane's flight. The cells of the matrices were filled in with the lowest values of these parameters that can still

preserve all ordinal relations within the rows and columns of the matrices. This system has the advantage of simplicity and that coefficients can be set based on relationships that multiple models can agree upon. An example of a set of matrices used in validating the model (Wickens, et al., 2001) is show in Table 4.1.

Parameter		AOI IP	AOI OW	AOI CDTI
Bandwidth (B)	Freeflight (Conf)	3	2	1
	Freeflight (Nconf)	2	1	0.5
	Baseline (Conf)	3	2	
	Baseline (Nconf)	2	1	

Relevance (R)				
	Aviate (FF)	2	1	0
	Navigate (FF)	1	2	2
	Aviate (Base)	2	1	
	Navigate (Base)	1	2	

Priority (V)	
Aviate (FF)	3
Navigate (FF)	2
Aviate (Base)	3
Navigate (Base)	1

Table 4.1. Parameter values assumed for two experimental conditions, free flight (FF) and baseline (Base). Higher numbers represent a higher ordinal ranking. From Wickens, et al., 2001.

Validating the model involved a series of experiments in which pilots had their eye movements recorded as they engaged in a series of flight simulations. The framework of a cockpit display of information was used. This included three areas of interest (the instrument panel, the outside world, and a cockpit display of traffic information or CDTI) to support two different tasks, aviating and navigating. Aviating was considered to be a more important task than navigating, and this task hierarchy defined the value (V) or importance of the AOIs that served each of the respective tasks. The relevance (R) parameter was determined by the number of sources of information each AOI contributed to a task, and the bandwidth parameter (B) was determined by the rate of change of that

information for each display, i.e., those displays with more information and faster rates of change were given higher ordinal rankings.

In the most complex validation of their model, there are four relevant AOIs, three tasks (aviate, navigate, and communicate), and five different conditions. Figure 4.2 shows the data for the scan measures (percentage of dwell time within each AOI) regressed against the model predictions for the different conditions. Differences in the percent dwell time in each of the three AOIs, across the various conditions were statistically significant (Helleberg & Wickens, 2001). The regression analysis of obtained on predicted scanning percentages revealed a very high degree of fit, with $r = 0.97$, or 95% of the variance explained. The percentage of variance explained for two other validation experiments was approximately 80%.

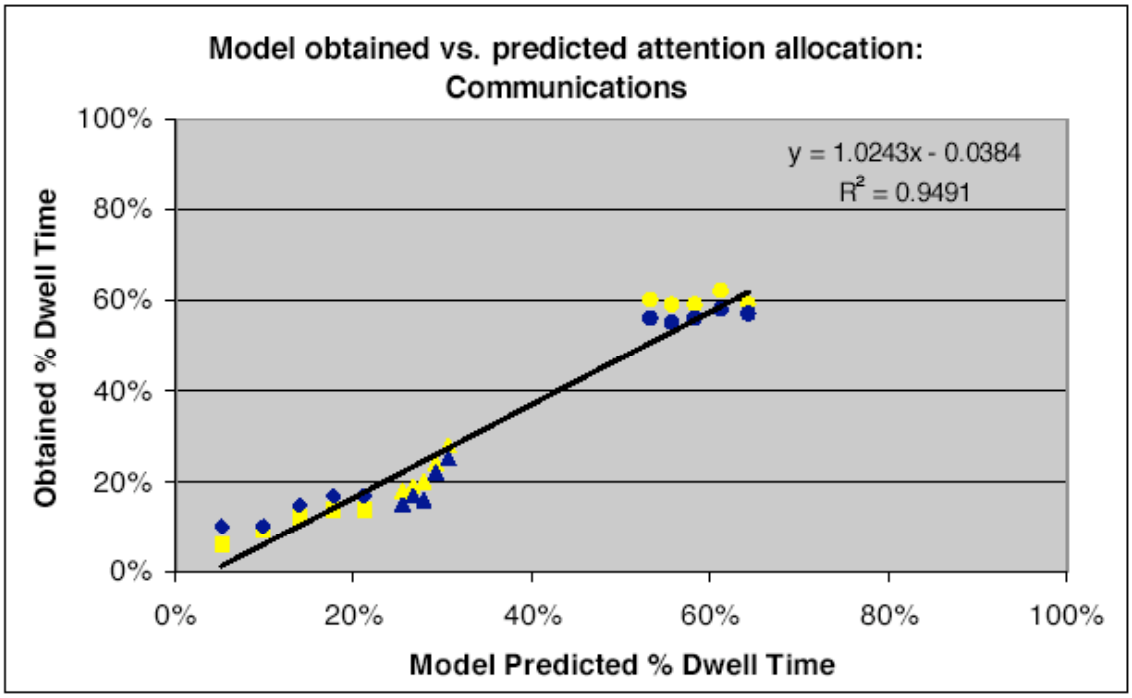


Figure 4.2. From Wickens, et al. 2001. Model fit of experiment. Each point represents the model predicted dwell time on an AOI. Different shapes represent different AOIs. The different conditions are not labeled but are represented by different shading (light or dark) and fall in monotonic order along the x-axis.

The results of the modeling effort suggested that well-trained pilots are quite optimal in allocating attention, as the model was normative in nature. This conclusion agrees with that offered by Moray (1986) in his comparison of model fitting results applied to novice subjects (substantial departures from optimality) and more trained experts (fewer departures). This conclusion further suggests that the two components of the descriptive SEEV model, conspicuity or salience and effort, either are not primary sources of variance among expert observers or that the function that describes human behavior with respect to these two components is flat. The model is able to account for 80% to 95% of the variance in visual scanning across multiple experiments without the

inclusion of these two parameters. With respect to salience, this may be due to the statistical averaging technique employed by the modelers (average percentage to dwell time across a trial). One role of salience, the attention capturing properties of an event, not the properties of a channel or AOI, would be to shorten the latency of attention allocation following an event, rather than to alter the overall distribution of attention to that AOI relative to others. In other words, observers may end up looking at channels where events frequently occur *less* often. It might be optimal to look at these channels less frequently, since attention is more likely to be captured by those more salient events in peripheral vision.

Effort may have played some role in influencing the scanning of participants. The authors observed relatively long dwell times, as long as six seconds, on a particular AOI, the instrument panel. They noted that during these longer dwells the observers looked at a number of different information sources, before moving on to a different AOI. They term this phenomena the "in the neighborhood" heuristic of scanning, whereby observers make several short saccades to information sources located near to each other (in the neighborhood) before making a longer saccade to a more distant information source. Such a strategy reduces the overall cost of information access.

Overall, the Wickens model exhibits several favorable comparisons with other models of visual sampling, such as that of Senders (1964) already discussed. First and foremost, the Wickens model is relatively simple, both in its mathematical structure and its employment. It involves only two parameters, expectancy and value, and a fairly simple heuristic for assigning those parameter values to particular conditions, according to the lowest set of values that preserve ordinal relations across tasks. This translates to a

model that is relatively easy to employ and is likely to show similar predictions when used by different modelers.

The Wickens model is flexible in that it may be extended by incorporating other visual sampling work. In the case that a set of parameters could not be agreed upon or easily calculated, a different model may be used in lieu of simple ordinal ranking. For example, the value parameter represents a judgment of the value of an AOI to a particular task. One way to calculate this value would be to use a queuing model algorithm, such as that of Carbonnell (1966) to calculate the probability that a significant event would be missed on another AOI when viewing any particular AOI. Those with a higher probability would be given a lower ordinal ranking.

The disadvantages of using the Wickens model are that any model is only as good as the parameter estimation that goes into it. The parameters of Wickens model are coarse, a rank ordering of value and relevance within a task. Indeed, this is one of the strengths of the model, the simplicity of estimating its parameters. Using a relatively simple parameter estimation system, the model is capable of making relatively accurate predictions of human performance. However, the model predictions cannot be made at a finer grain of analysis than the parameters are estimated at. Hence the predictions that it makes are also relatively coarse, i.e., the percent dwell time on regions of interest. It should be noted that this is a valuable and important metric of visual sampling performance. For example, it may be used to distinguish between expert and novice performance in a visual monitoring task (Moray, 1986; Wickens et al., 2001). However, it would be valuable to be able to make predictions on a finer grain of analysis, e.g. from

saccade to saccade or AOI to AOI. Whether the model may more accurately predict eye movement data given more precise inputs remains an open question.

The next model we will cover, though untried in the domain of visual sampling, has proven to be an accurate predictor of such fine grained behavior in other tasks.

5. Information Foraging Theory

Information Foraging Theory (IFT), developed by Pirolli and Card (1999), is an approach to understanding how strategies and technologies for information seeking, gathering, and consumption are adapted to the flux of information in the environment. The theory assumes that people, when possible, will modify their strategies or the structure of the environment to maximize their rate of gaining valuable information. Cognitive systems engaged in information foraging will exhibit such adaptive tendencies.

Information Foraging Theory stems from evolutionary ecological explanations of foodforaging strategies in anthropology (Smith & Winterhalder, 1992) and behavioral ecology (Stephens & Krebs, 1986). The theory is based on the analogy between the food foraging behavior of animals and the uniquely human search for information. Much of Pirolli and Card's introduction of the theory lies in elucidating this analogy. Similarly, much of the discussion here will focus on extending that analogy from the physical search for information to the visual sampling of information in the visual environment. It is our hypothesis that just as humans engage in a variety of information gathering activities, such as searching the internet or collecting references in a library, the human visual system is constantly engaged in the process of gathering information from the visual environment. Information foraging theory demonstrates that information-gathering activities settle on stable states that converge on maximizing the amount of information gained per unit of cost. At the root of our analogy is the concept that in a scenario where the statistical properties of the environment are well-understood, the visual sampling

patterns of an observer will converge on patterns that maximize the rate of information gain.

5.1 The Roots of Information Foraging Theory

Information Foraging Theory draws heavily upon models and techniques developed in optimal foraging theory (Stephens & Krebs, 1986), which seeks to explain adaptations of organism structure and behavior to the environmental problems and constraints of foraging for food. Optimal foraging theory originated in attempts to address puzzling findings that arose in ecological studies of food seeking and prey selection among animals (Stephens & Krebs, 1986). It has had an enormous impact in anthropology (Smith & Winterhalder, 1992), where it has been used to explain dietary choice (Kaplan & Hill, 1992), variations in land tenure and food sharing (Smith, 1987), group size (Smith, 1981), habitat choice (Cashdan, 1992), time allocation (Hames, 1992), and many other aspects of hunter-gatherer culture. Pirolli and Card (1999) propose that the information foraging adaptations they discuss are *exaptations* of the behavioral plasticity that humans evolved for food-foraging.

It is our hypotheses that if these exaptions of optimal tendencies of foraging behavior are evident in human information gathering activities, and there is strong evidence that they are (Pirolli & Card, 1999), then they should also be evident in visual behavior. Vision is the primary sense by which we acquire information from the environment. As such, a larger portion of our brains is devoted to visual processing than to any other sense. Indeed, it is possible that other foraging behaviors, such as those for food and information, are exaptions from visual behavior, not the other way around as Pirolli and Card (1999) suggest[2].

---

[2] Pirolli and Card (1999, page 6) state, "We would like to think that the information foraging adaptations we observe are *exaptations* of the behavioral plasticity that humans evolved for food-foraging, but it is unlikely that we will be able to obtain data relevant to tracing this evolution." Although we may not be able

One area of research from which IFT has drawn from is the field of evolutionary psychology. Evolutionary psychology (Barkow, Cosmides, & Tooby, 1992) argues for cognitive universals in mental architecture that have evolved as adaptations over evolutionary time, with an emphasis on specific complexes for such activities as social exchange (Cosmides & Tooby, 1992) and language (Pinker & Bloom, 1992). The general theory seeks to explain complex cognitive modules, typically by looking back at the physical and social environments that shaped human evolution during the Pleistocene (Tooby & Cosmides,1992). A motivating analogy for evolutionary psychology is the eye, an exquisite, function-specific, complex arrangement of tissue that has evolved due to strong selection pressures (Tooby & Cosmides, 1992). We propose that not only is the "hardware" of the eye exquisitely adapted to the environment, but the "software" or strategies that guide the behavior of the eye are also well-adapted to the physical environment.

Another theory related to IFT and that speaks to the adaptation of human cognition to the environment is Rational Analysis (Anderson, 1990), which focuses on general mechanisms that solve information-processing problems posed by the environment. For instance, the mechanisms of memory (Anderson & Milson, 1989) are characterized as adaptive solutions to the recurrent structure of events in the world (Anderson & Schooler, 1991), and mechanisms of categorization (Anderson, 1991) are characterized as adaptive solutions to recognizing naturally occurring phenotypes in the biological world. Similarly to evolutionary psychology, it is argued that the statistical properties of the environment exert strong selection pressures, which have shaped human cognitive abilities.

---

to trace the evolution of some information foraging behaviors partly because other animals do not engage in information gathering activities in the same manner as humans do, the visual system may provide such an evolutionary history. To the extent that humans exhibit optimality in visual scanning patterns, it may be possible to show that less-developed visual systems in more primitive animals exhibit less-than-optimal visual scanning tendencies.

The theory proposed here suggests that because vision is the primary sense by which humans acquire information from the environment, it is probable that human vision is sensitive to the statistical structure of the visual environment. Further, I propose that this adaptation is evident in the visual sampling strategies of humans in any scenario where the statistical properties of the environment are well understood. Specifically, that the visual sampling patterns of observers will approach optimality in terms of maximizing the rate of information gained per unit cost.

5.2 Analogies with Food and Information Foraging

In describing IFT, Pirolli and Card (1999) make use of several analogies between food foraging behavior as described by optimal foraging theory and information foraging behavior as described by IFT. Those analogies are provided here along with some discussion of our extension of the theories to visual sampling behavior.

Pirolli and Card (1999, page 8) first offer an analogy to describe the "essence of conventional models of foraging theory."

> Imagine a predator, such as a bird of prey, that faces the recurrent problem of deciding what to eat, and we assume that its fitness, in terms of reproductive success, is dependent on energy intake. Energy flows into the environment and comes to be stored in different forms. For the bird of prey, different types of habitat and prey will yield different amounts of net energy (energetic profitability) if included in the diet. Furthermore, the different food-source types will have different distributions over the environment. For the bird of prey, this means that the different habitats or prey will have different access or navigation costs. Different species of birds of prey might be compared on their ability to extract energy from the environment. Birds are better adapted if they have evolved strategies that better solve the problem of maximizing the amount of energy returned per amount of effort. Conceptually, the optimal forager finds the best solution to the problem of maximizing the rate of net energy returned per effort expended, given the constraints of the environment in which it lives. These constraints include the energetic profitabilities of different habitats and prey, and the costs of finding and pursuing them. This is the essence of conventional models in optimal foraging theory (Stephens & Krebs, 1986).
>
> An analogous situation in information foraging theory might be an office worker or academic researcher facing the recurrent problems of finding task-relevant information. Information flows into the environment to be represented in different types of external media, such as books, manuscripts, or on-line documents. The different information sources (or repositories) will have different profitabilities, in terms of the amount of valuable information returned per unit cost in processing the source. In addition, the different kinds of sources will be distributed in the task environment in different ways. Some will be more prevalent, or less effortful to access, than others. Conceptually, the optimal information forager is one that best solves the problem of maximizing the rate of valuable information gained per unit cost, given the constraints of the task environment. These constraints include the profitabilities of different sources, and the costs of finding and accessing them.

This analogy extends well to the domain of visual information. In this sense, information may be anything in the visual environment. This information flows to the observer via the eye. Different pieces of information may have different profitabilities. For instance, information that is highly task-relevant may be considered more "profitable" than information that is less task-relevant. For instance, in a typical visual search task studied in the laboratory, the observer may be told to locate a particular red object, such as a red "X" amongst other red objects and other green objects. Any red object may be considered more profitable than a green object, and will draw the attention of the observer, but red "Xs" will be considered the most profitable object. In a Senders-like task in which the observer is told to monitor a set of ammeter-like dials for alarms, a dial with a high bandwidth may be considered more profitable than a dial with a low bandwidth because the high-bandwidth dial provides more information per unit of time and because the opportunity cost of missing a critical event on another dial is lower when observing a higher bandwidth dial. Finally, the different objects or sources of information in the visual environment have different distributions in the environment. For one, the likelihood of encountering a certain object in the visual environment may change with the environment, and secondly, the cost of acquiring the relevant information may vary with the object. For example, some objects may require the observer to move their head in order to view the object, to push a button or change the display in order to view the information, or simply may need to be viewed for a longer period of time in order to extract relevant information (such as a low-bandwidth dial). Conceptually, the optimal visual system is one that maximizes the profitability of the objects it views relative to the costs of observing those objects.

Pirolli and Card (1999) discuss the decision problem of how to allocate time to different activities, which arises when the environment has a "patchy" structure. As an example of food-foraging in a patchy environment, they discuss a bird that forages for berries found in patches on berry bushes. "The forager must expend some amount of *between-patch* time getting to the next food patch. Once in a patch, the forager engages in *within-patch* foraging, and faces the decision of continuing to forage in the patch or leaving to seek a new one. Frequently, as the animal forages within a patch, the amount of food diminishes or depletes. For instance, our imaginary bird would deplete the berries on a bush as it ate them. In such cases there will be a point at which the expected future gains from foraging within a current patch of food diminish to the point that they are less than the expected gains that could be made by leaving the patch and searching for a new one. Quantitative formulations of patch models in optimal foraging theory determine the optimal policies for allocating time to foraging within a food patch vs searching for new patches." Information is commonly arranged in patches and the information forager must make similar decisions about how time should be allocated among between-patch foraging activities and within-patch foraging activities. For instance, the information forager must decide when to move from one WWW website to another or from one pile of documents to another relative to continuing to search within a particular website or stack of documents.

Observers engaged in visual monitoring tasks must make similar decisions of how to best allocate their time and energy. For instance, in a typical visual monitoring task, the observer must monitor an array of displays that are constantly changing. The observer must determine how long he or she should continue to look at a particular display and when to move on to a different display. As in information foraging, this decision will depend on the task at hand. Assuming that the display in which the observer is monitoring is constantly changing, then the observer will be gaining some degree of information as long as he or she continues to monitor that display. However, as he or she

continues to monitor a single display, their uncertainty regarding the other displays is increasing, as the other displays are also constantly changing. Once the degree of uncertainty of the other, un-monitored displays, reaches a level greater than the level of uncertainty for the currently observed display, then it would behoove the observer to move to a different display.

In IFT, the decision regarding when to leave an information patch and which patches to pursue are informed by proximal cues. For instance, which link to select on a WWW web page may be based on the text of the link and possibly the text surrounding that link. The text of link in this case acts as the proximal by which the decision-maker judges the profitability of selecting a particular WWW link. In IFT, these proximal cues are known as *information scent*. In other words, the proximal perception of information scent is used to assess the profitability and prevalence of information sources. These scent-based assessments inform the decisions about which items to pursue so as to maximize the information diet of the forager. If scent is sufficiently strong, the forager will be able to make the correct choice at each decision point. If there is no scent, the forager would perform a random walk through the information space.

In visual sampling, information scent is also inferred from proximal cues. These proximal cues may be derived from two sources. One source is the observer's knowledge about the visual environment. For instance, the observer may know that a desired source of information is found in a particular location in the environment, and he or she may direct visual attention to that source without any additional information from the environment. This is known as "top-down" processing of visual attention allocation, as the decision regarding the allocation of visual attention is based entirely on internally-derived information. Conversely, the decision of when and where to look next may be a function of the visual information presented in the environment. For instance, a sudden movement viewed in one's peripheral vision, may "grab" one's attention. This is known as "bottom-up" processing, as attention-allocation decisions are a function of the external

environment. Both bottom-up and top-down processing make use of the ability to "preattentively" parse the visual environment. It is well-established in vision research that some visual information is available to the visual system before anything in the environment has explicitly been the focus of attention (Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989; Wolfe, 1994). Often, when and where to look is a function of both externally and internally derived information. For instance, if an observer is told to find a red X in a field of green Xs and red Os, his or her visual system will be guided by the knowledge that they are searching for a red X as well as the properties of the display, i.e. the proportion of red Os and green Xs on the display (Bacon & Egeth, 1997; Poisson & Wilkinson, 1992; Shen, Reingold, & Pomplun, 2000). Indeed, researchers are working on developing models that predict how these types of information, internal and external, are combined to determine the allocation of visual attention (Wolfe, 1994; Wolfe & Gancarz, 1996).

Information derived via bottom-up or top-down channels comprise the proximal cues that guide the allocation of visual attention, in a similar manner to how information scent guides the behavior of an information forager. In a typical Senders task where the participant is instructed to detect "alarms" on a display of four simulated ammeters, which is the paradigm we are exploring, the participant is not able to guide their attention solely via bottom-up cues. The ammeters are spaced apart sufficiently such that they must be foveated for the participant to determine their value and whether that value represents an alarm. Rather, our adaptation of IFT predicts that the participant will be forced to rely on the top-down information, which will include their knowledge of the bandwidth of the dials, their alarm frequencies, and their relative values. These factors influencing the allocation of visual attention will be mediated by the effort required to view the ammeters.

5.3 Foraging (Optimization) Model Components

In order to create an optimization model of foraging behavior, three components must be present.

Decision Assumptions must specify the decision problem to be analyzed. In the context of visual sampling, this would be how long to look at a region and where to look next.

Currency Assumptions identify how choices are to be evaluated. Visual foraging activities will assume the value of information at the locus of visual attention to be the relevant currency. Similarly to IFT, choice principles include maximization, minimization, and stability of that currency.

Constraint Assumptions limit and define the relationships among decision and currency variables. In IFT, "these include constraints that arise out of the task structure, interface technology, and the abilities and knowledge of a user population" (Pirolli & Card, 1999, page 6). In visual sampling, the primary "interface technology" is the human eye. Constraints include how long it takes to change the locus of visual attention (make a saccade) and how long it takes to perceive information at a given location. Naturally, these constraints are also a function of the visual environment as well. How long it takes to change the locus of visual attention will depend on the eccentricity of the to-be-attended object from the current focus of visual attention. How long a region must be attended in order to obtain useful information may be a function of the bandwidth of the information source and whether the information source must be the foveated in order for the observer to obtain useful information.

5.4 The Conventional Model

Conventional models of information foraging rest on a few strong assumptions, but their results appear to be broadly applicable even when the assumptions are relaxed

(Stephens & Charnov, 1982). An initial assumption we begin with is that a forager's activities may be divided into two mutually exclusive sets: (1) between-patch activities (i.e., searching for the next item) and (2) within-patch activities (i.e., exploiting an item). The notion of a patch is loosely defined. For instance, in information foraging activities a patch may be considered a collection of documents or a collection of content contained within an individual document. Similarly, in visual information foraging activities, a patch may be considered an individual object, such as an ammeter, or a collection of objects.

Let R be the rate of gain of valuable information per unit cost. This is operationalized by Hollling (1959) as the ratio of the total net amount of valuable information gained, G, divided by the total amount of time spent between-patches, T_B, and exploiting within patches, T_W,

$$R = G / (T_B + T_W) \text{ information-value-units/cost units.} \qquad (5.1)$$

Including additional assumptions, we can construct a version of the model based on averages. In order to do so, we must assume that (1) the number of patches processed is linearly related to the amount of time spent in between-patch foraging activities, (2) the average time between processing patches is t_B, (3) the average gain per item is g, and (4) the average time to process patches is t_w. Then the average rate of encountering patches can be defined as

$$\lambda = 1/ t_B \qquad (5.2)$$

The total amount of information gained can be represented as a linear function of between-patch foraging time as,

$$G = \lambda T_B g \quad (5.3)$$

The total amount of within-patch time can be represented as,

$$T_W = \lambda T_B t_w \quad (5.4)$$

Substituting terms, Equation 1 may be re-written as

$$R = (\lambda g) / (1 + \lambda t_w) \quad (5.5)$$

The profitability of information patches, π, can be defined as the ratio of net value gained per patch to the cost of within-patch foraging,

$$\pi = g / t_w \quad (5.6)$$

This is intuitively informative, as according the equations just outlined, increasing the profitability of within-patch activities increases the overall rate of gain, *R*. Decreasing the between-patch costs, t_B, or increasing their prevalence, λ, increases the overall rate of return, *R*, towards an asymptote equal to the profitability of patches, $R = \pi$ (Pirolli & Card, 1999).

5.4.1 Time Allocation Within and Between Information Patches

The conventional patch model of optimal foraging theory (Stephens & Krebs, 1986) is an elaboration of Equation 4.5. It addresses the optimal allocation of total time to between-patch activities vs within-patch activities, under certain assumptions. Rather

than having a fixed average gain per patch and a fixed average within-patch cost the patch model assumes that (a) there may be different kinds of patches and (b) that the total gains from a patch depend on the within-patch foraging time, which is under the control of the forager. The patch model assumes that there may be different kinds of information patches, which may be denoted using i = 1, 2, ...P. Further, the patch model assumes that the forager must expend some amount of time going from one patch to the next. Once in a patch, the forager faces the decision of continuing to forage in the patch or leaving to seek a new one. For a particular type of patch, the function g_i (t_{wi}) represents the cumulative amount of valuable information returned as a function of within-patch foraging time t_{wi}. In this example, there is a linear increase in cumulative within-patch gains up to the point at which the patch is depleted.

If we assume that patches of type *i* are encountered with a rate λ_i as a linear function of the total between-patch foraging time, T_B. Now imagine that the forager can decide to set a policy for how much time, t_{wi}, to spend within each type of patch. The total gain could be represented as,

$$G = \sum_{i=1}^{P} \lambda_i T_B g_i(t_{wi}) \tag{5.6}$$

$$= T_B \sum_{i=1}^{P} \lambda_i g_i(t_{wi}) \tag{5.7}$$

Likewise, the total amount of time spent within patches could be represented as,

$$T_W = \sum_{i=1}^{P} \lambda_i T_B t_{wi} \tag{5.8}$$

$$= T_B \sum_{i=1}^{P} \lambda_i t_{wi} \tag{5.9}$$

The overall average rate of gain would be:

$$R = \frac{G}{T_B + T_W} \tag{5.10}$$

$$= \frac{T_B \sum_{i=1}^{P} \lambda_i g_i(t_{wi})}{T_B + T_B \sum_{i=1}^{P} \lambda_i t_{wi}} \tag{5.11}$$

$$= \frac{\sum_{i=1}^{P} \lambda_i g_i(t_{wi})}{1 + \sum_{i=1}^{P} \lambda_i t_{wi}} \tag{5.12}$$

5.5 Adapting the Model to the Visual Environment

Our analysis of IFT adapted to the visual environment will focus on two metrics, the proportion of fixations devoted to each AOI and the average dwell time on each AOI.

IFT predicts that attention is allocated to different AOIs based on the rate of information gained, R. As such, it predicts that the proportion of fixations and time allocated by participants to the different AOIs will be a ratio of the rate of gain of an individual AOI to the sum of information gained from all AOIs. The transition matrix between the AOIs is calculated by multiplying the predicted proportions of fixations devoted to each AOI.

Two different calculations of the rate of gain of an AOI will be examined. One method used will employ equation 5.5 from above. In this formulation, T_B, the time between AOIs, will be constant for each of the AOIs and can hence be removed from the equation. The time within an AOI, T_W, will be determined empirically through examination of the eye data. The information gained for an area of interest, G, will be calculated in terms of an expected value, similar to the notion of expected value in the

SEEV model. The expected value of an AOI will be equal to the bandwidth, alarm frequency, or some combination of alarm frequency and bandwidth of the dial multiplied by its value.

The second formulation of the equation that will be examined is based on equation 5.6, which incorporates different notions for the rate of gain of an AOI and for the average time to encounter a different information source, t_B. As shown in equation 5.4, t_B may be considered the product of the average rate of encountering information, λ, and the average time to process an information source, t_w. Again, t_w was determined empirically from the eye tracking data. The average rate of encountering information, λ, will be considered to be either the bandwidth, the alarm frequency, or the ecological validity of the dial depending on which source of information participants are most sensitive to. The information participants should be concerned with is the alarm frequency of the dial. As such, the average rate with which they should expect to encounter this information is directly related to the average alarm frequency of each dial. However, as discussed in the next section alarm frequency is a distal cue that may be difficult to assess, and as such participants may be influenced by the bandwidth of the dial, a proximal cue.

6. Impetus for Experimental Manipulatoins

The results of Senders's experiments (Senders, 1964, Senders, 1983) strongly suggest that the bandwidth of an instrument is a key factor influencing the allocation of visual attention in a visual monitoring task. Senders defined bandwidth as the "information generation rate of a signal." This definition is sufficiently broad to acknowledge that a number of factors may contribute to the information rate of a signal and hence, influence the scanning patterns of individuals. However, the notion that patterns of visual sampling are dependent upon the information rate of a signal was only

tested with one notion of the bandwidth of an instrument. Bandwidth in Senders's experiments was simply the rate of movement of the signal. We have identified four additional factors, other than rate of movement that may contribute to the information rate of a signal. Three of these are components of the SEEV model, effort, value, and salience, and one, alarm frequency, indicates a possible confounding variable in the Senders's experiments. A brief description of each of the four hypothesized factors follows.

Alarm frequency: the rate at which the signal enters a state indicating an alarm.

Value: Some information sources may be considered to be more important to the task at hand than others, and hence the observer may place a higher value on them, which may lead to more frequent sampling of that source.

Effort: Different information sources may require different levels of effort from observers in order to be adequately sampled. The effort required may be physical, such as requiring an observer to adjust their position in order to view the source, or cognitive, such as the effort stemming from sources which are difficult to read or interpret.

Salience: Some sources of information, or at least pertinent information at that source, may not require that the observer frequently sample the source to be aware of them. Information sources with high visual salience should be sampled more or less frequently than other sources depending on the relationship of the high salience source to the task at hand.

6.1 General Procedures

Each of the experiments followed the same basic paradigm, which is an adaptation of the paradigm developed by Senders and colleagues (1964). This paradigm will be described in some detail here.

Each participant was presented with four virtual ammeters on a computer screen (see Figure 6.1). The task of the participants was to press a key whenever the pointer on any of the dials passed a predetermined set point on the dial. (Passing the set point on the dial would constitute an "alarm" and will be referred to as such.) On the dials, the region constituting an alarm was represented by a darker shade of gray; i.e. whenever the needle of a dial entered the dark gray area, the participant should have pressed the alarm key corresponding to that dial.

The keys corresponding to the upper-left and the lower-left dials were the "A" and "Z" keys, respectively. The keys corresponding to the upper-right and the lower-right dials were the apostrophe (') and slash (/) keys, respectively. Note that the location of these keys on the keyboard corresponded to the location of the dials on the screen, i.e., the "A" key is the upper-left of the four keys and corresponds to the upper-left of the four dials. Participants were instructed to keep their fingers on these four keys throughout the course of the experiment.

The movement of each of the dials was driven by quasi-random forcing functions (sum of sinusoids), which have been shown to appear random to human observers (Moray, 1986). The four dials had bandwidths of 1/2, 1, 2, and 4 radians per second. These bandwidths are the same as those used by Senders's (1964). They were chosen because they push the observers' sampling abilities to the theoretical limits as calculated

by Senders, i.e., a participant will just be able to sample each of the dials at the Nyquist interval (see Section 3 for a discussion of the Nyquist interval as it relates to optimal visual sampling behavior).

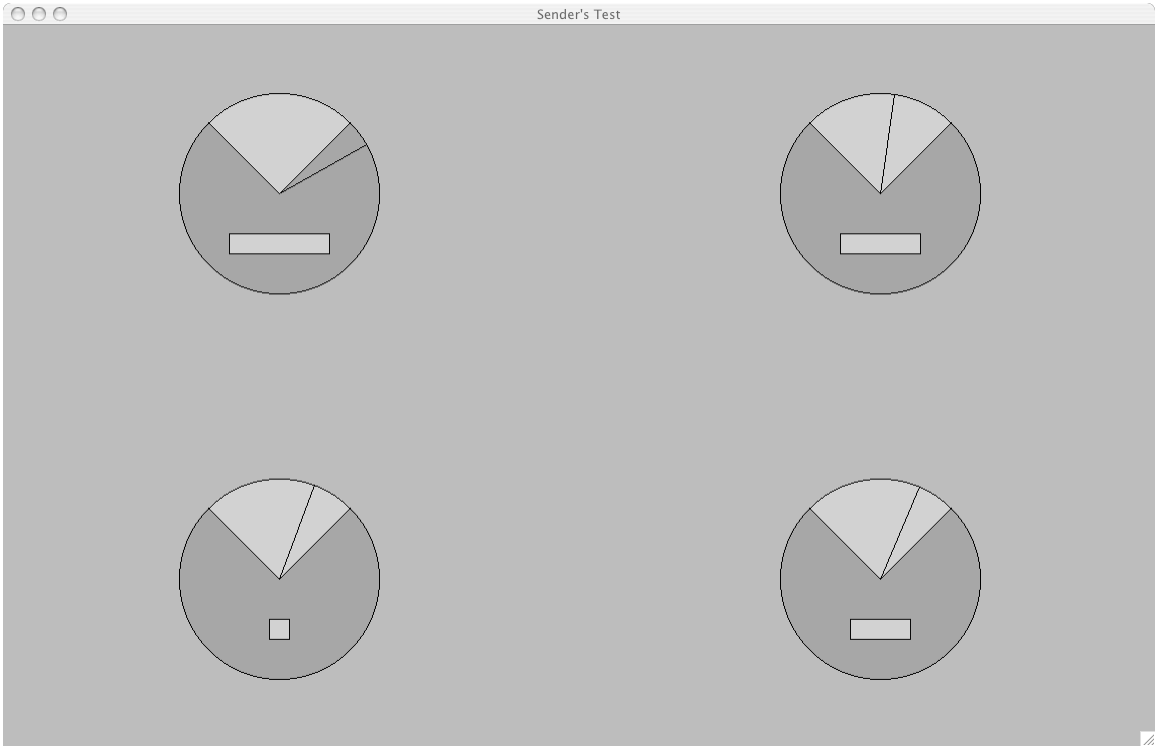


Figure 6.1. The basic set-up used in the experiments. An alarm has occurred on the upper-left dial. The screen shot was taken from Experiment 3.

The participant was positioned such that he/she was viewing the screen from a distance of approximately 20 inches. At this distance, each ammeter was separated by approximately 12 degrees of visual angle horizontally and 7 degrees of visual angle vertically, with the intention that a participant would only be able to observe one dial at a time. Each dial subtended 6 degrees of visual angle.

Scoring:

Participants were awarded points based on the correct detection of alarms. Each dial had a point value associated with it, and participants were given the number of points for a dial whenever they correctly detected an alarm on that dial. Participants were told the point value of each of the dials, and they were present on the screen in the form of bars of different lengths. In Figure 6.1, the bars in the alarm range of each gauge were used to indicate the number of points each gauge was worth. In this case, the longest bar was worth 10 points, the bigger medium-sized bar was worth 6 points, the smaller bar was worth 4, and the smallest bar was worth 2 points. Each time a participant correctly detected that a gauge had gone out of bounds they received the number of points indicated by the length of the bar. Conversely, they received negative points for misses (any time that an alarm occurred on a dial and they did not detect it) and false alarms (when they pushed the button associated with a dial, yet there was no alarm on that dial). The number of negative points they received for each dial is the same as the number of positive points they would have received for the correct detection of an alarm on that dial.

The participants' overall score was the positive points they earned for each time they correctly detected that a gauge had gone out of bounds minus the points lost for both misses and false alarms. Since there were many ways to lose points, it was possible for participants to receive a negative overall score on a trial. Their total score was given to them at the end of the each 5-minute trial.

A response was scored as correct if it occurred within 1 second of a significant deviation of a pointer, an alarm. The participants were allowed to predict an alarm based on prior observations, even though he/she may not be observing the dial in question at the

time of the response. Participants were given a bonus incentive (a candy bar) for each session on which they correctly identified 95% or more of the alarms on all dials. The participants in each experiment also competed against each other for a larger reward (a gift certificate). Of the five participants in each experiment, the participant with the highest average score across all trials received a $15 gift certificate. The second and third highest average scores received gift certificates of $10 and $5, respectively.

7. Specific Methods and Techniques

A series of three experiments were conducted in which each of the aforementioned factors, alarm frequency, value, salience, and effort, were systematically varied. The design elements common to each experiment are discussed here.

7.1 Participants

Five participants took part in each of the three experiments (15 participants total). Each participant took part in an experiment for one hour per day on five different days within a ten-day period. During each one-hour session, each participant spent 40 minutes (8 trials of 5 minutes each) participating in the experimental task, during which time their eye movements were recorded. Each participant was allowed to take a break from the task between trials whenever they felt it was necessary. The lone deviation from this pattern was the first experimental session, in which participants only completed 6 trials. Thus, each participant spent 3 and 10 minutes engaged in the task. This provided the participants with ample time to develop a consistent sampling pattern, as Senders (1964 and 1983) found that participants sampling patterns stabilized after approximately 2 to 3 hours of practice at the task.

7.2 Eye tracking

Participants's eye movements were tracked using an ISCAN RK726/RK520 HighRes Pupil/CR tracker with a Polhemus FASTRACK head tracker. Head-mounted optics and a sampling rate of 60 Hz were used in the experiment. This system, like many other laboratory eye trackers, works by shining an infrared light on the eye and taking a video image of the eye. From that image, it is possible to determine the pupil center and the point on the cornea closest to the camera (the corneal reflection) and take the vector between them. This vector changes as the eye orients to different positions on the screen, and with calibration to known points, it is possible to compute visual point of regard (POR, also referred to as "point of gaze"). The magnetic polhemus is used to compensate for head movements. POR reports by the eye-tracking equipment are typically accurate to within one-half degree of visual angle.

The POR was overlayed in real time on a monitor showing the screen that the participant was observing. The contents of this monitor were recorded to digital video.

An experimenter monitored each experimental trial and recalibrated the eye tracking equipment if there appeared to be a sizable disparity between reasonable expectations about where participants were looking and the position reported by the tracker.

7.2.1 Analysis Technique

From each of an individual participant's 38 trials (6 trials from the first session and 4 additional sessions with 8 trials per session), one minute was selected at random for further analysis. Thus, approximately 38 minutes of video were analyzed for each participant.

The digital video was analyzed using Videoscript software. Videoscript was used to track the position of the crosshairs representing the POR of the participant in the videos of their eye movements. The output of the software was the location on the screen (in x and y coordinates) of the POR for each frame of digital video, which was analyzed at a rate of 30 Hz.

Once the POR locations were extracted from the video data, metrics of the allocation of the visual attention of the participants were calculated. Software was written (in Videoscript) to calculate the number of frames in which the POR of the participant fell within each of the four quadrants of the screen containing a dial, the transition matrix between these quadrants (the number of transitions from looking at any one dial to any other dial, including the currently attended dial), and number of consecutive samples within each region (i.e., the dwell duration). These metrics provided the basis for comparing participant performance to the predictions made be each of the aforementioned models of visual attention.

7.3. Calculation of Model Predicted Sampling Patterns

In previous sections of this document, three different theoretical models of visual sampling behavior were described. In Section 3, we described the basic Senders' Model and its adaptation in the Constrained Random Sampler. In Section 4, the SEEV model and its implications in the visual sampling domain were discussed. In Section 5, Information Foraging Theory (IFT) was introduced as a possible model of visual sampling behavior, albeit one that had yet to be evaluated in such a context.

In the three experiments conducted, several statistical properties of the dials were manipulated, namely, bandwidth (BW), alarm frequency (AF), value (V), the effort

required to sample a dial (E), and the visual salience of a dial (S). None of the proposed theoretical models of the task incorporate all of the variables manipulated. (The SEEV model incorporates four of the five manipulations, but it does not distinguish between BW and AF, nor has it been evaluated with effort and salience as factors.) Hence, in order to account for the experimental manipulations, the models or the variables in the models must be combined. This section describes how the models were combined, and the eventual models that were produced. Although various models were attempted, seven models were eventually used in the final analysis of the data. This section describes how the proportion of attention to a quadrant was calculated in each of the seven models.

Each of the models calculates a relative "expected value" for each of the sources of information (dials, in this case) in the display. We make the strong assumption that the proportion of visual attention devoted to any source of information is directly proportional to the relative expected values of the dials. Hence, the predicted proportion of visual attention devoted to each dial was calculated as the expected value of each dial divided by the sum expected values of all of the dials.

7.3.1. Combining Senders' CRS Model and the SEEV Model

As noted, the original Senders' models only incorporated BW as the lone factor influencing the sampling strategy of participants. Additionally, the model had only been evaluated in a context where AF and BW were perfectly correlated. In order to account for the disassociation of BW and AF and the additional experimental manipulations, the basic Senders' model was combined with the theoretical sampling factors from the SEEV model, specifically, alarm frequency and value. These were combined in a multiplicative manner, and four models were derived, multiplicative models of BW and value (BW*V),

alarm frequency and value (AF*V), and two models combining all three of the factors, a BW*AF*V model, and a model that averaged the contributions of BW and AF, (V*(BW+AF)/2).

7.3.2 Information Foraging Models

The original information foraging theory model was not developed for application in visual monitoring tasks, and the parameter of the model needed to be adapted to a new domain. Section 5.5 covers two possible manners of adapting the parameters of the model to a visual sampling environment. The equation that was used in Experiments 1 and 3 was based on equation 5.5:

$$R = (\lambda\, g) / (1 + \lambda\, t_w) \qquad (5.5)$$

Where R is the rate of information gain, λ is the rate of encountering information, g is the gain of an item, and t_w is the time within an information patch. In the context of the experiments, each dial was considered to be an information patch, and the rate of encountering information was assumed to be either the bandwidth of a dial, its alarm frequency, or some combination of the two—all three formulations of the rate of encountering information were used and separately evaluated. The gain of an item was considered to be a dial's point value, and the time within an information patch was the average dwell duration of participants on a dial.

For Experiment 2, in which only one dial was visible at a time, and the time between when a dial was selected for view and when it became viewable (referred to as

the delay), the equation could be simplified. In this formulation the time between encountering information, t_B was considered to be the delay time, and the IFT equation became:

$$R = \lambda\, g / (t_B + t_W) \qquad (7.1)$$

Three variations of the IFT model, which differed only in their notion of λ, were evaluated. The three different models used either BW (referenced as the IFT (BW) model), AF (IFT(AW) model), and the average contributions of BW and AF (IFT(BW+AF) model).

8. Experiment 1

In Experiment 1, each dial had a bandwidth (rate of movement) that was different from the alarm frequency of that dial, such that the overall ecological validity (the correlation between bandwidth and alarm frequency) of the 4-dial matrix was 0.50. A point value was also associated with each dial. The pairings of bandwidth (BW), alarm frequency (AF), and value for each dial are shown in Table 8.1.

Dial	BW (radians/second)	AF (alarms/sec)	V (Points)
Upper-Right	1	0.02	2
Lower-Right	4	0.08	4
Lower-Left	2	0.12	6
Upper-Left	0.5	0.05	10

Table 8.1 Bandwidth, alarm frequency, and point value associated with each dial.

8.1 Scoring Results

In general the scoring data is presented as a proportion of the maximum possible score on each trial (Figure 8.1 is the only exception). In order to calculate the maximum score, for each dial the total number of alarms on that dial was multiplied by the point value of the dial. The maximum score attainable on any trial was the sum of these products. The maximum score for each trial in Experiment 1 was 474 points. Because participants received negative points for both false alarms and misses, they often had negative scores during the early part of the experiment.

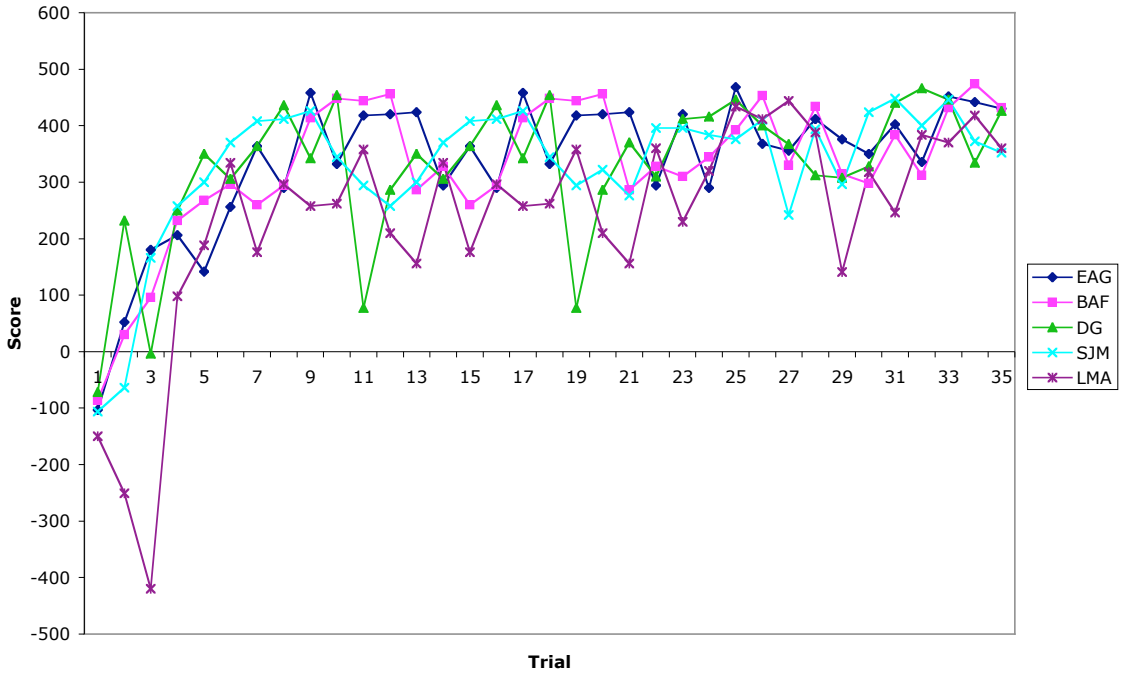


Figure 8.1 Experiment 1, score of individual participant by trial.

Examination of the participants' scores from Experiment 1 revealed a reliable effect of trial, $F(6, 24) = 6.55$, $p < 0.001$, indicating that participants' scores improved over the course of the experiment (Figure 8.1). Also revealed was a reliable effect of session, $F(6, 24) = 40.55$, $p < 0.001$, again indicating that the participants' performance improved over the course of the five sessions (Figure 8.2). Much of this improvement occurred within the first session, and was revealed in a reliable session by trial interaction, $F(24, 96) = 4.72$, $p < 0.001$ (Figure 8.3). This improvement in performance was revealed to be linear in nature, as evidenced by a significant linear effect of trial, $F(1, 4) = 16.79$ $p < 0.05$, and of session, $F(1, 4) = 67.44$, $p < 0.001$.

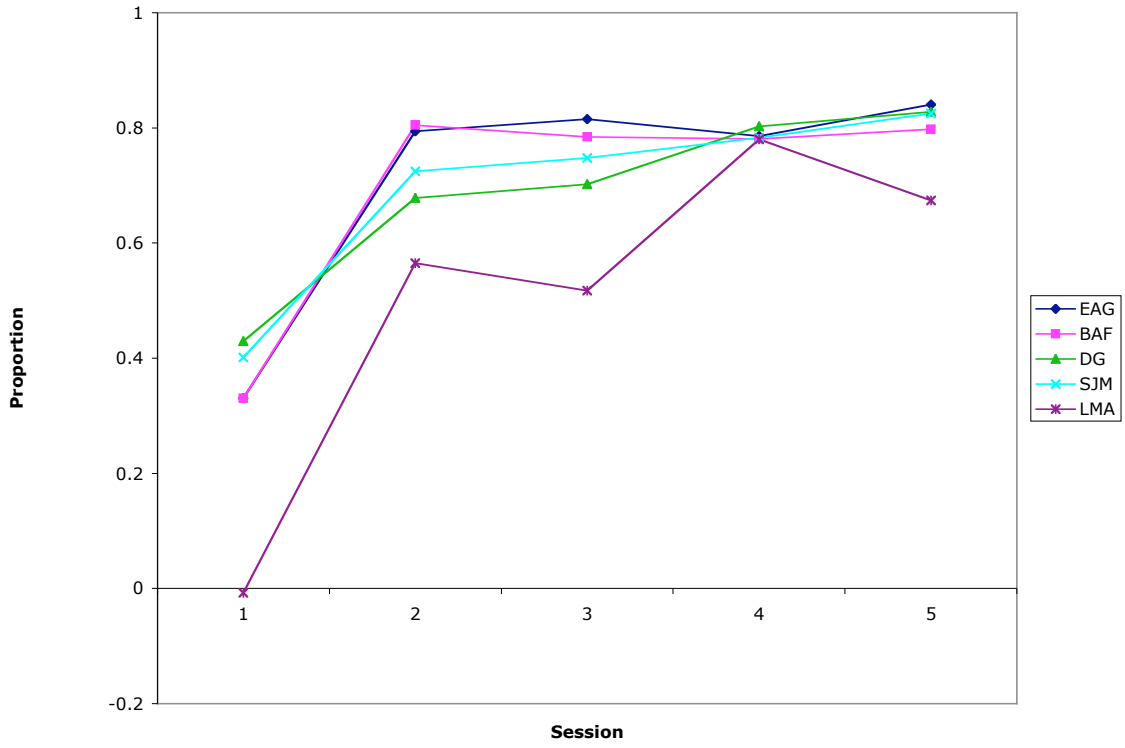


Figure 8.2 Experiment 1, participants scores as a proportion of the maximum score. Presented by session.

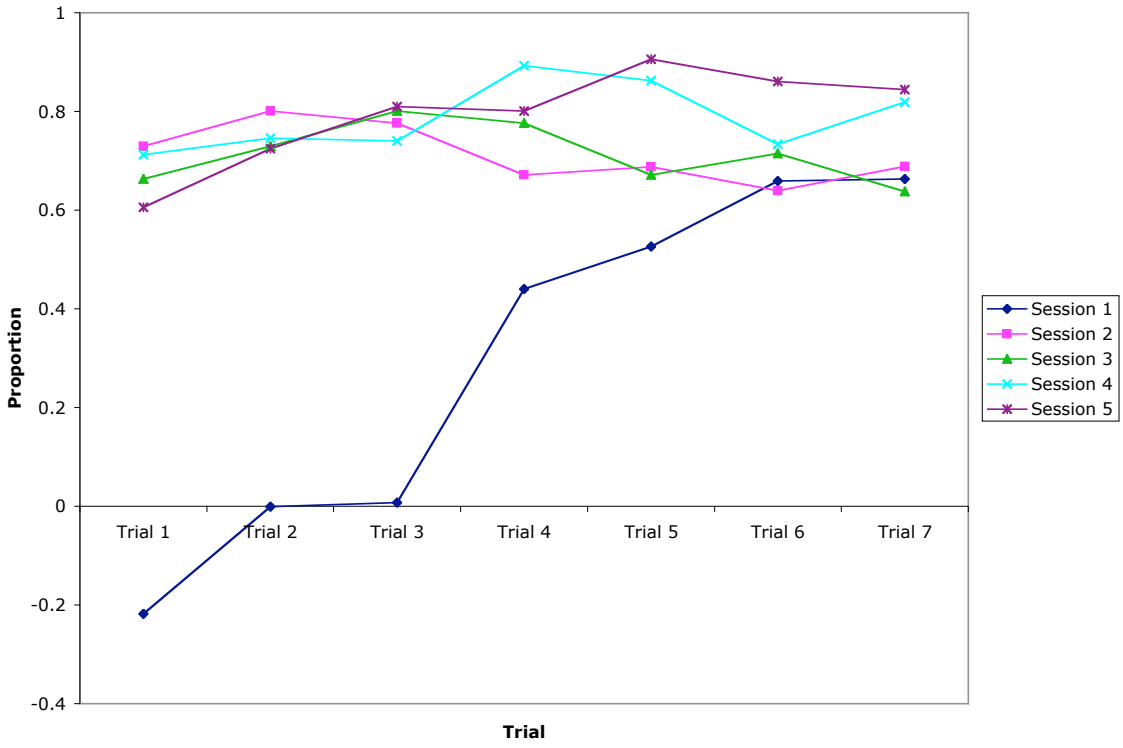


Figure 8.3 Experiment 1, participant scores as a proportion of the maximum score, presented by trial and session.

8.2 Proportion of Attention Allocated to the Four Quadrants

Due to an unforeseen error in the data collection software, the eye-movement data from the first session was lost. Hence the following analyses only represent the final four sessions of the experiment.

In order to examine the attention allocation patterns of participants, and in order to make comparisons with the predicted attention allocation patterns of the models, the proportion of time in which a participant's point-of-gaze was directed at each of the four quadrants of the screen, as recorded during the experiment, was examined in some detail. In the following figures in this section, the observed proportion of gaze durations (i.e.,

total time) directed by participants to each of the four quadrants of the screen is presented alongside one or two basic model predictions. As noted previously, the optimal sampling strategy of participants, if neither effort to sample nor salience of the visual stimuli are factors in how they allocate their attention, should be based on the product of the alarm frequency and value of each of the dials (AF*V), where alarm frequency is measured as the number of alarms per second on any dial and value is the point value associated with any dial. Using AF*V as the optimal criterion, the proportion of time participants should spend sampling any region should be directly related to the relative values of AF*V as a proportion of the sum AF*V for all four dials. Also presented in some instances is the product of the bandwidth and value associated with each dial (BW*V). Though not the optimal sampling criterion, BW*V is expected to be an influence on participants' sampling strategies, and it is shown here for comparison purposes.

The average (and standard deviation) of the proportion of time during which participants' gaze was directed at each of the four quadrants is shown in table 8.1.

	Upper-Right	**Lower-Right**	**Lower-Left**	**Upper-Left**
Average	0.15	0.32	0.28	0.25
Std. Dev.	0.01	0.03	0.03	0.03

Table 8.1 Average and standard deviation of proportion of time during which participants' gaze was directed at each of the four quadrants.

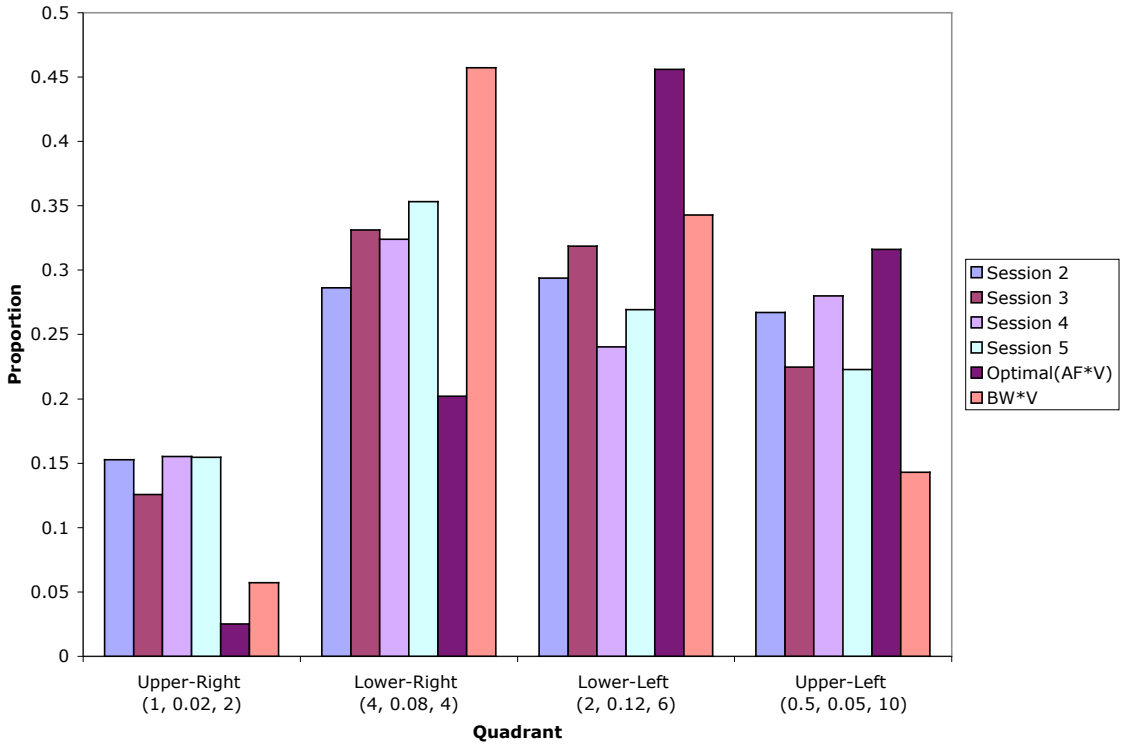


Figure 8.4 Average proportion of gaze duration by quadrant and session for all participants. Note that listed below the name of each quadrant are the bandwidth (radians/second), alarm frequency (alarms/second), and point value of the respective dials.

As anticipated, participants devoted different proportions of attention to each of the four quadrants, as evidenced by a reliable main effect of quadrant, $F(3, 12) = 7.52$ $p < 0.01$. Because the proportion of attention devoted to each quadrant must sum to 1.0 across the four quadrants, a significant main effect of session would not be found in the data. However, analyses did reveal a significant session by quadrant interaction, $F(9, 36) = 3.00$ $p < 0.01$, indicating that the proportion of attention devoted to each quadrant varied as a function of the experimental session.

Because we anticipated that there would be differences in the sampling strategies of individual participants, we examined the sampling patterns of each of the five participants. A qualitative examination of their individual patterns was quite revealing and is discussed here.

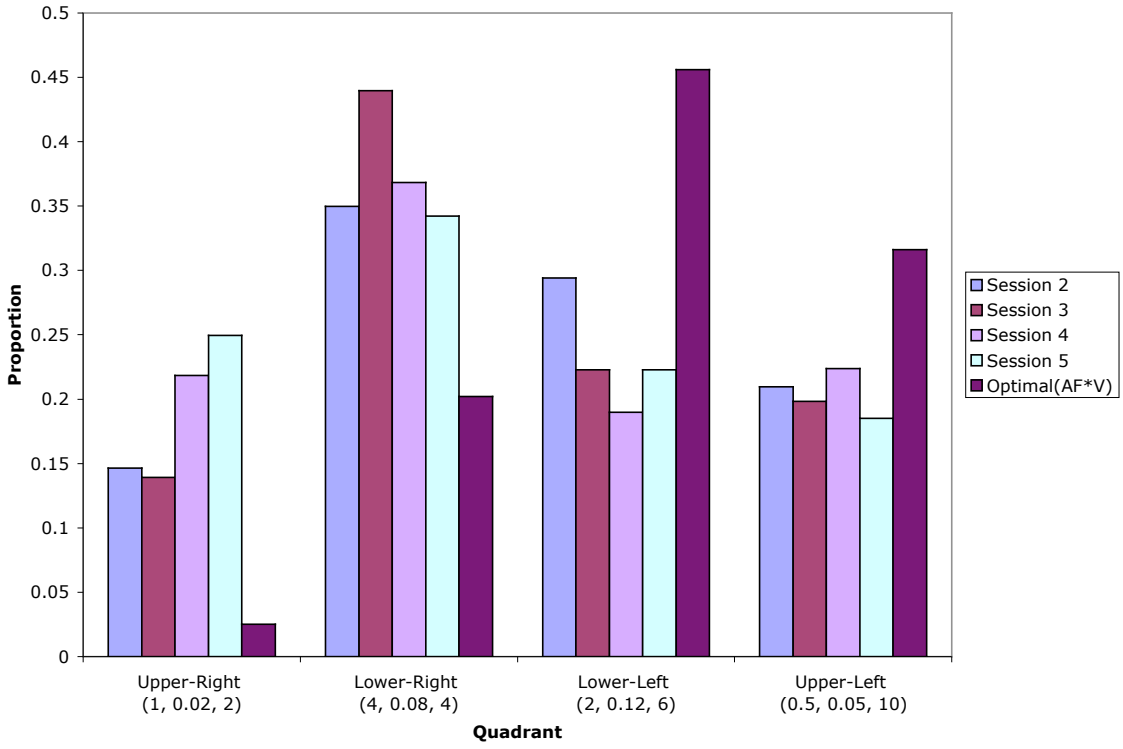


Figure 8.5 Participant 1 (BAF), Proportion of Gaze Duration on Four Quadrants, by Session.

By the final experimental session, participant BAF had settled on a strategy in which he sampled each of the dials at approximately the same frequency, with slightly more attention to the lower-right (LR) dial. This resulted in a strong tendency to

oversample (relative to the optimal criterion) the upper-right (UR) and LR dials, and undersample the LL and UL dials.

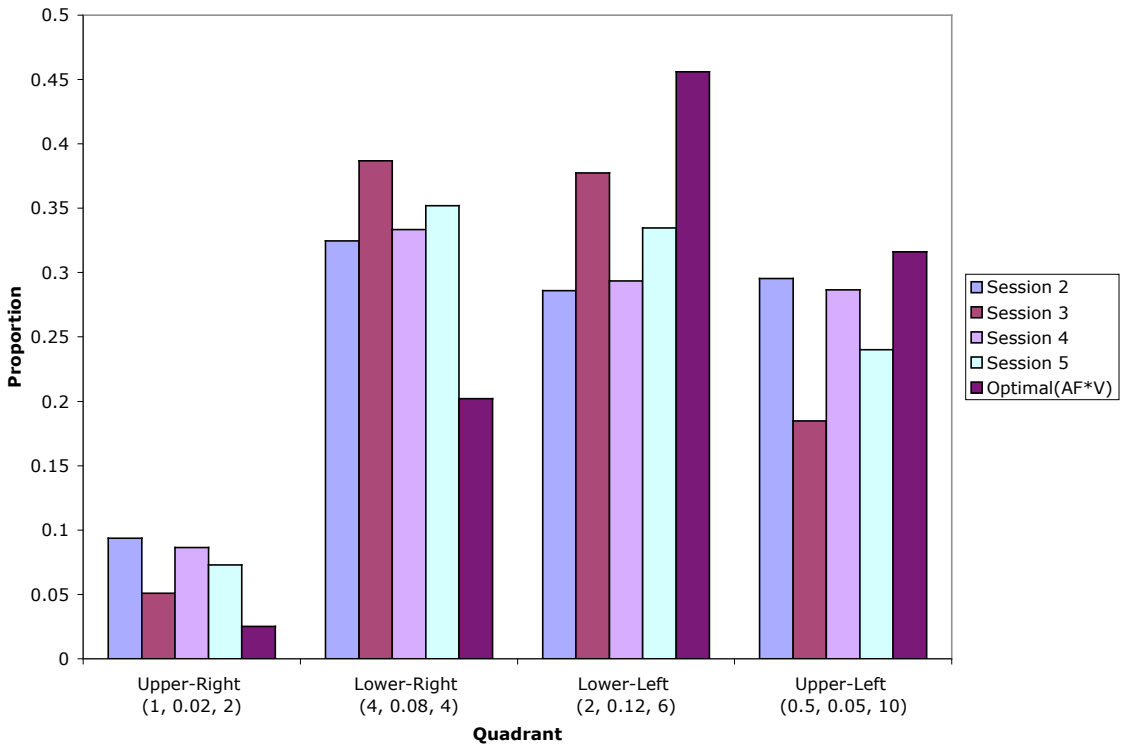


Figure 8.6 Participant 2 (DG), Proportion of Gaze Duration on Four Quadrants, by Session.

Participant DG showed a sampling pattern consistent with the general trend of that predicted by the optimal AF*V model. Her largest discrepancy from the predicted patterns was on the LR dial, which she exhibited a tendency to oversample.

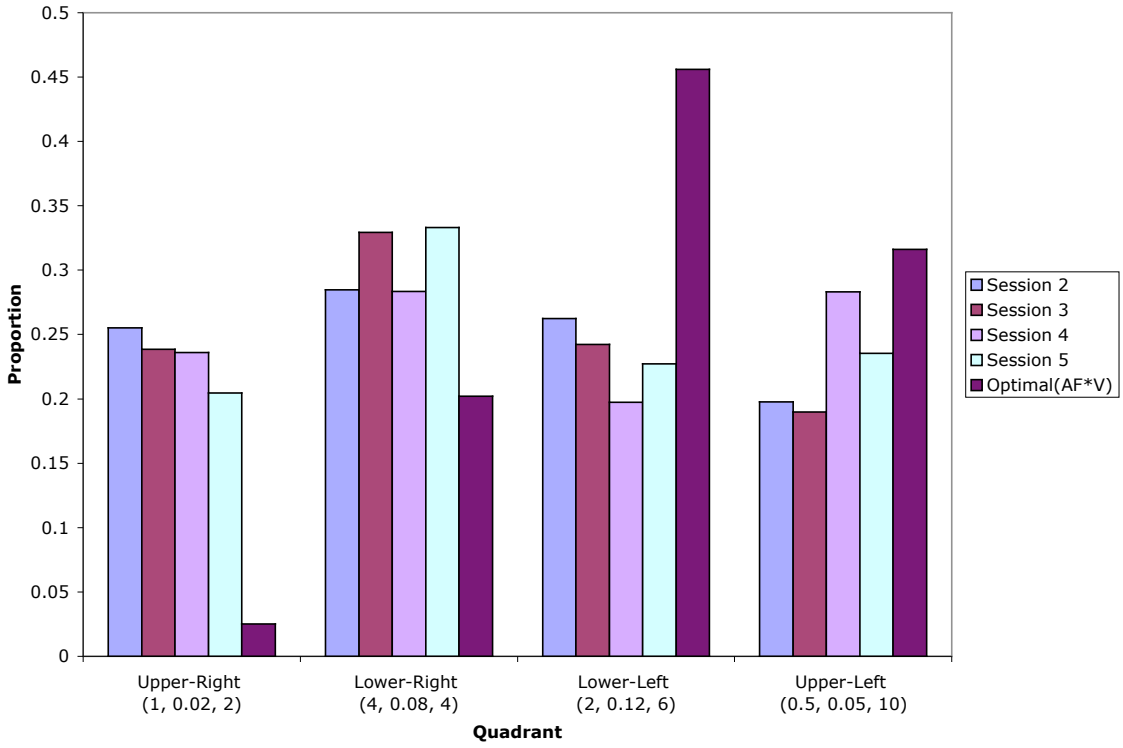


Figure 8.7 Participant (EAG), Proportion of Gaze Duration on Four Quadrants, by Session.

Participant EAG did not show much variation in his sampling pattern across the four dials, i.e., each dial was sampled at approximately the same frequency.

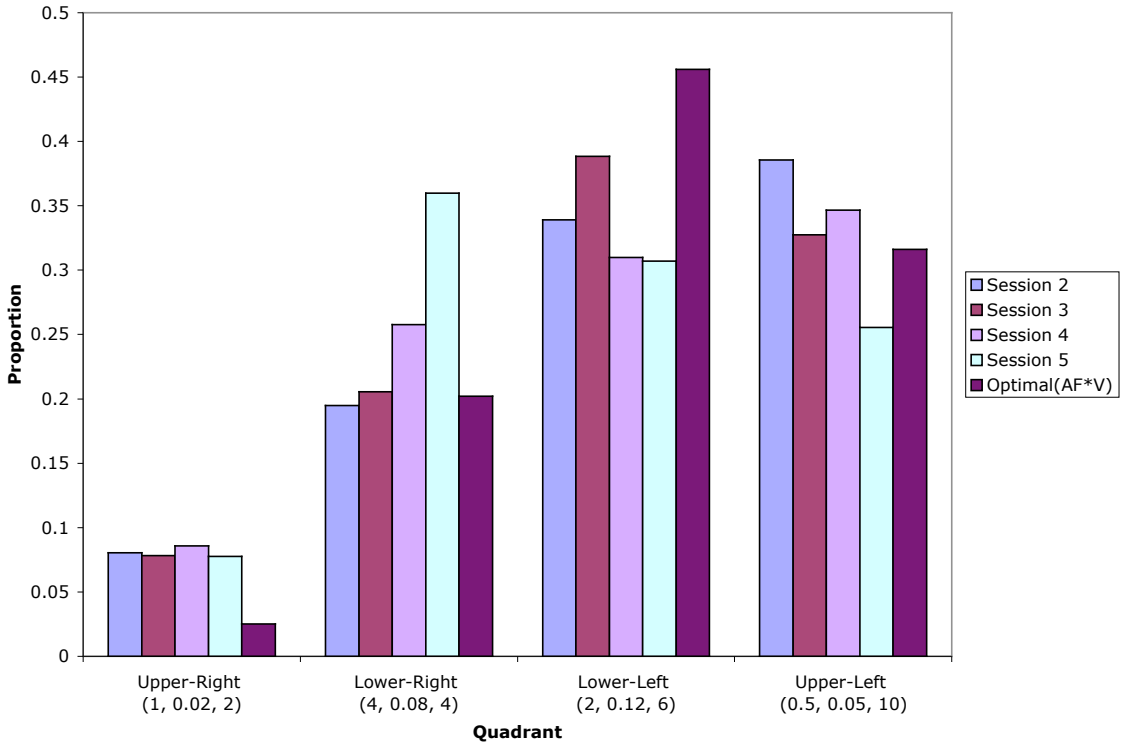


Figure 8.8 Participant 4 (LMA), Proportion of Gaze Duration on Four Quadrants, by Session.

Participant LMA exhibited a sampling pattern generally in accordance with the AF*V optimal model, although she showed a consistent tendency to oversample the upper-right dial and undersample the lower-left dial.

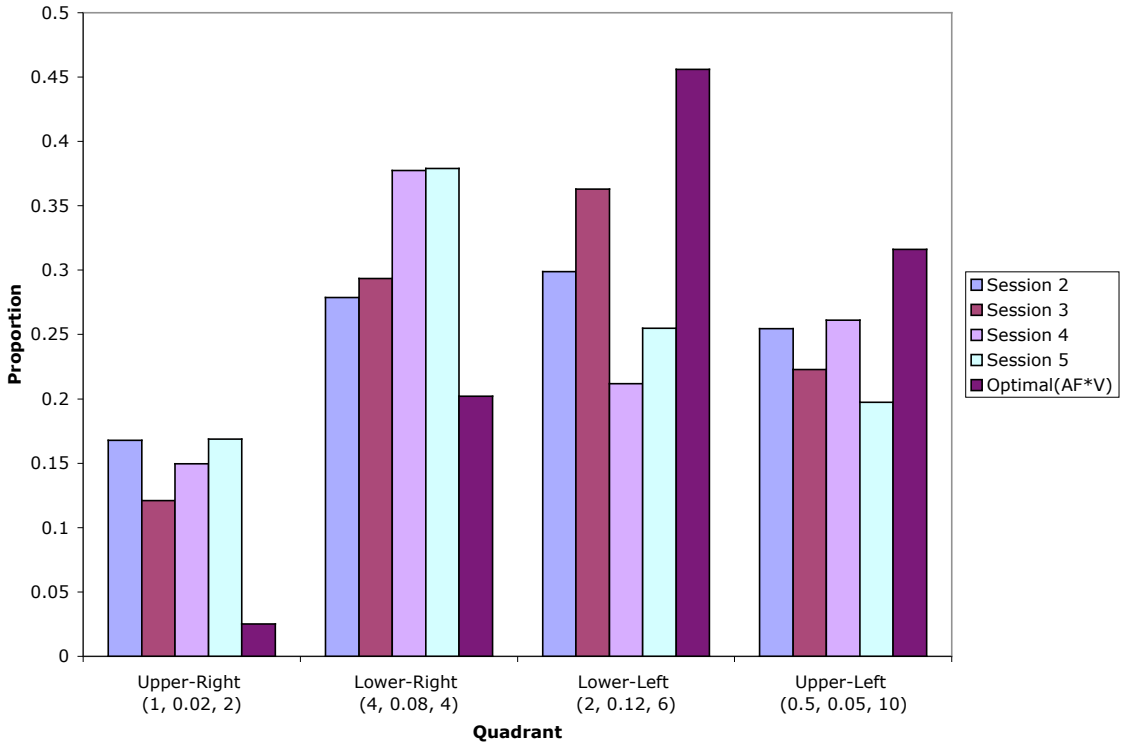


Figure 8.9 Participant 5 (SJM), Proportion of Gaze Duration on Four Quadrants, by Session.

The sampling pattern exhibited by participant SJM exhibited the same general pattern as predicted by the optimal criterion, but she had a tendency to oversample the LR dial and undersample the LL dial on the latter two sessions.

8.3 Dwell Duration Analysis

The average dwell duration of participants, as a function of quadrant, was examined in a similar manner as the proportion of gaze duration by participants as a function of quadrant (Figure 8.10).

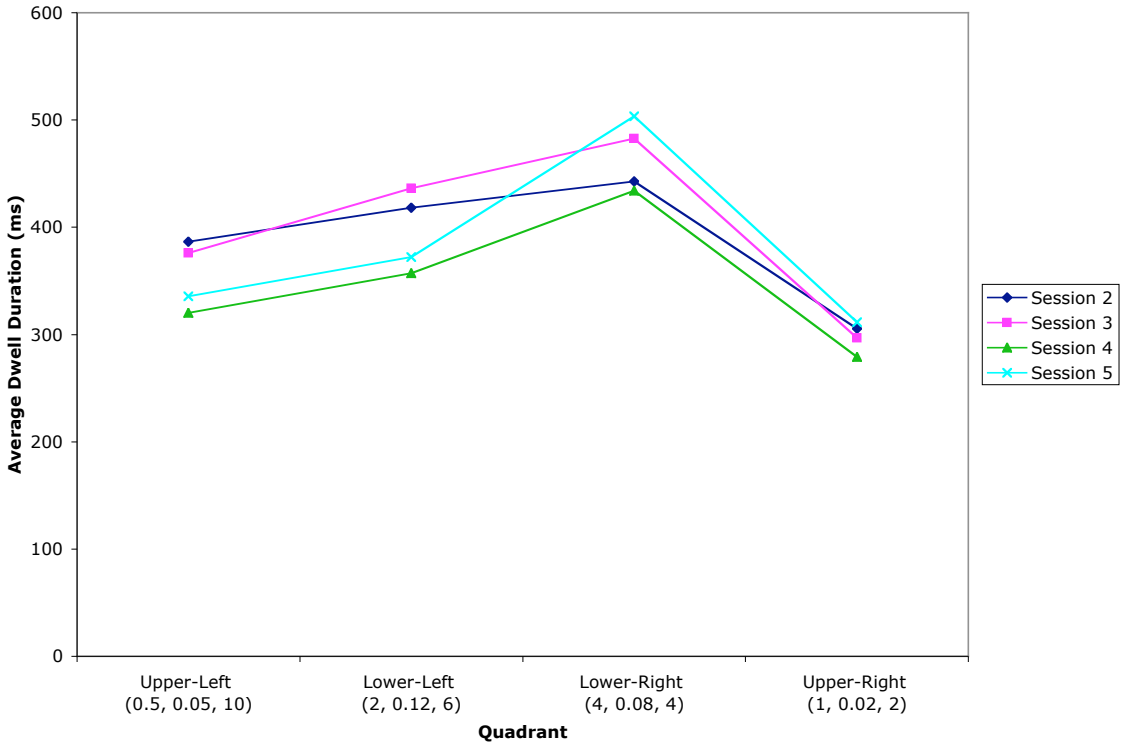


Figure 8.10. Average dwell duration by experimental session and quadrant.

Analysis of the average dwell duration of participants (the average amount of time that the point-of-gaze remained within a quadrant before moving to another quadrant) revealed the average dwell duration varied by quadrant, $F(3, 12) = 12.01$, $p < 0.001$. Additionally, the dwell duration varied by session such that participants had shorter average dwell times in the latter sessions than in the earlier sessions, $F(3, 12) = 3.67$, $p < 0.05$. Analyses did not reveal a session by quadrant interaction, $F(9, 36) = 0.79$, $p = 0.69$, indicating that the pattern of relative dwell durations associated with each region remained relatively constant across the experimental sessions.

Examining the patterns of dwell duration across the four quadrants for each individual participant indicated that except for LMA, the general pattern of relative

durations holds across participants (Figure 8.11). For each of the other four participants, the relative dwell duration associated with each region is ordered such that the lower-left region had the longest average dwell duration, the upper-left region the lowest, and the upper-right region the second lowest dwell duration.

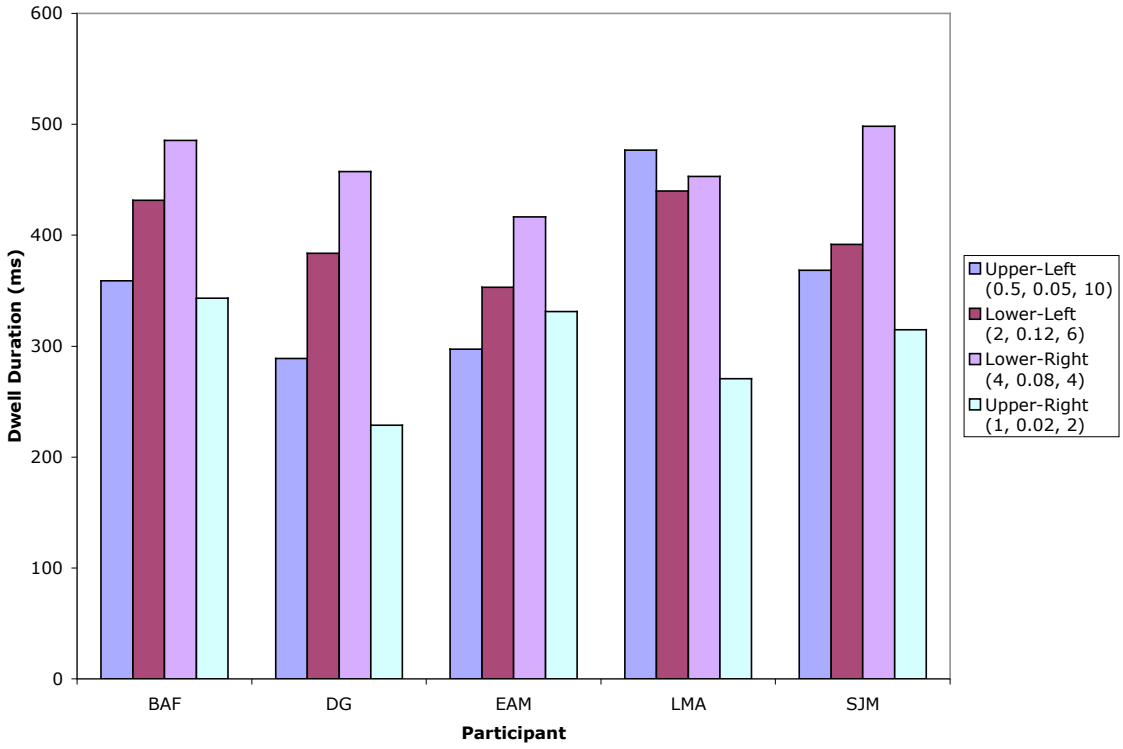


Figure 8.11 Average dwell duration by participant and quadrant.

8.4 Correspondence between Observed Sampling Patterns and Model Predicted Sampling Patterns

In order to examine the degree to which each of the model predicted sampling patterns (i.e., the proportion of attention devoted to each quadrant) correspond to the observed sampling patterns of participants, we examined the proportion of variance explained in the participant data by each of several different models.

Examining the proportion of variance explained by each of the models parsed by experimental session, but collapsed across all five participants (Figure 8.12), it is apparent that the AF*V model did not explain a great deal of variance in the sampling data, with an average R^2 of 0.40 across the four experimental sessions. Similarly, the BW*AF*V model did not explain a large proportion of the variance in the experimental data. The three models that explained a high proportion of the variance in the data were both versions of the IFT models and the (V*(AF+BW)/2) model.

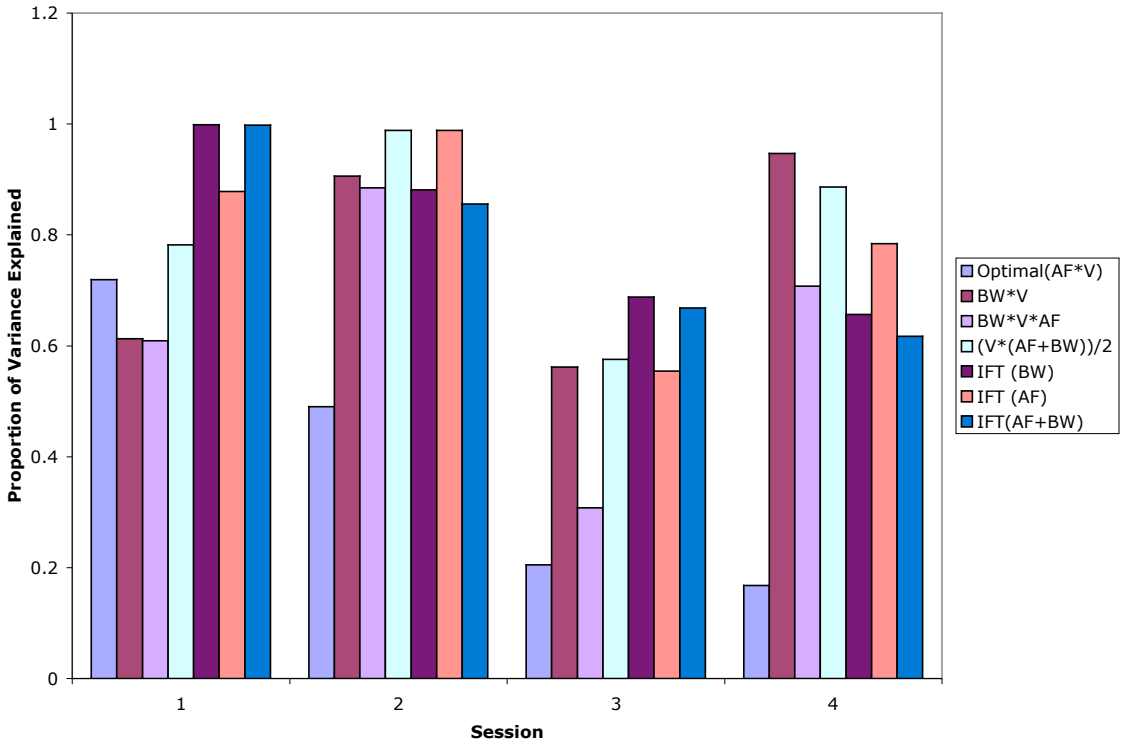


Figure 8.12 Proportion of variance explained in observed sampling patterns explained by seven different theoretical models.

Model	AF*V	BW*V	BW*AF*V	V*…/2	IFT (BW)	IFT (AF)
R^2	0.39	0.85	0.70	0.90	0.88	0.86

Table 8.1 R^2 for each of six models relative to the average gaze patterns of all participants across all experimental sessions.

The proportion of variance explained by each of the models reflects the vast individual differences in sampling patterns found across the group participants.

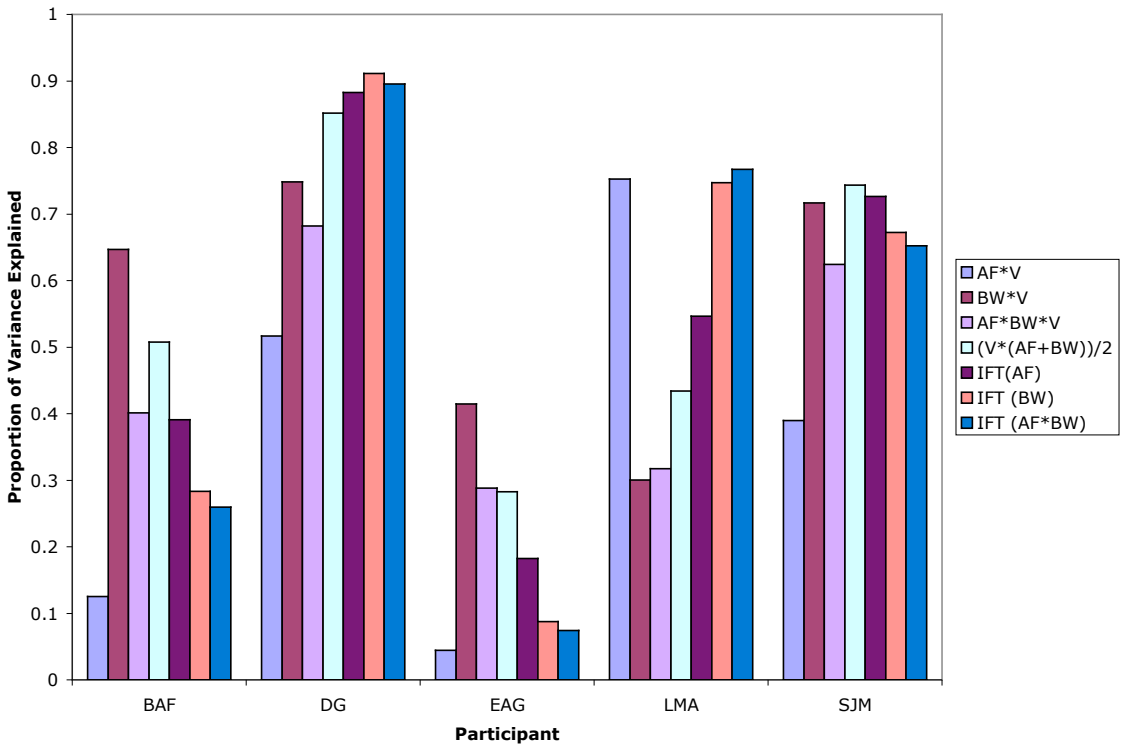


Figure 8.13 Proportion of variance explained by model and participant.

Examining the fit of each model to the data by participant (Figure 8.13) clearly reveals the capability of the models to adequately describe and predict the attention allocation patterns of some participants and an inability to predict the patterns of others.

Specifically, participants 1 and 3 do not correspond well to any of the models. In contrast, participants 2, 4, and 5 all exhibit sampling patterns in accordance with some of the models, particularly the information foraging models.

8.5 Discussion of Experiment 1 Results

The scoring patterns from Experiment 1 suggest that much of the learning of the task occurs over first eight trials, at least in terms of the scoring criterion. After trial 8, the average of all five participants was nearly identical to that of the five participants on the final trial of the experiment (Figure 8.3). Unfortunately, without the eye movement data from the first experimental session, which was lost due to a technical problem, it cannot be determined if there were substantial changes in the sampling patterns of participants over this time. The following two experiments, particularly Experiment 3 provide sufficient data to evaluate this question in some detail.

The average dwell durations of participants were, in order from shortest to longest, associated with the upper-right, upper-left, lower-left, and lower-right regions. This ordinal pattern does not correspond to any order described by AF, BW, or V alone. However, it corresponds well with the ordinal ranking of dials based on the product of AF and BW (Table 8.2). One hypothesis for such an ordinal pattern is that it may be an indication that participants sampling patterns are dependent on the conditional properties of the display. Specifically, it appeared that participants would often spend more time looking at a dial that was near to an alarm region, relative to one that was equidistant from the two alarm regions. If this were the case, then all other things being equal, we would expect to find that dials with the highest instance of “near alarms” (which would include the set of real alarms, as they were once near alarms also) would be associated

with the longest dwell durations. We did not measure the instance of near alarms in the experiment; however, we can assume that it is a product of AF and BW. Dials with a high AF would, by definition, be near the alarm region a relatively high proportion of time. And dials with a high BW would also be expected to approach the alarm regions more frequently, simply by their nature of having a higher rate of movement. The hypothesis that dwell duration is a function of these two elements is investigated further in the subsequent experiments. However, it should be noted that this observation, that participants may be sensitive to BW and AF in conjunction provides some credence for a theoretical model that encompasses both factors.

Dial	BW (radians/second)	AF (alarms/sec)	V (Points)	AF*BW (seconds)
Upper-Right	1	0.02	2	0.02
Lower-Right	4	0.08	4	0.32
Lower-Left	2	0.12	6	0.24
Upper-Left	0.5	0.05	10	0.25

Table 8.2 Bandwidth (BW), alarm frequency (AF), point value (V), and product of AF and BW for each of the four dials in Experiment 1.

One notable aspect of the experiment data was that the scoring patterns of participants did not show much difference across participants, yet their sampling patterns did. This is an indication that the task may not have been difficult enough to force participants to adapt to the statistical properties of the display. More specifically, the performance of participants may have been under constrained by the task, in that they could adopt a wide variety of sampling strategies and still perform well at the task.

One aspect revealed in the experiment analysis that speaks directly to the under constrained nature of the task was that two participants that had very good scores (the highest and third highest average scores over the course of the experiment), both exhibited sampling strategies that were poor fits to optimal sampling models. Fortunately, participants were interviewed at the end of each experimental session regarding their sampling strategies. In the case of both participants, after the second session, they described what they believed to be statistical properties of display which are not revealed in the models. Specifically, both of them were looking for (and believed they had found) relationships, or correlations, between the dials. Essentially, they were looking for cues on any of the dials, i.e., patterns of the movement of the needle, that might provide some information as to when to look at another dial and which dial to look at. For example, Participant 1, BAF, described one pattern as, "When the lower-right dial go like this [indicates approaching the left alarm region], then suddenly jump [sic] back over here [right alarm region], then I know it's not going to go [into alarm region] but that one of these two is [indicates upper and lower-right dials]."

As discussed in Section 6, the movements of the needles on each of the dials were derived such that they appeared random to observers. However, because of the computing time needed to derive patterns that met the experimental conditions (alarm frequency, bandwidth, and ecological validity), only four patterns were recorded, and these were "replayed" to participants twice during each experimental session. This exact methodology had been employed successfully by Miller and Kirlik (2004). However, it is possible that some participants may have been able to remember the characteristics of the four recorded patterns and were able to recognize and take advantage of the remembered

patterns in subsequent trials. When asked, participants did not admit that they recognized any specific patterns. However, this does not necessarily indicate that they were unable to do so.[3] If it is the case that these participants were able to recognize such patterns, then it is likely that their sampling patterns would not be reflected in the predictive sampling models, as no such correlations between dials are accounted for by any of the models.

[3] Their inability to recall and describe any particular pattern does not necessarily indicate that they were not able to recognize a pattern when it occurred. For one, recognition is often superior to recall in memory tasks. Second, knowledge of such patterns and how to respond to them is likely stored as procedural knowledge, which often cannot be accurately verbalized, relative to declarative knowledge.

9. Experiment 2

Experiment 2 was designed to explore the hypothesis that the effort of acquiring information via the eyes influences an observer's patterns of visual attention allocation. The pairings of each dial with an associated bandwidth, alarm frequency, and point value remained identical to those in Experiment 1.

Experiment 2 consisted of two experimental stages. In the first stage, the task of the participant and the experimental manipulations remained identical to those of Experiment 1. This stage lasted for a total to ten trials, all of the six trials from the first session and half (4) of the trials from the second testing session. The purpose of this stage was to allow the participants to gain some familiarity with the properties associated with each dial, specifically their bandwidth and alarm frequency. A duration of 10 trials was chosen for this stage based on the scoring from Experiment 1. The scores of participants in Experiment 1 had begun to "plateau" at approximately 8 to 10 trials, depending on the participant (see the Results section of Experiment 1 for more detail).

In the second condition, the basic task remained the same, accumulating points by detecting alarms. However, in this version of the task, only one of the four dials was visible at any time (see Figure 9.1.1 for an example of the experimental screen). Participants selected which dial they wanted to be visible by pressing the key corresponding to a particular dial. For instance, when they wanted to view the upper-left dial, they would press the "A" key. In order to indicate that an alarm had occurred on a dial, they pressed the key corresponding to the visible dial. For instance, when the upper-left dial was visible, pressing the "A" key indicated that they had detected an alarm on that dial. Hence, they could only indicate alarms on the dial that was currently visible.

Pressing a different key, other than the key corresponding to the visible dial, would switch the visible dial to that dial associated with the recently-pressed key. For instance, pressing the "Z" key when the upper-left dial was visible caused the view to change such that the lower-left dial became visible and the upper-left dial was occluded from view. The second experimental stage occupied the remaining 3.5 experimental sessions (3.5 hours) remaining after stage 1.

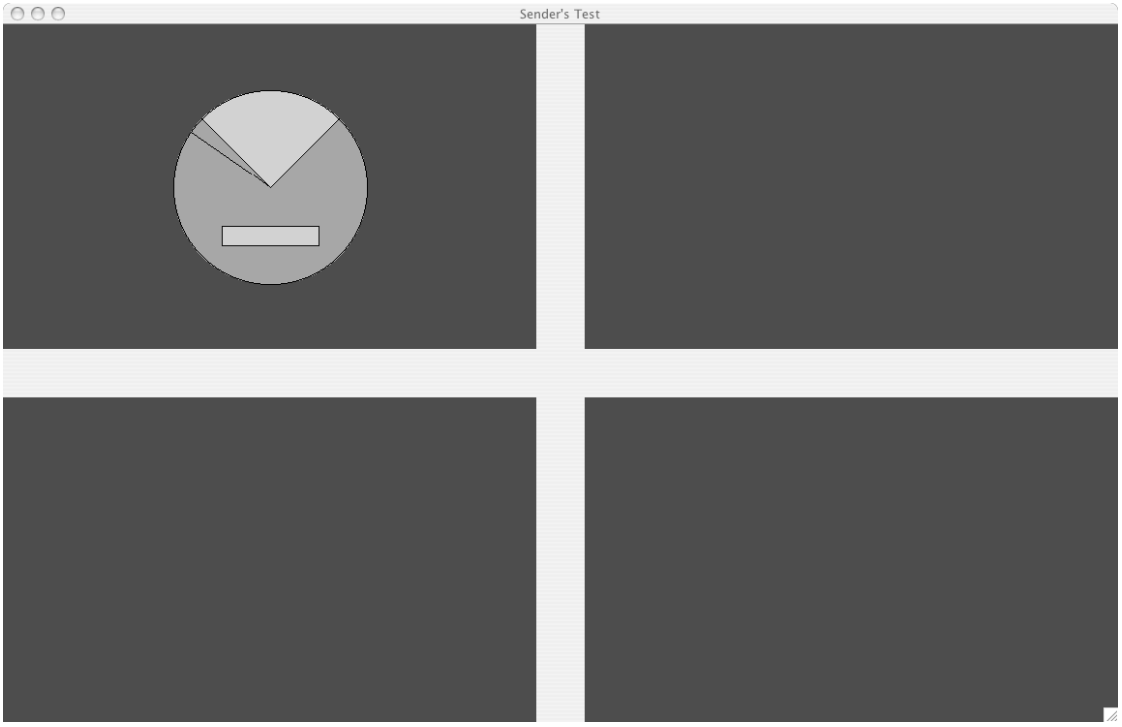


Figure 9.1.1. The experiment screen, second stage of Experiment 2, one-dial-visible condition. An alarm has occurred on the visible dial.

In order to manipulate the "effort" required to view a dial in a measured, systematic manner, we manipulated the time delay between when a participant indicated that they would like to view a new dial (pressed a key corresponding to a dial other than

the currently visible dial) and when the new dial became visible. The time delay for each dial is presented in Table 9.1.1.

Dial	BW (radians/second)	AF (alarms/sec)	V (Points)	Delay (seconds)
Upper-Right	1	0.02	2	0.0
Lower-Right	4	0.08	4	1.0
Lower-Left	2	0.12	6	0.25
Upper-Left	0.5	0.05	10	0.5

Table 9.1.1 Bandwidth, alarm frequency, point value, and time delay for each dial.

9.1 Experiment 2 Scoring Results

The results from Experiment 2 are divided into two sets of analyses corresponding to the two stages of the experiment. Recall that for the first ten trials of the experiment (all of session 1 and half of session 2), the task and experimental design was identical to that of Experiment 1. For the latter 28 trials (the latter half of session 2 and sessions 3, 4, and 5), the task of participants remained the same, accumulating points by detecting alarms. However, in the second stage of the experiment, only one of the four dials was visible at any moment, and participants had to choose which dial they wanted to view. The delay between when they selected a dial and when it became visible was manipulated. Once again, because participants received negative points for both false alarms and misses, they often had negative scores during the early portion of the experiment.

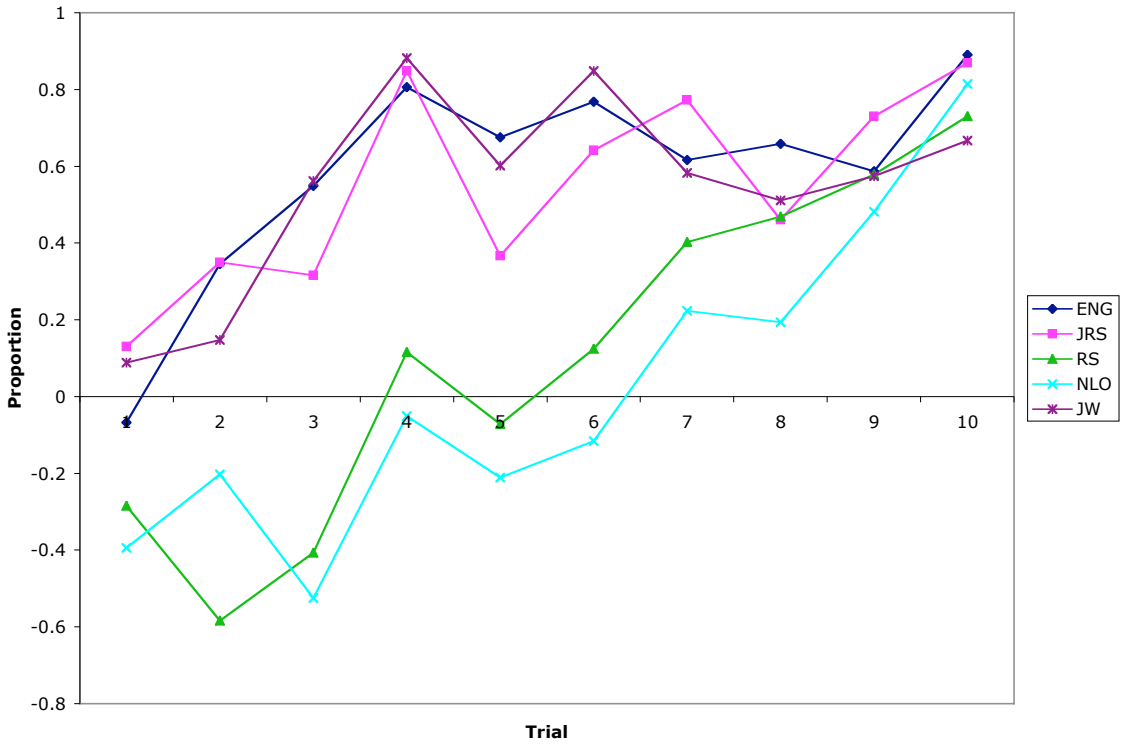


Figure 9.1 Scores as a proportion of maximum on the first 10 trials (Stage 1) of Experiment 2.

Examination of the participants' scores on the first stage of Experiment 2 revealed a reliable effect of trial, $F(9, 36) = 9.65$, $p < 0.01$, with Greenhouse-Geisser correction, indicating that participants' scores improved over the course of the stage (Figure 9.1). This improvement in performance was revealed to be linear in nature, as evidenced by a significant linear effect of trial, $F(1, 4) = 21.26$, $p = 0.01$. Note that two participants, RS and NLO, initially show poorer performance on the task than the other three participants, but their performance has reached that of the others by trial 10. Hence, each participant moved onto the second stage of the experiment having achieved approximately an equivalent level of performance at the task.

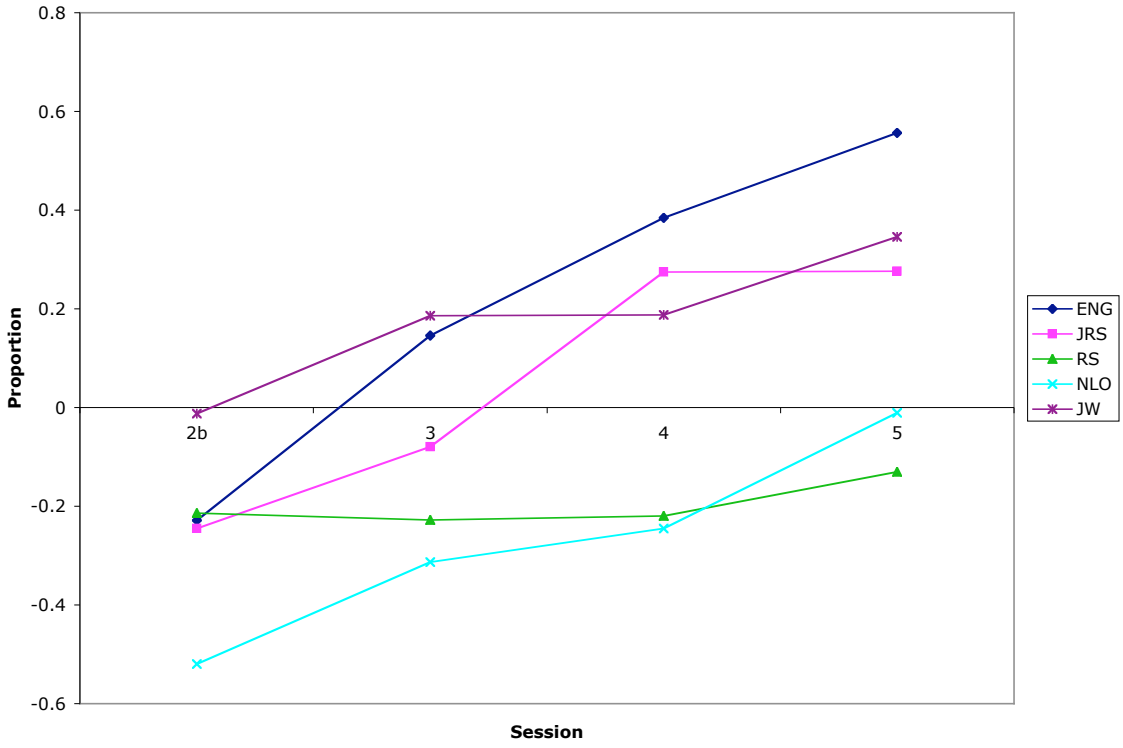


Figure 9.2 Average scores as a proportion of maximum by session and participant for the second stage of Experiment 2.

Examination of the participants' scores on the second stage of the experiment revealed there was a reliable effect of trial, $F(6, 24) = 2.49$, $p = 0.05$, indicating that participants' scores improved over the course of the experiment (Figure 9.2). Also revealed was a reliable main effect of session, $F(3, 12) = 8.94$, $p < 0.01$, again indicating that the participants' performance improved over the course of the four sessions (Figure 9.2). This improvement in performance was linear, as evidenced by a significant linear effect of trial, $F(1, 4) = 12.80$, $p = 0.02$. In addition, the analyses revealed a trial by session interaction, indicating that the improvement in performance from trial to trial was contingent upon the experimental session, $F(18, 72) = 3.44$, $p < 0.001$. More specifically,

the linear rate of improvement varied across the four sessions, as indicated by a linear session by linear trial effect, $F(1, 4) = 12.36$, $p = 0.03$ (Figure 9.3; note that Session 1 refers to the initial stage of the experiment, and is considered separately in the statistical analyses.)

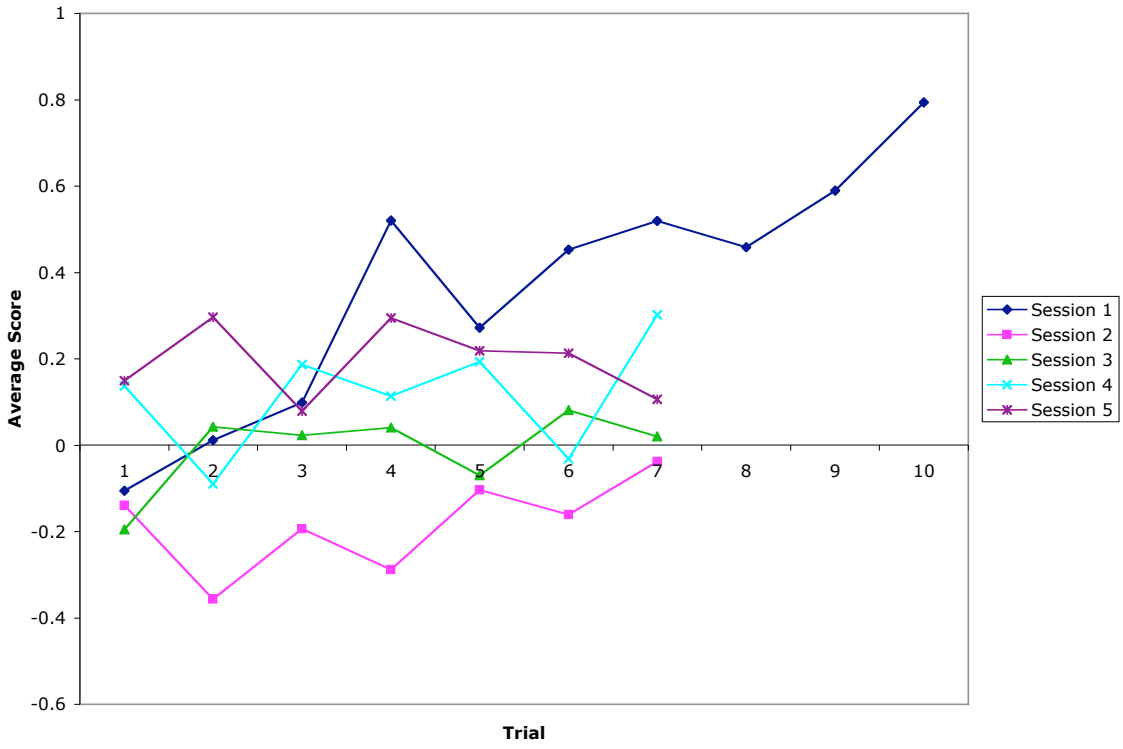


Figure 9.3 Average score as a proportion of maximum by session and trial.

9.2 Dwell Analysis

The same general patterns of behavior, in terms of dwell duration, that were found in Experiment 1 were also found in the first stage of Experiment 2. The average dwell duration varied by quadrant, $F(3, 12) = 8.82$, $p < 0.01$, and reflects the same pattern of relative durations as was found in Experiment 1, i.e., the lower-right dial was associated with the longest average dwell duration, the lower-left the second longest, and so on. Unlike Experiment 1, the analyses of the first stage of Experiment 2 did not reveal an

effect of session, nor an interaction between session and quadrant. This may be due to the reduction in statistical power resulting from only comparing dwell durations across two sessions.

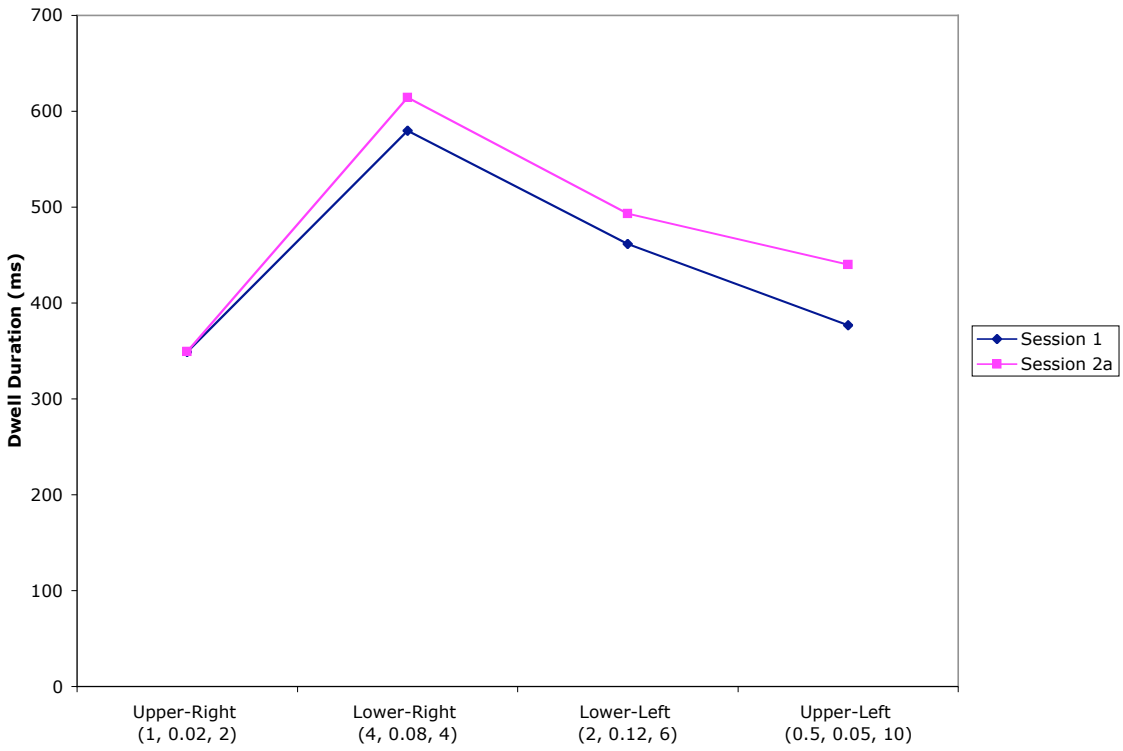


Figure 9.4. Average dwell duration by quadrant, averaged over session 1 and the first stage of session 2.

Dial	BW (radians/second)	AF (alarms/sec)	V (Points)	Delay (seconds)	BW*AF
Upper-Right	1	0.02	2	0.0	0.02
Lower-Right	4	0.08	4	1.0	0.32
Lower-Left	2	0.12	6	0.25	0.24
Upper-Left	0.5	0.05	10	0.5	0.25

Table 9.1 Bandwidth, alarm frequency, point value, delay time, and product of alarm frequency and bandwidth associated with each dial in Experiment 2.

In the second stage of Experiment 2, the average length of the dwell durations increased more than three-fold, from under 500ms to over 1500ms, although the relative durations across the four quadrants maintained the same general ordinal pattern, with the longest durations on the lower-right quadrant and the second-longest durations on the lower-left quadrant.

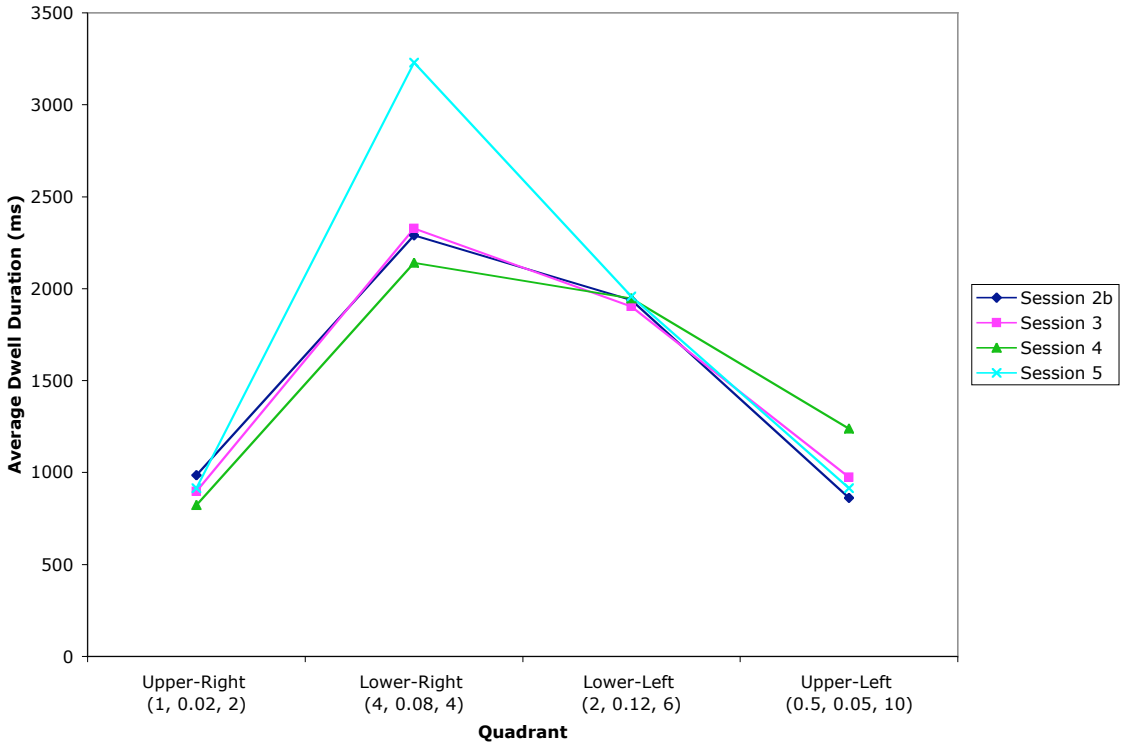


Figure 9.5 Average dwell durations on each quadrant for each session of the second stage of Experiment 2.

Analysis of dwell durations in the second stage of Experiment 2 yielded similar results to those found in stage one. Participants had reliable different dwell durations

across the four quadrants, $F(3,9) = 8.56$, $p = 0.03$, with Huynh-Feldt correction. No effect of session nor a session by quadrant interaction was found. Examination of dwell durations across the four quadrants, when parsed by the different participants (Figure 9.6), indicates a great degree of variation in the duration patterns of participants. This is most noticeable with respect to participants, JRS and JW, who had very low dwell durations on the upper-right quadrant relative to the other participants. This is likely due to differences in sampling strategy, which is examined further in the next section.

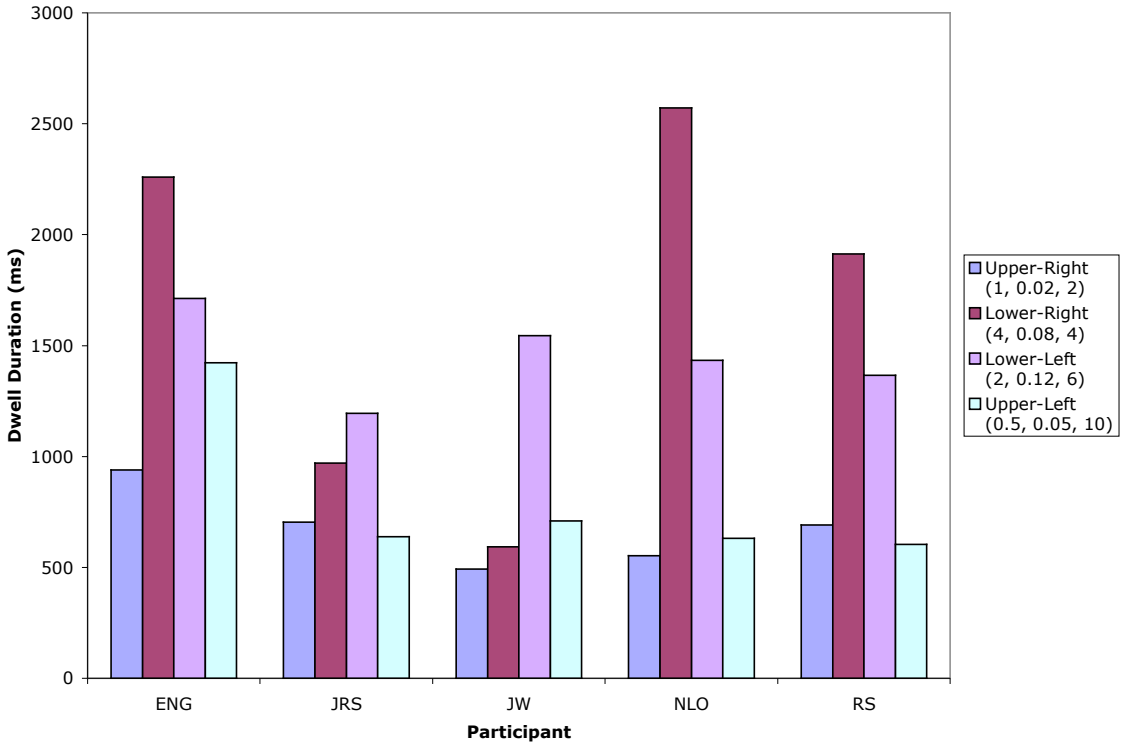


Figure 9.6 Average dwell durations in Stage 2 of Experiment 2 by participant and quadrant.

9.3 Proportion of Attention to the Four Quadrants

The attention allocation patterns of participants, which are assumed to be indicative of their sampling strategies, are examined here in some detail. As in the results section of the previous experiment, the attention allocation patterns of participants are shown alongside the predictions of the two most basic models, AF*V and BW*V.

In Table 9.2, the average proportion of attention devoted to the four quadrants in session 1 and the first half of session two are presented. Note that the proportions to each of the quadrants are almost identical for the two sessions, but that the average score of participants on the task differs by nearly a factor of three.

	Quadrant				Average
	UL	**LL**	**LR**	**UR**	**Score**
Session 1	0.11	0.42	0.33	0.14	98.77
Session 2a	0.10	0.41	0.33	0.16	278.75

Table 9.2 Average proportion of attention devoted to the two quadrants for the first session and the first half of session 2, which was a replication of the first ten trials of experiment 1. The last column shows the average score of participants for the two sessions.

The average proportion of time devoted by the five participants to the four quadrants in the second stage of the experiment is presented in Figure 9.7. As anticipated, participants devoted different proportions of attention to each of the four quadrants, as evidenced by a reliable main effect of quadrant, $F(3,12) = 8.69$ $p < 0.01$. However, analyses did not reveal a significant session by quadrant interaction, $F(9,36) = 1.26$, $p = 0.29$, providing no evidence that the proportion of attention devoted to each quadrant varied as a function of the experimental session.

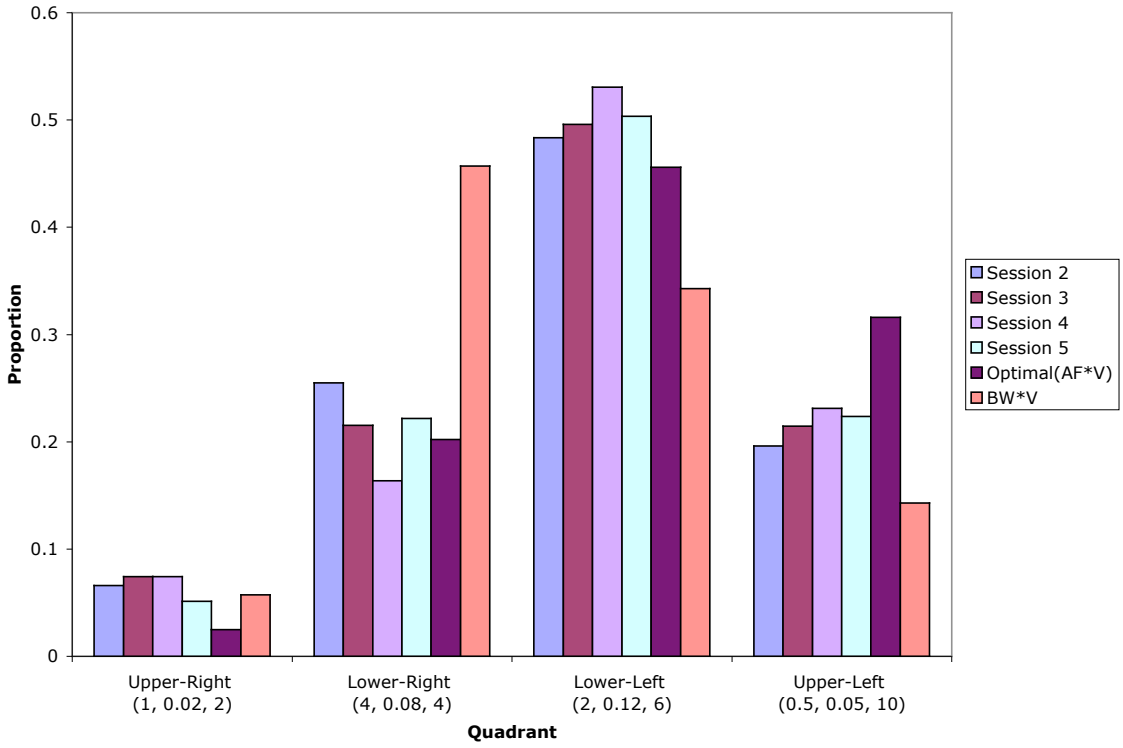


Figure 9.7 Average proportion of gaze duration by quadrant and session, collapsed across all participants.

As in Experiment 1, we anticipated that there would be differences in the sampling strategies of individual participants. The sampling patterns of each of the five experiment participants are examined at a qualitative level subsequently.

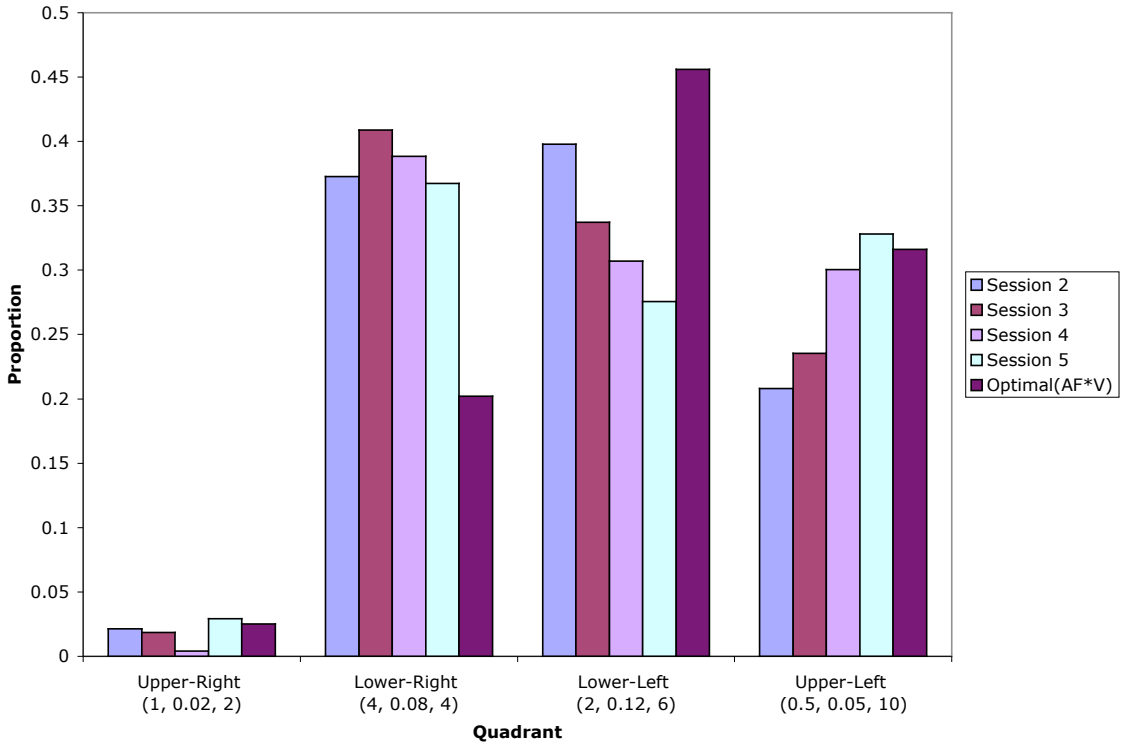


Figure 9.8 Participant 1, ENG, proportion of gaze duration on four quadrants by session.

The sampling pattern of Participant ENG display a moderate fit to the optimal criterion, in the sense that the proportion of attention paid to the upper dials is quite close to optimal, and the proportions paid to the lower dials are less so.

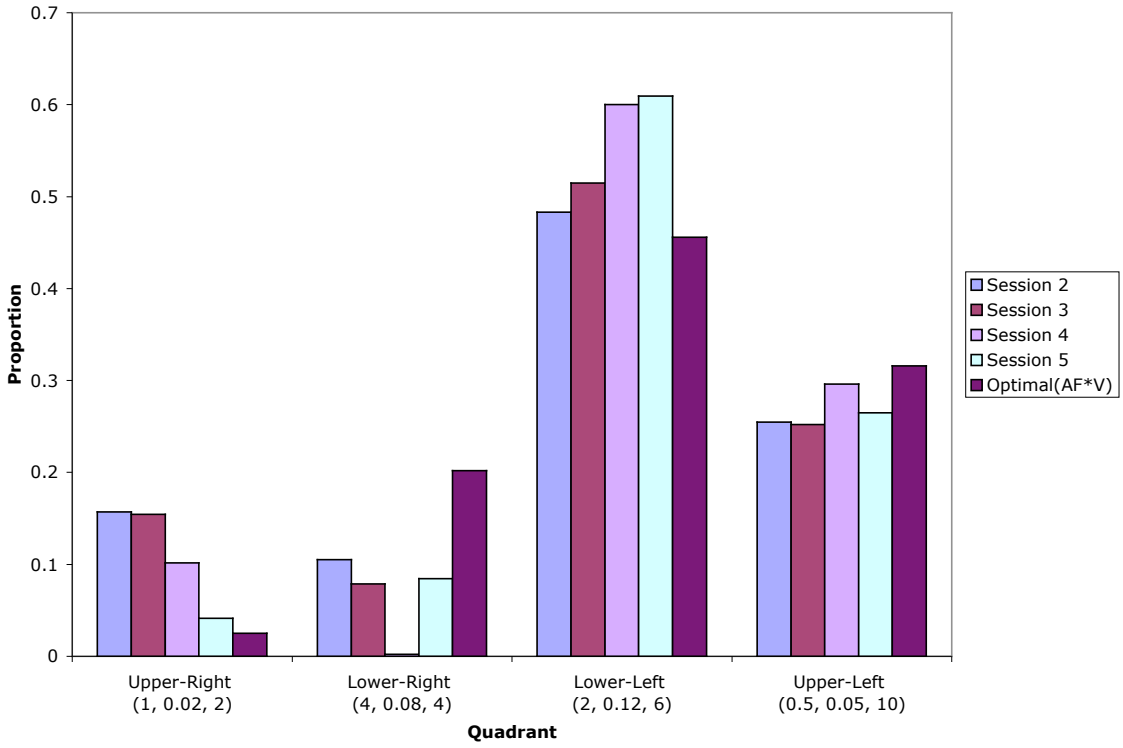


Figure 9.9 Participant 2, JRS, proportion of gaze duration on four quadrants by session.

Participant JRS displayed a tendency to undersample the lower-right and upper-left dials, but matched the general trend of the AF*V model quite well.

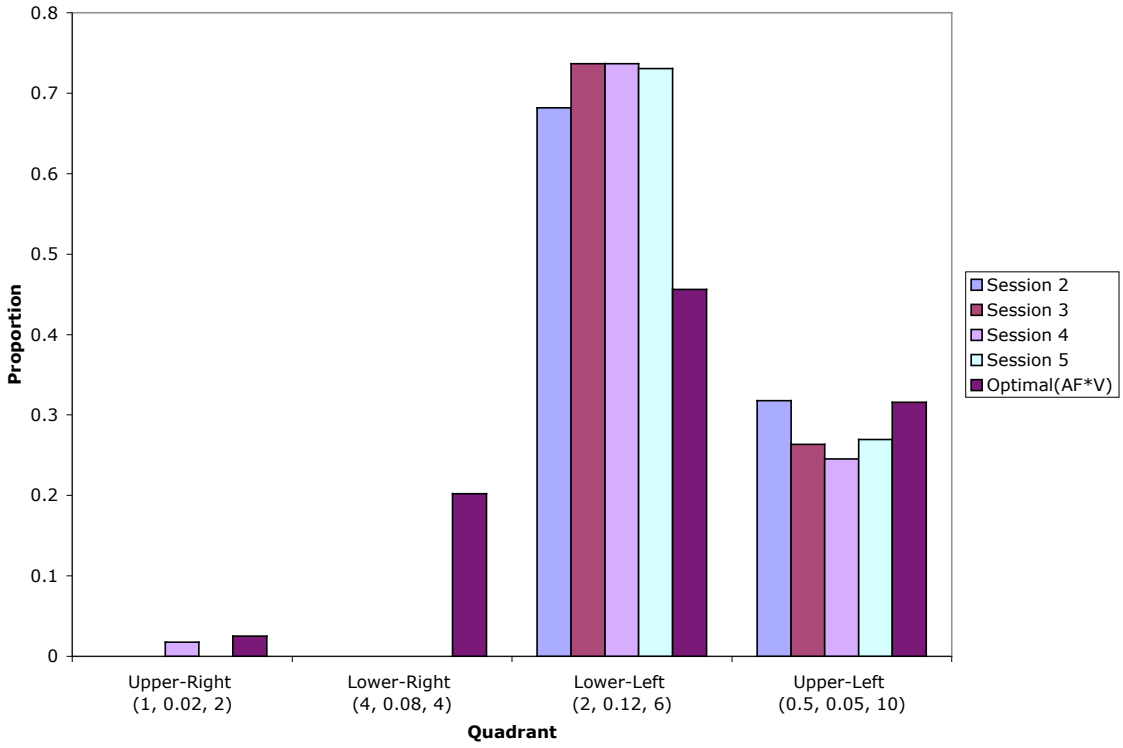


Figure 9.10 Participant 3, JW, proportion of gaze duration on four quadrants by session.

Participant JW chose a different sampling strategy than all of the other participants. He chose to focus solely on the left two dials and almost entirely ignored the dials on the right side of the display. Interestingly, this placed much lower cognitive demands on the participant and still allowed him to score at a decent rate (second highest average score of the five participants). Also it is not greatly discordant with the optimal criterion.

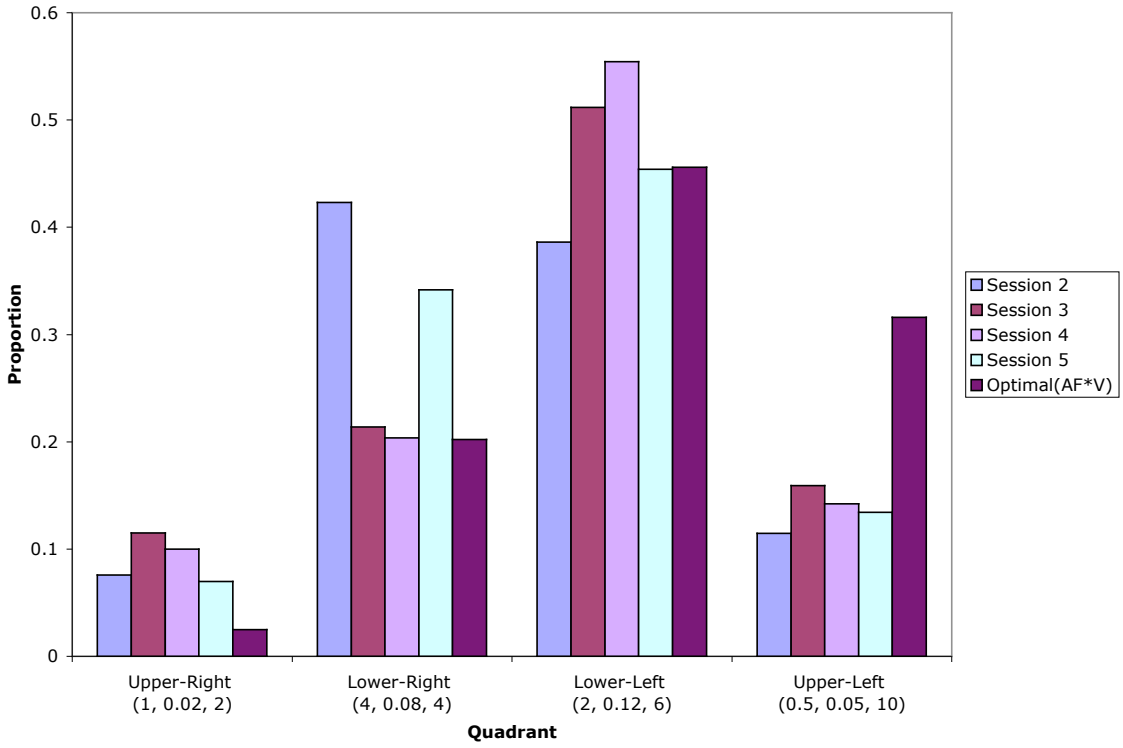


Figure 9.11 Participant 4, NLO, proportion of gaze duration on four quadrants by session.

Participant NLO displayed a high degree of correspondence to the optimal criterion except on the upper-left dial, which she under sampled.

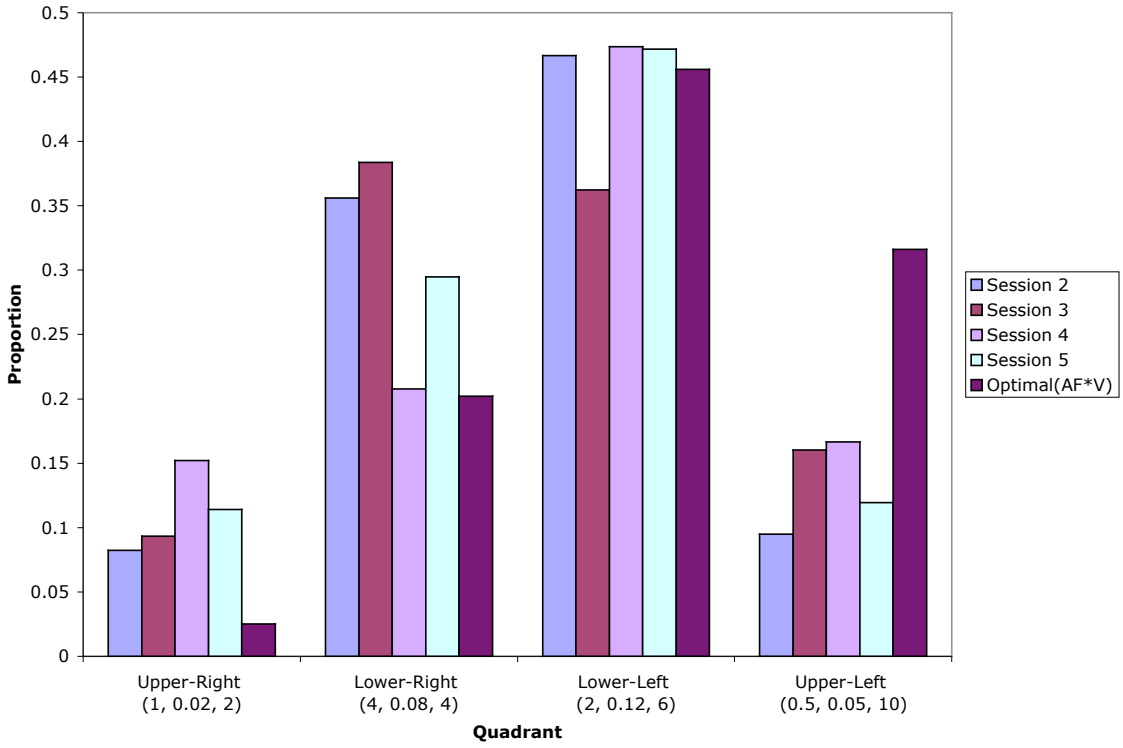


Figure 9.12 Participant 5, RS, proportion of gaze duration on four quadrants by session.

The pattern of attention allocation displayed by participant RS is very similar to that displayed by participant NLO, strong correspondence to the optimal criterion except for a systematic undersampling of the upper-left dial.

9.4 Proportions of Variance Explained by Models

As in Experiment 1, examining the fit of each model to the data by participant (Figures 9.13 and 9.14) clearly reveals the capability of the models to adequately describe and predict the attention allocation patterns of some participants and an inability to predict the patterns of others. Unlike Experiment 1, in which some of the participants exhibited sampling patterns that did not correspond well with any of the theoretical

models, all of the participants in Experiment 2 exhibited patterns that conformed well to at least one the models. However, there was little agreement among participants as to which model most adequately predicted their attention allocation patterns.

Model	AF*V	BW*V	BW,AF*V	V*…/2	IFT (BW)	IFT (AF)	IFT(AF,BW/2)
R^2	0.86	0.30	0.60	0.49	0.52	0.81	0.91

Table 9.3 Average R^2 for each of six models collapsed across experimental sessions and participants.

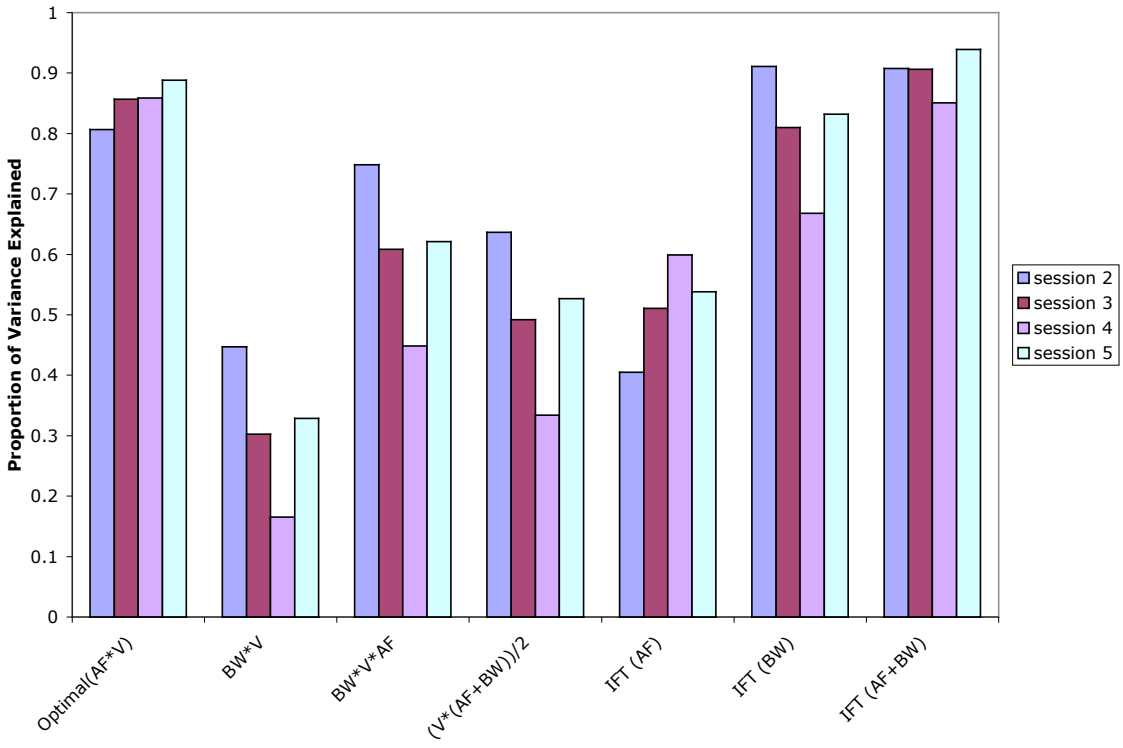
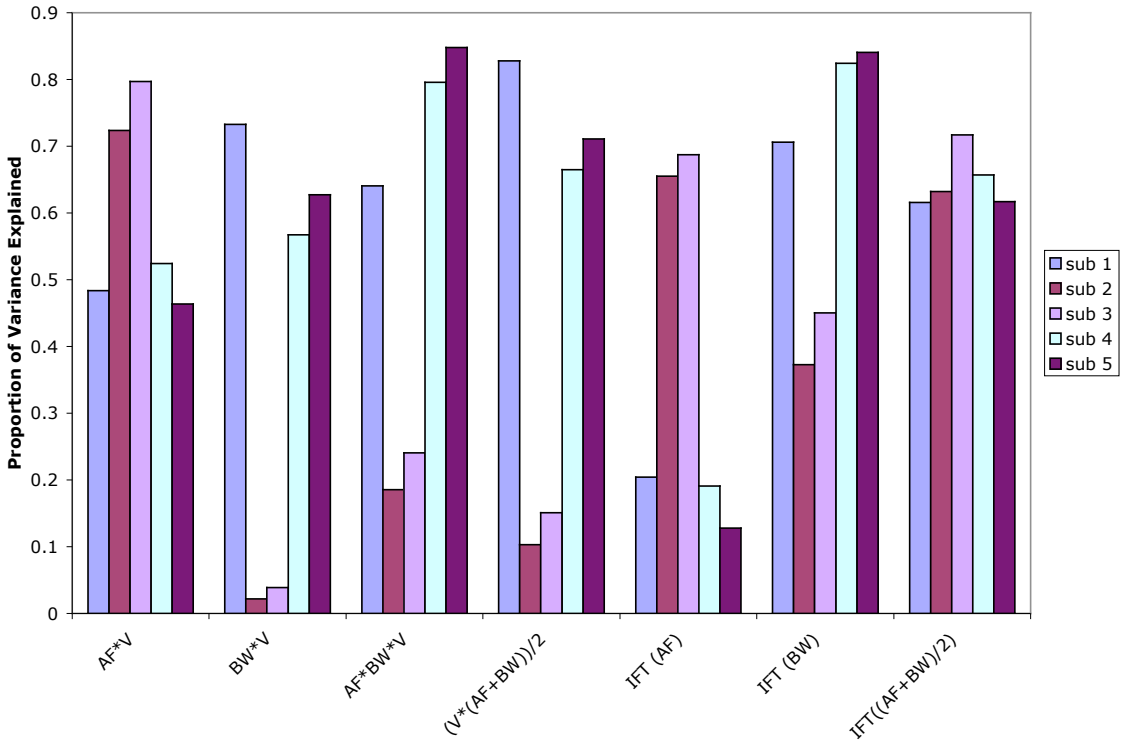


Figure 9.13 Proportion of variance explained by model for each session of Stage 2.



9.14 Proportion of variance explained in sampling pattern for each model and participant.

9.5 Discussion of Experiment 2

In Experiment 1, we were unable to examine the eye movement patterns of participants in the first session due to a loss of data. However, in Experiment 2, we have an exact replication of the first ten sessions of Experiment 2. In both experiments, the scoring patterns of participants revealed that the participants were able to learn to do the task, or at least score well at it, within ten trials. The question remained open from Experiment 1 as to exactly what it was that participants learned over the course of the ten trials. Were participants simply becoming better at the task of detecting alarms and pressing the correct button associated with an alarm, or were they adapting their sampling strategies to the statistical properties of the display? Evidence from this experiment suggests that they were not significantly altering their sampling strategies. Despite a

difference in the average score of participants of nearly three-fold from session 1 to session 2, the sampling patterns of participants remained almost identical over the two sessions. This suggests that participants adopt a sampling strategy early on in the experimental process, and only make slight adjustments to that strategy over time. This is further suggested by the results of the second half of Experiment 2, in which a reliable trial by session interaction was not found in the data, suggesting that the sampling pattern adopted by participants did not change over the course of the four sessions.

In Experiment 1, the same ordinal relationship was found between dwell duration and the product of BW and AF. It was hypothesized that dwell duration was a function of "near alarms" and that the product of BW and AF described the number of near alarms. In the first stage of Experiment 2, this same relationship was found. However, in the second half of Experiment 2, the ordinal matching between BW*AF and dwell duration did not hold. There are several possible reasons for this. First, the task changed substantially from stage 1 to stage 2. Participants now had to incorporate the component of effort into their sampling strategies. As a result, the sampling strategies in session 2 were likely more conscious than those in stage 1, and thus may have been less sensitive to the statistical properties of the display. (The level of conscious decision making in strategy selection is discussed in greater detail subsequently.)

Also due to the manipulation of sampling effort in the task, participants had much longer dwell durations on each of the four quadrants, i.e., because of the increased cost of changing dials, participants switched dials less often. Hence, they were less likely to view a dial when it was in a near-alarm state, and as a result the influence of near alarms on sampling strategy would decrease. In sum, we would not expect the frequency of near

alarms to drive the dwell duration patterns of participants nearly as much as in the first Experiment.

In the second stage of the experiment, there were large individual differences in the proportion of attention that participants devoted to each of the four quadrants. For instance, participant JW chose to essentially sample only two of the four dials. When asked about his strategy, he responded that he believed that was the strategy that would help him achieve the highest average score at the task. Clearly, his selection of such a strategy was a conscious decision. It has been well documented that people are not optimal conscious decision makers (Moray, 1986; Rasmussen, 1983; Senders, 1983). Hence, because participants were consciously selecting sampling strategies in the second stage of Experiment 2, it may be expected that the participants in the task would not approach optimality in their sampling patterns.

The influence of less-than-optimal conscious decision making may have also been manifested in the large individual differences in the sampling strategies of the individual participants. It would appear that some participants, such as JW and JRS were quite sensitive to the alarm frequencies of the dials, as both of their patterns of attention allocation were well-described by the AF*V and IFT(AF) models. In contrast, the patterns of NLO and RS were well-described by models that incorporated BW instead of AF, indicating that these participants were heavily influenced by the rate of movement of the needles on each dial. Interestingly, the two participants who were sensitive to the BW of the dials had the lowest average scores at the task in both the first and second stages of the task, and manifested a lower learning rate at the task as depicted in Figure 9.1.

Participant ENG's sampling pattern was most closely approximated by models incorporating BW, however like two participants in Experiment 1, he explicitly noted that he was looking for correlations between the movements of the dials. Also like the two participants in the Experiment 1, his scoring average was above that of the other participants, which may indicate the he did, indeed, discover some correlations between the dials. Of course, the models assume that no such correlations exist, and hence, they do not describe his attention allocation patterns well.

10. Experiment 3

Experiment 3 replicated Experiment 1 with two exceptions. First the dial with the lowest alarm frequency was altered. This ammeter differed in that whenever an alarm state occured on it, the dial flashed (background color changed to white) for the duration of the alarm (while the needle was in the alarm region). It was assumed that this flashing would be sufficient to alert participants of the occurrence of an alarm, i.e. constitute "perfect" visual salience. As such, we expected that sampling of this dial would approach zero. The other alteration of from Experiment 1 was the point values of the dials, each dial was associated with a different point value (Table 10.1).

Dial	BW (radians/second)	AF (alarms/sec)	V (Points)
Upper-Right	1	0.02	8
Lower-Right	4	0.08	6
Lower-Left*	2	0.12	2
Upper-Left	0.5	0.05	10

Table 10.1 The BW, AF, and Vs associated with each dial. *The lower-left received the manipulation of visual salience, whereby alarms were signaled by a flashing of the dial.

10.1 Scoring Reults

The scoring data is presented as a proportion of the maximum possible score on each trial. The maximum score for each trial in Experiment 3 was 414 points, which was lower than Experiment 1 (474 points) due to different pairings of point values with different dials.

The general pattern of scoring by the participants was quite similar to that of Experiment 1, and this was confirmed by a similar set of statistical analyses. Examination of the participants' scores revealed a reliable effect of trial, $F(6, 24) = 30.67$, $p < 0.001$, indicating that participants' scores improved over the course of the experiment. Also revealed, was a reliable effect of session, $F(4, 16) = 11.61$, $p < 0.001$, again indicating that the participants' performance improved over the course of the five sessions (Figure 10.2). Much of this improvement occurred within the first session, and was revealed in a reliable session by trial interaction, $F(24, 96) = 4.50$ $p < 0.001$. (Figure 10.1). This improvement in performance was revealed to be linear in nature, as evidenced by a significant linear effect of trial, $F(1, 4) = 70.97$ $p < 0.001$, and of session, $F(1, 4) = 31.32$, $p < 0.01$.

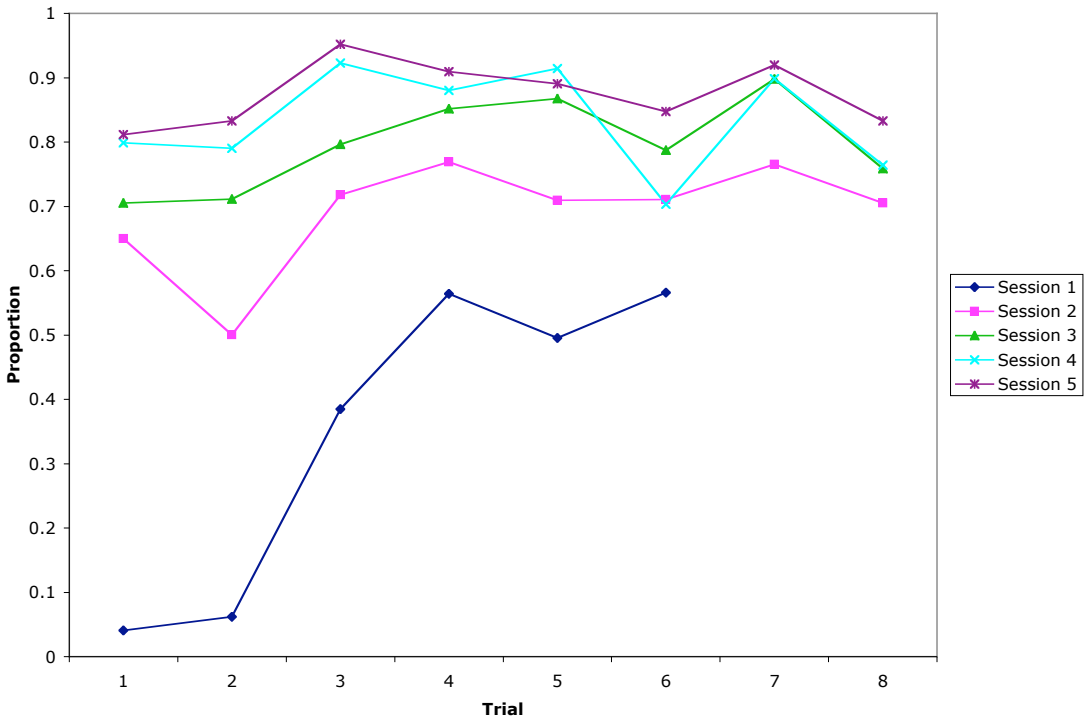


Figure 10.1 Average score as a proportion of maximum by trial and session. Session 1 consisted of six trials, while the other sessions consisted of 8 trials.

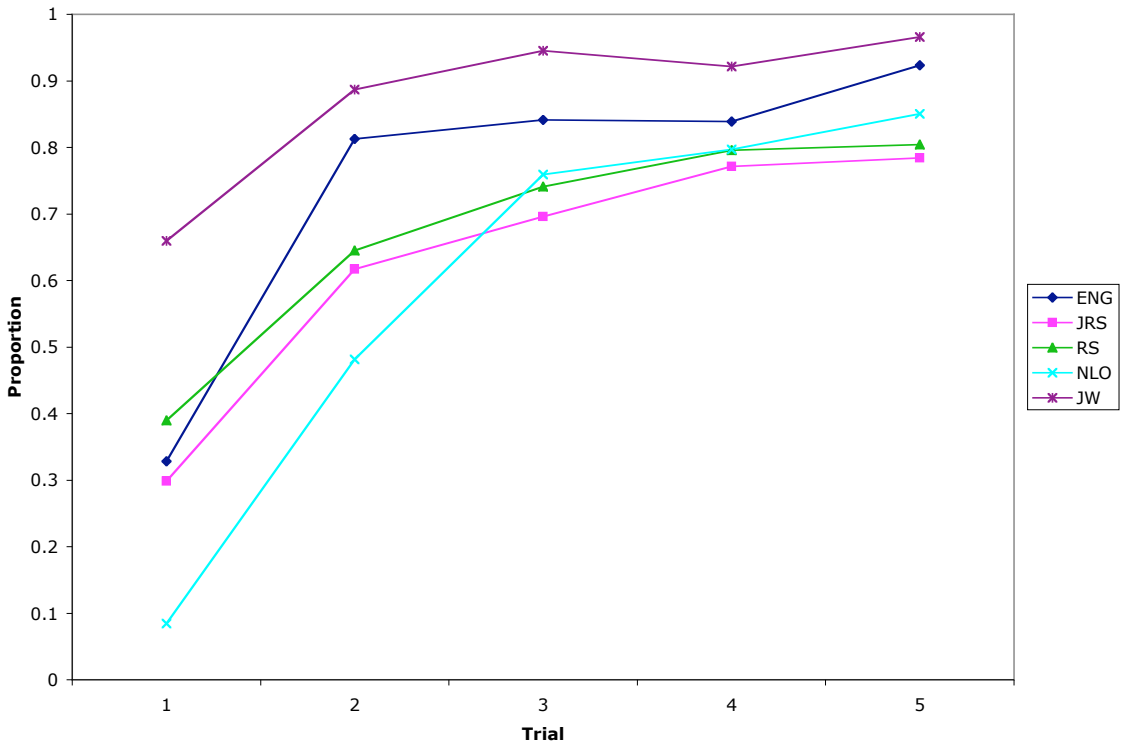


Figure 10.2 Average score by session as a proportion of maximum.

10.2 Dwell Analysis

Analysis of the average dwell duration of participants revealed that the average dwell duration varied by quadrant, $F(3, 12) = 12.17$, $p < 0.001$ (Figure 10.3). However, the analyses did not reveal a reliable effect of session, $F(4,1\ 6) = 0.78$, $p = 0.55$, nor did they reveal a reliable session by quadrant interaction, $F(12, 48) = 5.85$, $p = 0.19$ with Greenhouse-Geisser correction, indicating that the pattern of relative dwell durations associated with each region remained relatively constant across the experimental sessions.

In addition the pattern of dwell durations relative to one another remains constant across all participants. Ranging from the longest average dwell duration to the shortest

average dwell duration, the pattern for each of the participants was LL, UR, UL, LR (Figure 10.4).

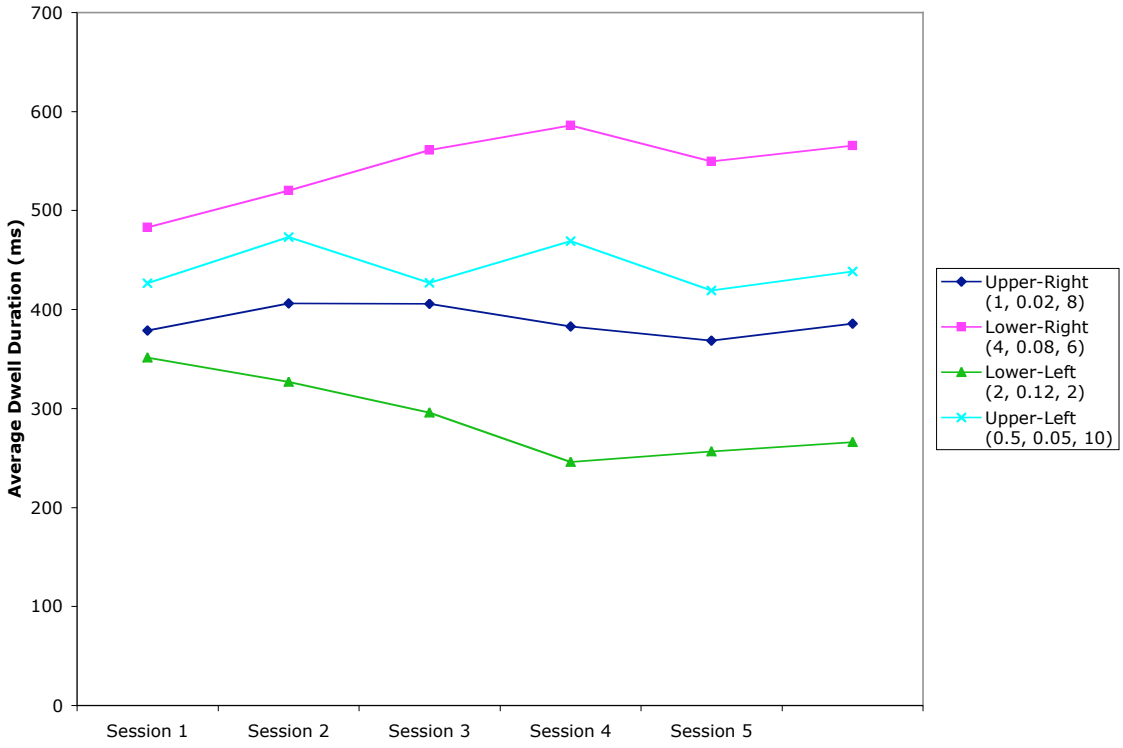


Figure 10.3 Average dwell durations by session and quadrant.

Dial	BW (radians/second)	AF (alarms/sec)	V (Points)	BW*AF
Upper-Right	1	0.02	8	0.02
Lower-Right	4	0.08	6	0.32
Lower-Left*	2	0.12	2	0.24
Upper-Left	0.5	0.05	10	0.25

Table 10.2 The BW, AF, and Vs associated with each dial. *The lower-left received the manipulation of visual salience, whereby alarms were signaled by a flashing of the dial.

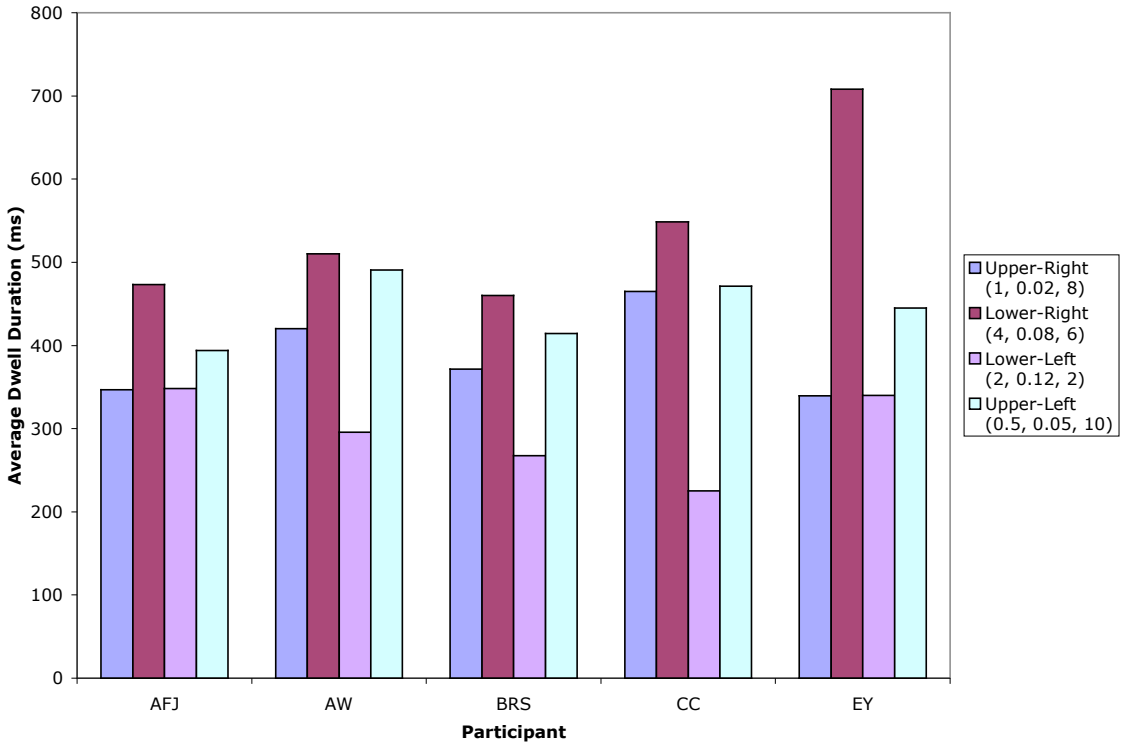


Figure 10.4 Average dwell duration on the four quadrants by participant.

10.3 Proportion of Attention Allocated to the Four Quadrants

As in the analysis of the data in the previous experiments, the proportion of time participants allocated to each quadrant is presented alongside the optimal criterion (AF*V) as a benchmark. Note that the AF*V criterion shown in the charts does not show any attention allocated to lower-left quadrant. This is the quadrant in which the visual salience of alarms was altered by having the dial flash when an alarm occurred. This prediction was accounted for in the models by discounting the value of the dial to zero, so that the predicted proportion of attention to the flashing dial was 0% percent. (Setting the BW or the AF of the dial to zero would have given the same prediction.)

Participants devoted different proportions of attention to each of the four quadrants, as evidenced by a reliable main effect of quadrant, $F(3, 12) = 14.56$ $p < 0.001$

(Figure 10.5). Analyses did not reveal a significant session by quadrant interaction, $F(12, 48) = 0.87$, $p = 0.58$, providing no evidence that the proportion of attention devoted to each quadrant varied as a function of the experimental session. It is clear from a visual examination of the data (Figure 10.5) that the largest discrepancy between the sampling pattern predicted by the optimal criterion and that exhibited by participants is an undersampling of the LR dial and an oversampling of the UL dial.

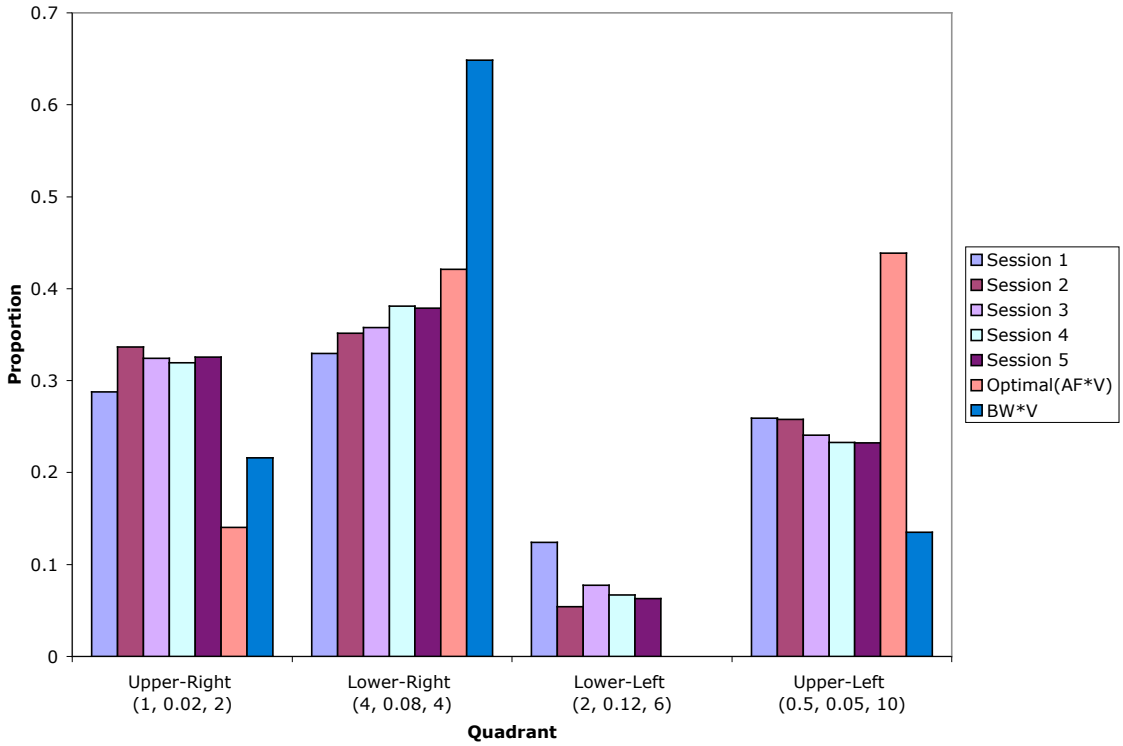


Figure 10.5 Average proportion of gaze duration by quadrant and session collapsed across all participants.

In order to examine the differences in the sampling strategies of individual participants, we present a qualitative discussion of the sampling patterns of each of the five experiment participants.

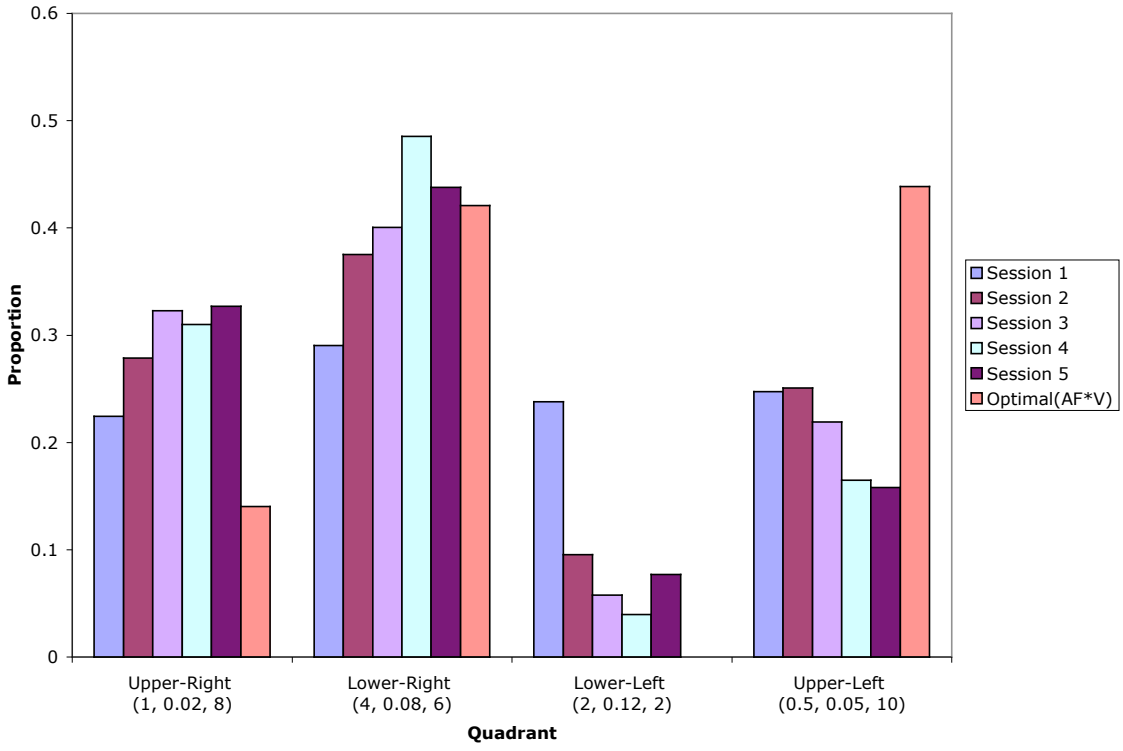


Figure 10.6. Participant 1, AFJ, Proportion of gaze duration on four quadrants by session.

The sampling pattern of AFJ appeared to change after the first session, after which he sampled the lower-left dial at a much lower rate. He also exhibits a tendency to sample the upper-left dial less frequently over the course of the sessions, which is contrary to the prediction made by the AF*V model. He shows a tendency to severely over sample the upper-right dial and undersample the UL dial relative to the AF*V prediction.

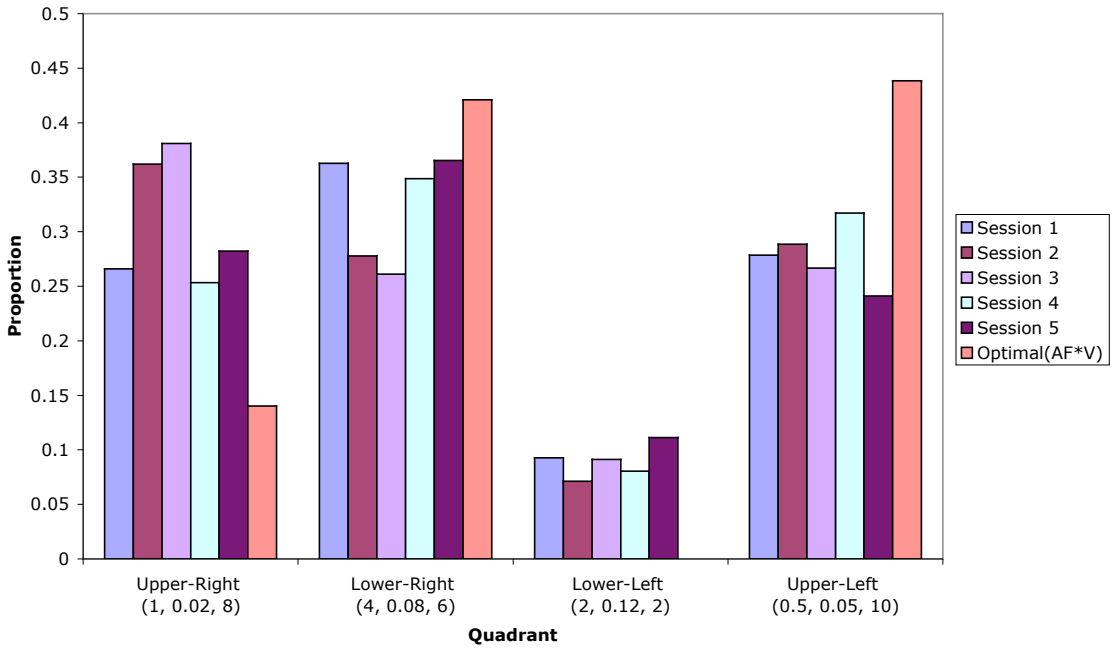


Figure 10.7. Participant 2, AW, Proportion of gaze duration on four quadrants by session.

Similar to participant 1, AFJ, participant AW displays sampling tendencies which are not in accord with the AF*V model. She also shows a tendency to oversample the upper-right dial and undersample the upper-left dial.

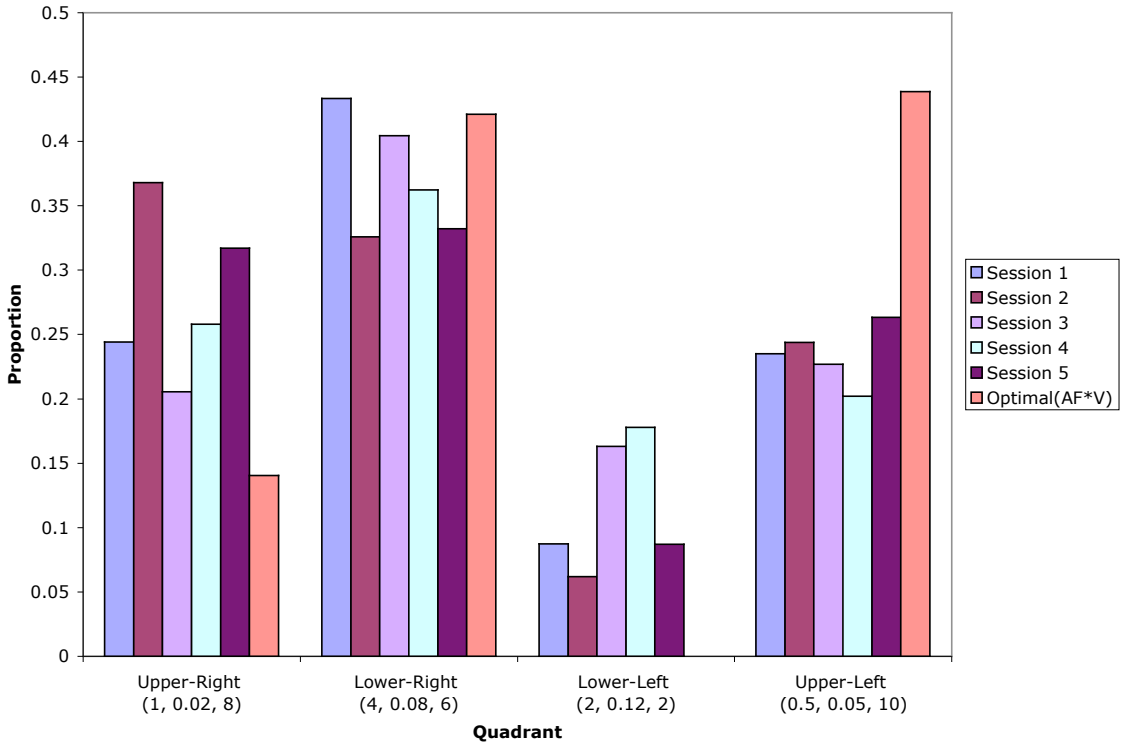


Figure 10.8. Participant 3, BRS, Proportion of gaze duration on four quadrants by session.

Participant 3, BRS, also displays a sampling pattern similar to the previous participants, with a tendency to oversample the upper-right dial and undersample the upper-left dial.

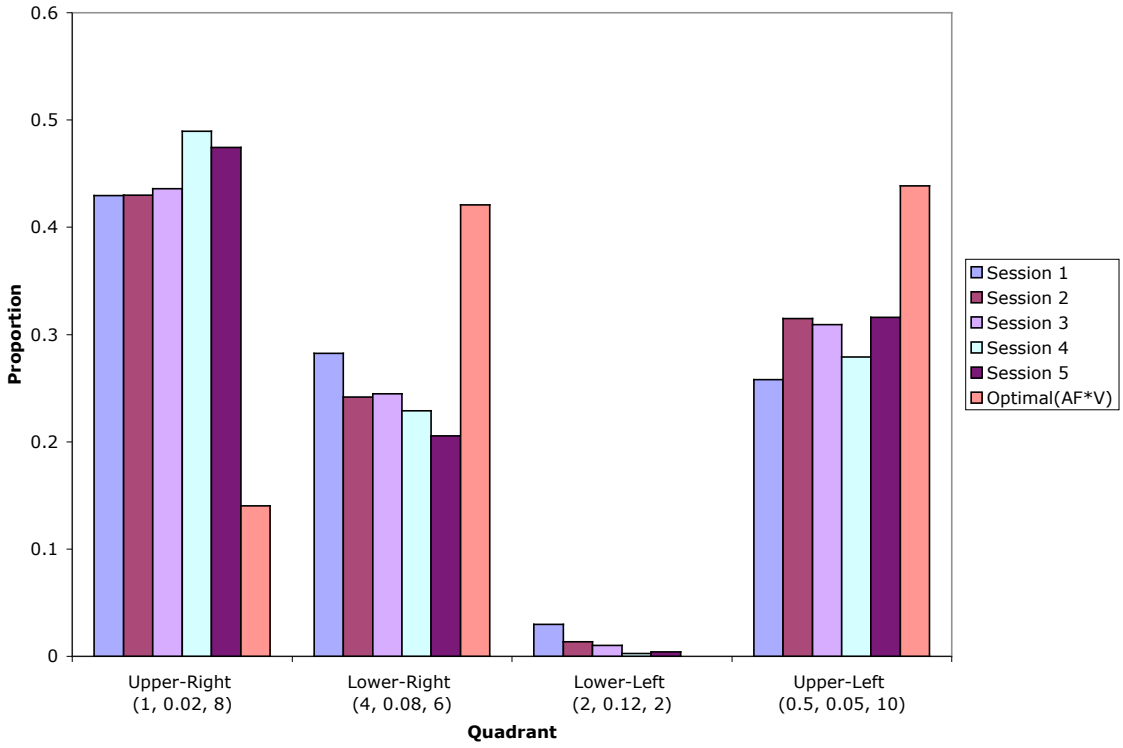


Figure 10.9. Participant 4, CC, Proportion of gaze duration on four quadrants by session.

Like the other participants, CC exhibits a tendency to oversample the upper-right dial and undersample the upper-left dial. However, his pattern is more extreme than the other participants in that he greatly oversamples the upper-right dial. He also is quite efficient in his sampling of the lower-left dial, which he very rarely overtly samples.

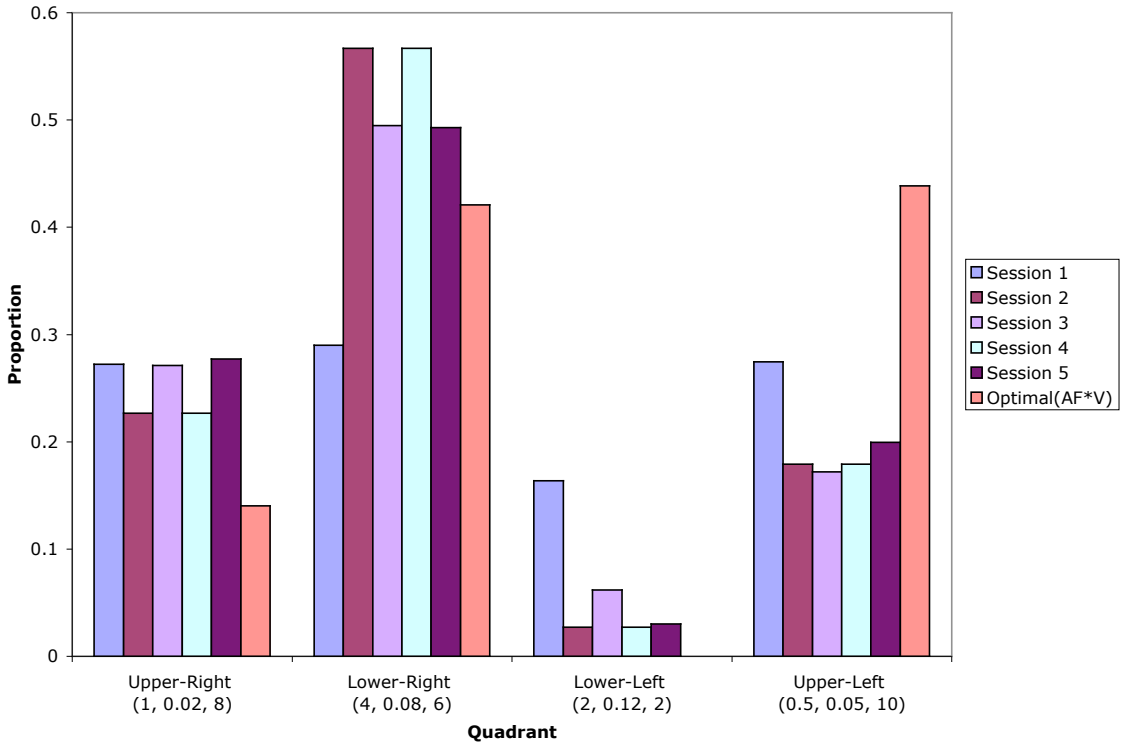


Figure 10.10. Participant 5, EY, Proportion of gaze duration on four quadrants by session.

After session 1, Participant EY does not sample the LL dial, except very infrequently. Her discrepancies from the AF*V model are primarily exhibited in a tendency to undersample the upper-left dial.

10.4 Analysis of the Proportion of Variance Explained by the Models

Examining the fit of each model to the data by participant (Figures 10.12) reveals the capability of the models to adequately describe and predict the attention allocation patterns of some participants and an inability to predict the patterns of others. Like Experiment 1, in which some of the participants exhibited sampling patterns that did not correspond well with any of the theoretical models, participant CC exhibited patterns of

sampling behavior which did not correspond well to any of the models. The models which performed the best in terms of describing the participant data were the IFT((AF+BW)/2) model and the IFT(BW) model. This suggests that BW was a key determinant of participants sampling behavior, as it was in Experiment 2.

Model	AF*V	BW*V	BW,AF*V	V*…/2	IFT (AF)	IFT (BW)	IFT(AF,BW/2)
R^2	0.37	0.63	0.47	0.65	0.31	0.72	0.74
- CC	0.43	0.72	0.55	0.74	0.35	0.80	0.83

Table 10.3 Average proportion of variance explained by the models (R^2). The second line is the average proportion of variance explained by the models when participant CC is excluded from analysis.

Analyses of the proportion of variance explained by the models across the different sessions (Figure 10.13) did not reveal a reliable effect of session, $F(4 ,16) = 0.53$, $p = 0.71$, nor was there a reliable session by model interaction, $F(24 ,96) = 0.63$, $p = 0.90$, indicating that there is no evidence that the pattern of fit for the seven models to the data changed over the course of the five sessions. There was, however, a reliable effect of model, $F(6, 24) = 6.27$, $p < 0.001$, indicating that there was a difference in the level of R^2 ascribed to the different models.

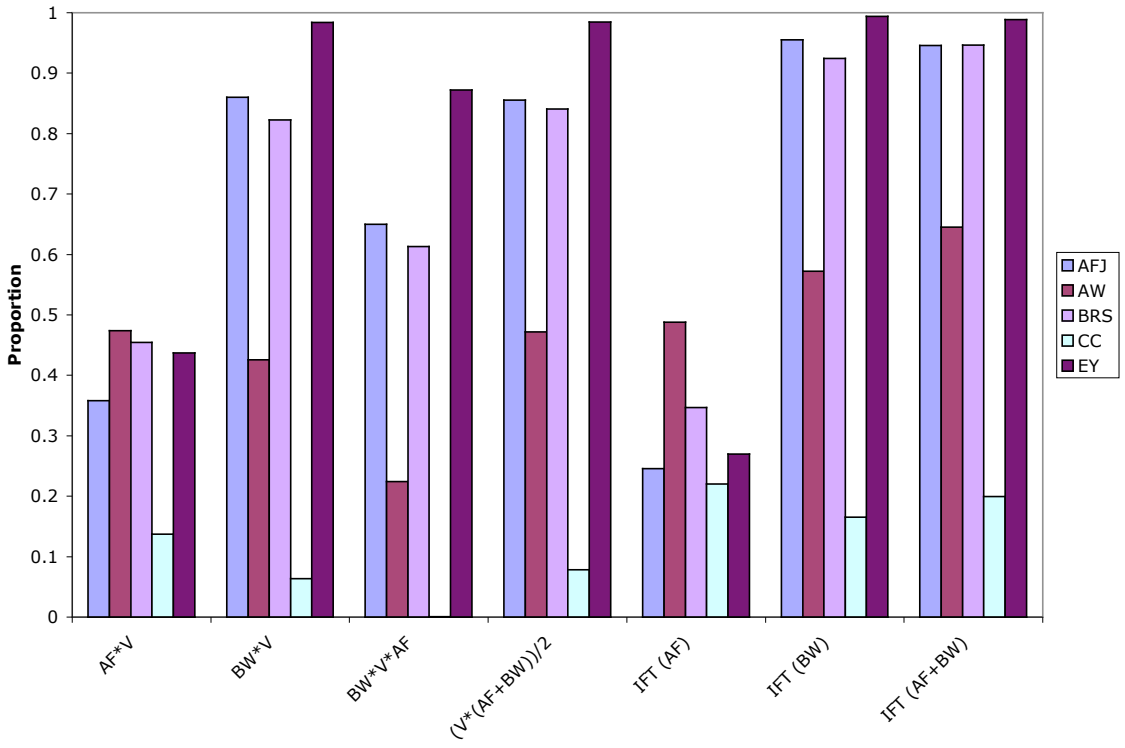


Figure 10.11 Proportion of variance explained by participant and model.

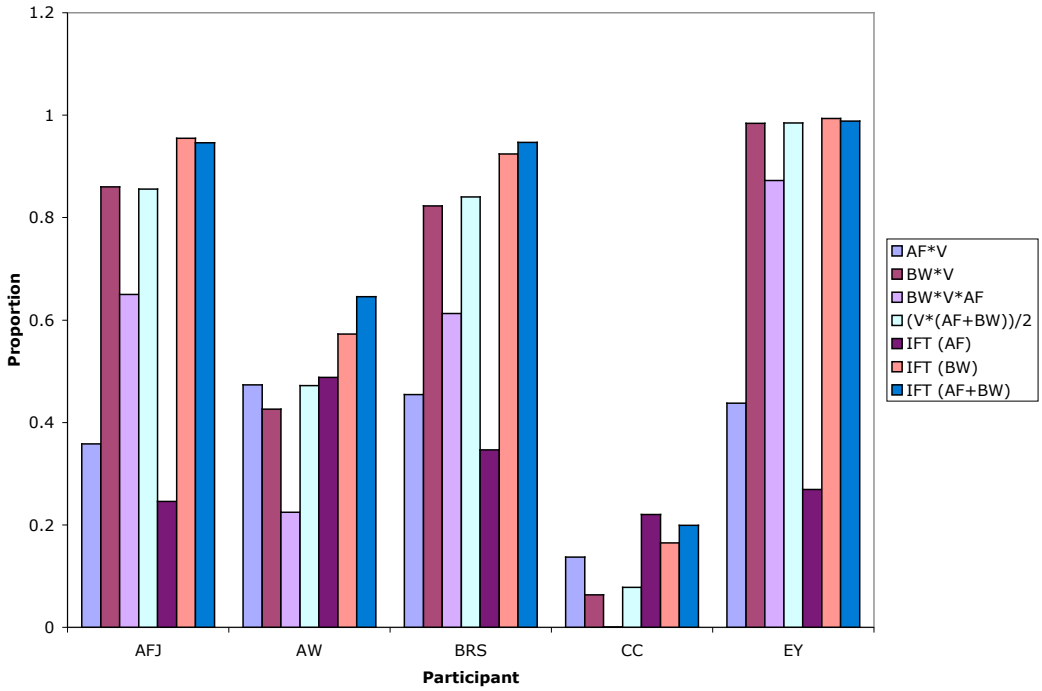


Figure 10.12 Proportion of variance explained by participant and model.

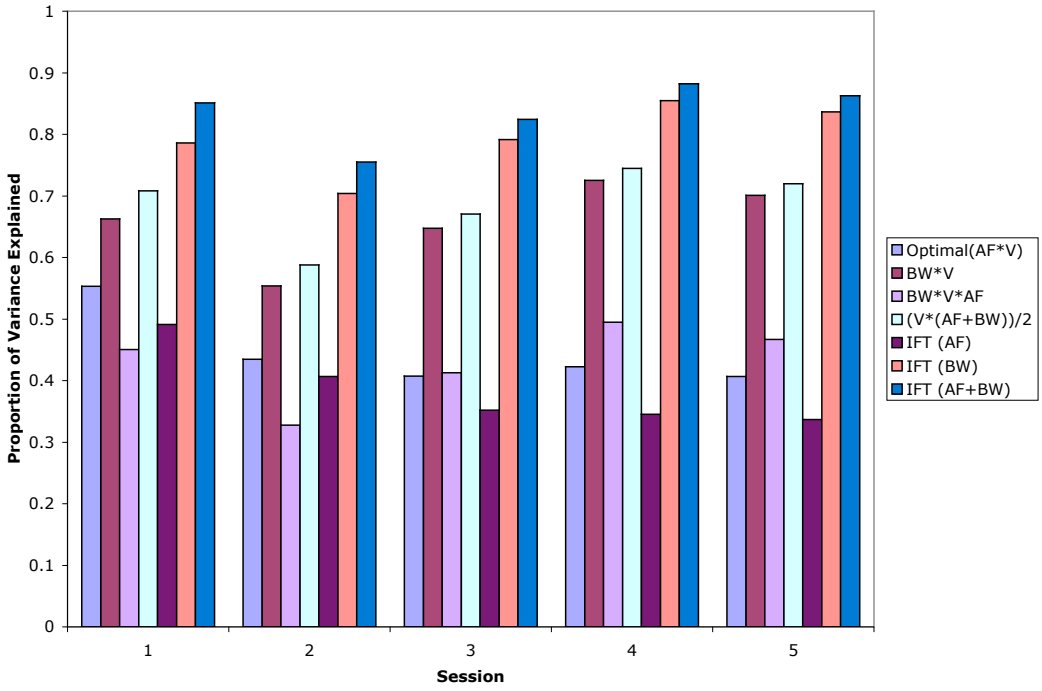


Figure 10.13 Average proportion of variance explained by models, collapsed across participants.

10.5. Experiment 3 Discussion

The primary purpose of Experiment 3 was to examine the influence of highly visually salient information in the environment on the visual sampling behavior of participants. The secondary purpose of the experiment was to examine whether any of the proposed visual sampling models could be adapted in a simple straightforward manner to account for the influence of a highly salient information source. In this experiment, the lower-right dial flashed whenever an alarm occurred on it, and remained flashing until the needle of the dial had moved out of the alarm region.

It is evident from the examination of the scoring data from Experiment 3 that the addition of the flashing dial made the experiment easier for participants. As a group, the participants in Experiment 3 had a higher average score as a proportion of the maximum score than the participants in Experiment 1, $t(54) = 2.13$, $p < 0.05$. Presumably, the addition of the high VS dial freed up attention to be devoted to other dials, and as participants spent more time looking at the other dials, their detection of alarms on the other dials improved. Hence detection rates on the other dials improved, without sacrificing detection rates on the flashing dial.

In the discussion of previous experiments, it was noted that the AF*BW function described well the ordinal ranking of dwell durations on the dials. Specifically, those dials with the longest average dwell duration, relative to the other dials, had the largest product of AF times BW. It was hypothesized that this was due to the greater number of near alarms on those dials with a higher BW*AF product. With one caveat, that relationship was once again found in this experiment. The caveat is due to the flashing dial, which had the lowest average dwell duration among participants, yet had the second

highest BW*AF product. It was expected that this dial would have both a low sampling frequency and a low dwell duration, as participants quickly learned that they did not need to sample the dial. If this dial is removed from the analysis, the BW*AF function once again predicts the ordinal ranking of dwell durations perfectly.

One of the primary purposes of Experiment 3 was to examine if people continued to sample the flashing dial. The experiment was designed with the goal that this information source essentially had perfect visual salience, i.e., that participants could detect alarms on the dial without ever overtly sampling it. This prediction was accounted for in the models by discounting the value of the dial to zero, so that the predicted proportion of attention to the flashing dial was 0% percent. (Setting the BW or the AF of the dial to zero would have given the same prediction.) Without altering the expected value of the dial in the AF*V model, the model would have predicted that participants sampled the dial 17.4% of the time. Over the final four sessions of the task, participants actually sampled the dial about 5% of the time, which was found to be reliably different from both 0%, $t(24) = 6.43$, $p < 0.001$, and 17.4%, $t(24) = 8.27$, $p < 0.001$. It was found the all of the models were a better fit to the participant data, in terms of the proportion of variance explained, when a 0% prediction was used. Granted, simply predicting that participants will never look at the dial is a very strong assumption and may be considered a crude method of fitting the data to the model. Undoubtedly the models would have been a better fit to the data if this parameter were adjusted somewhat. However, the power of the models to predict a very high proportion of variance explained in the data (approaching 90%) using such a simple method is quite promising.

Similar to previous experiments, one of the participants in Experiment 3 scored very well on the task but had a sampling pattern that was not predicted well by any of the models. Participant CC had the highest average score among the five participants, and yet had the lowest R^2 on all of the models relative to the other participants. It should be noted that CC, similar to the participants in the previous experiments who exhibited the same relationship between high scores and low degree of model fit, also described noticing correlations between specific patterns on the dials and when alarms occurred. He conveyed his sampling strategy as stemming from the upper-left dial. According to him, he could determine when alarms were likely to occur on the UR and LR dials based on the movement of the UL dial. And because he was able to score so well using such a strategy, it is possible that he did indeed discover some correlations in the dial movements. No other participants indicated that they noticed any such correlations.

In the analysis of the proportion of variance in the data explained by the models, it was clear that the models which fit the data best were models that incorporated BW. Models which only incorporated AF did a poor job of explaining the variance in the participant data. One possible reason why the AF models did not do well in this task may be due to the under constrained nature of the task. As in Experiment 1, participants may have felt that they had ample time to sample the dials at a sufficient rate. Some evidence for this idea may be found in the scoring patterns of participants; on two trials in Experiment 1 participants were able to score a perfect score, i.e., detect 100% of the alarms. In Experiment 3, participants achieved a perfect score no less than eight times. Of course, observing all of the alarms is necessary to score perfectly, but not sufficient; participants must also press the correct button and do so within one second of identifying

the alarm. There were several times where participants knew they had missed a perfect score by one or two alarms, and even know exactly which two alarms they had missed. This may provide some anecdotal evidence that participants were able to sample all of the dials at a sufficient rate. If they felt that this was the case, then sampling according to in an optimal and efficient manner would not be an important concern. If that is the case, then we would not expect AF to be a constraint on the sampling behavior of participants. Similarly, we would not expect sampling effort to be a resource that they felt the need to allocate efficiently. In an environment where sampling effort is not a factor, then models that attempt to account for effort, as the IFT models do, will not be more predictive (and may be less so) than models which do not explicitly incorporate it. If the task had been more difficult, say by increasing the overall alarm frequencies of the dials, then perhaps effort and AF would have played a greater role and models that account for it would have been superior to alternative models. Indeed, the IFT models exhibited superior performance in both experiments 1 and 2.

If the task was not under constrained such that participants felt the need to allocate their visual attention efficiently, then we might expect BW to be a primary factor in their attention allocation strategy. BW and V are the two most easily accessible cues of a dials expected value available to participants. If the participants exhibited tendencies to be "cognitive misers" then we might expect them to use the cues most easily available to them. It is worth noting that all of the participants did well at the task, despite the fact that none of them exhibited sampling patterns that corresponded well to the AF models.

11. General Discussion

11.1 Research Goals

The primary goals of this project were to evaluate and refine several models that may describe and predict visual sampling behavior in a variety of contexts. Those contexts include altering the ecological validity of the set of information sources, which, in this case was the correlation between bandwidth and alarm frequency over the set of four dials. Also manipulated was the visual salience of an information source. In the third experiment, one of the dials was given "perfect" visual salience, perfect in the sense that it was designed so that observers could detect relevant events on the information source without overtly sampling the dial. The final context in which the set of sampling models were evaluated involved manipulating the effort required to sample an information source.

Previous research had not examined mathematical models of visual sampling behavior in such contexts. The original set of experiments using the same paradigm as that employed in this series of studies did not vary the ecological validity of the set of dials. In fact, the bandwidth and the alarm frequency on each dial were perfectly correlated. More recent work conducted by Miller and Kirlik (2004) manipulated the ecological validity of the set of four dials. They examined the sampling patterns of participants at ecological validity values of 1.0, 0.75, and 0.25. The value of 0.5 used in this series of experiments adds an additional ecological validity manipulation that may provide a clearer picture of how the relationship between bandwidth and alarm frequency influences the sampling patterns of individuals.

Similarly, previous research had not examined the influence of visual salience and effort on visual sampling behavior in a Senders-like, four-dial paradigm. Although the SEEV model posits that the visual salience of relevant visual events and the effort required to sample different information sources are factors that may influence the ways in which people sample information in supervisory control tasks, no studies had as-of-yet been conducted in which these properties of the visual environment were manipulated.

With respect to refining the models of visual sampling under investigation in the set of experiments reported here, to the extent that the experimental manipulations influenced the behavior of participants, we needed to account for the factors that were manipulated in the models. In their formulations prior to this set of studies, none of the models took each of the manipulated variables into account. Of the factors manipulated, Senders' models only took BW into account. The Information Foraging Theory models had never been adapted to be used with a visual monitoring task, and whether they even applied to such a context was an open question. The SEEV model theoretically accounts for the influence of salience and value on sampling behavior, but it did not speak to the disassociation of BW and AF. Indeed, none of the models spoke to the relative influence of bandwidth and alarm frequency on sampling behavior when the two factors were disassociated.

Evaluating the different models involved comparing the ability of the models to describe and predict the behavior of participants across the different contexts just mentioned. We were specifically interested in which models could account for the greatest degree of variance in the participant data. The goal of that aspect of the study is to determine which models best apply to different contexts.

11.2 Experimental manipulations and their influence on sampling behavior.

Experiment 2 was designed to investigate the influence of the effort required to sample information on sampling strategies of participants. It is clear from the results of the experiment that effort was a factor in the visual sampling of behavior. Relative to Experiment 1, which was identical to Experiment 2 except for the effort manipulation, the participants of Experiment 2 exhibited different patterns of visual attention allocation, dwell duration, and scoring. Other than making the task more difficult, as evidenced by the lower average scores of participants, the change in the task environment likely resulted in the development of more conscious sampling strategies. Participants were able to describe their sampling strategies explicitly, and also exhibited strategies that differentiated from typical visual sampling strategies, such as participant AW who chose to only sample two of the dials on many of the experimental trials.

One additional goal of Experiment 2 was to evaluate the Information Foraging Theory class of models, which explicitly incorporate sampling effort as a model parameter. Depending on what individual participants were most sensitive to, BW or AF, the IFT models did slightly better than the simple multiplicative models (AF*V or BW*V) at accounting for the proportion of variance explained in the data. This of course may be expected due to the formulation of the IFT models. The parameter for the time needed to extract information from a source (t_w) was inferred from the dwell durations of the participants. For each dial, its t_w parameter was the average dwell duration of all of the participants in the experiment on that dial. Hence, inherent in the IFT models were additional information on how long participants looked at the dials. And of course, we

would expect models with more information about the task to do a better job of predicting performance in the task, which they did. However, the sampling strategies of individual participants still appeared to be dominated more by AF or BW and less by sampling effort (as approximated by the time within dial and time between dial parameters), as evidenced by the only slight increase in the proportion of variance explained by the IFT models relative to models that did not incorporate effort.

The primary goal of Experiment 3 was to investigate the influence of visual salience on sampling behavior. As opposed to Experiment 2, which was a more difficult task for participants, the task of Experiment 3 was easier for participants, as evidenced by their higher average scores as a proportion of the maximum possible score. On the whole, participants were best described by models which incorporated BW instead of AF. This may have been due to the fact that the task was easier for participants. Participants were not constrained to be efficient in their attention allocation patterns. One possible explanation for their reliance on BW in the face of an easier task is that they gravitated to the most easily accessible elements of the display, AF and V.

In each of the three experiments, the dwell durations of participants on the four different quadrants corresponded to the ordinal pattern described by the product of AF and BW. Specifically, the dials with the highest AF by BW product had the longest dwell duration and vice versa. It was hypothesized that this was because the AF*BW product corresponded to the incidence of near alarms on each of the dials. Future research will involve examining this hypothesis in the data collected by Miller and Kirlik (2004).

The scoring patterns of participants indicated that they were able to learn the task quickly. In Experiments 1 and 2, after the first 8 to 10 trials, all of the participants were

scoring at approximately the same level as they did on the last 8 to 10 trials of the experiments. However, although large differences were found between the scores of participants on the first session and on the last session of the experiments, large differences between the first and last session were not found in their sampling data. The implications of such small changes in sampling strategy despite large-scale gains in scoring are discussed in the following section.

11.3 Discussion of the Model Results

One of the primary goals of this series of experiments was to determine which models best describe human performance. In order to examine this question, we examined the proportion of variance explained by several formulations of the models in the participant data from each of the three experiments. No single model outperformed all of the other models across all three experiments in terms of explaining the variance in the participant data. In fact, there was not really a "winning" model in any of the three experiments. However, for nearly every participant in the experiments, there was at least one model that described their individual data very well.

This finding, that different models described the behavior patterns of different individuals raises the question as to which model describes which type of person. Interestingly, the participants seemed to fall roughly within three categories, which can be described by the interaction of their scoring patterns on the task and the type of model that best describes their visual sampling pattern. The sampling behavior of participants who scored poorly on the task were generally described best by models in which the bandwidth of the dial was the dominant factor determining the predicted patterns of attention allocation (Table 11.1) The group of participants who scored in the middle,

relative to the other participants, had sampling patterns that were best described with models in which the alarm frequency of the dials was the dominant factor. And, perhaps most interestingly, those participants who had the highest scores at the task, were not well-described by any of the models.

Participants	**Score Index**	**Proportion of Variance Explained by Model**			
		AF*V	**IFT(AF)**	**BW*V**	**IFT(BW)**
BAF, EAG, ENG, CC, JRS, JW	1.36	0.42	0.44	0.43	0.39
AW, AFJ, SJM , DG	0.97	**0.56**	**0.85**	0.56	0.71
EY, BRS, RS, LMA, NLO	0.62	0.36	0.49	**0.65**	**0.85**

Table 11.1. Participants in all three experiments divided into three groups based on their scoring index, which was calculated as the average score of a participant divided by the average score of all the participants in the same experiment. The groups were divided by the top six scorers, the next four highest scorers, and the lowest five scorers.

It makes intuitive sense that participants who scored well at the task were most sensitive to the alarm frequency of the dial. Since the task was one of detecting alarms, being sensitive to the relative rates of alarms on the different dials would likely result in excellent performance at the task. Indeed, the "optimal criterion" that we have used as a benchmark to qualitatively describe the performance of individuals is based on the alarm frequency of the dials multiplied by the value of the dials, which essentially gives us the expected value of a dial.

It also makes intuitive sense as to why participants who were most sensitive to the BW of the dials did not score particularly well at the task. Because bandwidth and alarm frequency are only correlated at the rate of 0.50, then BW acts as a sort of red herring in this context. Bandwidth is a proximal cue, which is readily available to observers and can be monitored at any time, but which does not accurately predict the frequency of alarms,

which is a distal cue that can only be observed via infrequent, discrete events. It is worth noting that BW and AF are often highly correlated in the environment, i.e., information sources that are highly volatile also have a tendency to reach extreme values more frequently than information sources which are not so volatile. However, it is not the case that these elements are necessarily always highly correlated.

It is less clear why participants who performed the best at the task, i.e., scored at the highest rate, were not well described by any of the models. I believe it is because these participants were actually able to learn and remember correlations or even specific patterns between the dials. The types of patterns that participants described learning were of the type where a specific series of movements on one dial would be a cue to them that an alarm was about to occur, or had just occurred, on another dial. On each trial the movements of each of the dials were replayed to participants based on prerecorded movement patterns. The patterns were recorded such that each dial had a specified bandwidth and alarm frequency and such that the group of four dials had an ecological validity of 0.50. For each experiment, four different dial movement patterns were recorded and they were replayed to participants twice during each experimental session in a constrained random order (constrained such that each pattern was played once before any pattern was played twice). It would appear that the participants who scored extremely well at the task were able to remember the prerecorded patterns. This explanation is based on discussions with four of the participants in which they specifically mentioned patterns that they recognized in the dials. Only four participants mentioned observing such patterns in the dial movements, but six participants, the top six scorers at the task when indexed relative to the other participants in their experiment, did not conform well

to any of the models despite their high scoring capacity. It is worth noting that when asked at the end of the experiment, participants had all noticed that they saw the same movement patterns multiple times over the course of the experiment, but no participant had noticed that each of the prerecorded patterns were replayed twice in each experimental session.

It would be beneficial to the modeling endeavor of this study if we could identify the correlations or specific patterns that participants found in the dials and explore models that may take those types of patterns into account. This is difficult to do for two reasons. First, we do not really know what type of patterns to look for. Since the movements of the dials were prerecorded and replayed multiple times, then by definition, there are patterns in them. It is possible that if participants watched a single recorded movement series enough times, then they would be able to learn when and where the alarms occurred on that trial. However, simply by looking at the movements of the dials, we cannot determine which patterns they were able to recognize and learn, and which they were not. In order to do that we would need to know, at the very least, the values of the dials in conjunction with when participants were looking at them. Since we do not have that data, examining which patterns people were sensitive to becomes a difficult, or perhaps even impossible, task.

Recording the value of the dial as participants looked at it also has implications in terms of the type and complexity of the model used to describe and predict the eye movement data. The models employed in this series of experiments were only able to predict the attention allocation patterns of participants at a rather gross level, the proportion of time participants observed each dial over the course of an experimental

session. However, it was hoped that we would be able to make some finer-grained predictions regarding the eye movement patterns of participants. Specifically, we would have liked to accurately describe the transition probabilities between dials, i.e., when a participant was looking at any one dial, what is the likelihood that a particular dial would be the next dial they looked at. Unfortunately, attempts at this type of analysis yielded unsatisfactory results and discussion of them was not included in this document. Similarly, Miller and Kirlik (2004) attempted to examine the likelihood that the sampling patterns of participants in their experiments could be better predicted based on the proximity of the needle on a dial to an alarm region. However, they did not find a predictive relationship between the proximity of an alarm and the attention allocation patterns of participants. Contrastingly, the results of the experiments described herein do suggest a relationship between near alarms and sampling behavior. We did not record near alarms in the data, but we make the argument that it is approximated by the product of bandwidth and alarm frequency. This assumes that dials with relatively higher bandwidths and relatively higher alarm frequencies will have higher incidences of near alarms. We did not find the number of near alarms influenced the proportion of time participants spent observing a particular dial over the course of a trial or session, but it was a strong predictor of the dwell duration of participants on individual dials.

It is noteworthy that only the participants who performed at the highest level at the task mentioned that they were cuing their sampling behavior off of patterns that they had found in the dials. In the sense that achieving high scores was the goal of the task, these participants may be considered the "expert" participants of the entire group of participants. Hence, this finding is particularly noteworthy as the domains in which we

would like to apply models of sampling behavior are domains that are dominated by experts, e.g., airplane pilots and power plant operators. Further, the contexts in which these experts operate, such as the cockpit of an airplane, often contain sources of information that are not independent of one another. One possible definition of expert performance in these types of environments may include having knowledge not only of the statistical properties of the information sources, but having knowledge of the relationships between information sources. If that is the case, then in order to describe and predict the attention allocation patterns of experts, we would need to develop models of sampling behavior that incorporate the correlations between information sources. Senders (1983) began to develop such models, although, to my knowledge, they have never been formally evaluated.

The results of this series of experiments suggest that no single model accounts well for the sampling patterns of all participants, but rather that specific models account for different types of participants. Further, there is some evidence that which type of model is going to best describe a participant may be determined early on in the testing process. Contrary to expectations, the sampling patterns of participants did not appear to change much over the course of the experiment. Regarding the proportion of time that participants allocated to different quadrants, the analyses of the first and second stages of Experiment 2 and of Experiment 3 did not reveal a reliable quadrant by session interaction, indicating that there is no evidence that the relative amount of time participants allocated to the four quadrants changed as a function of the experimental session. Only in Experiment 1 was there a reliable quadrant by session interaction. However, visual inspection of the data does not indicate that the participants were

converging on any particular strategy in Experiment 1. Rather, their data manifests a quartic function, in which when a high proportion of time was allocated to a quadrant in one session, a low proportion was allocated to the same quadrant in the following session, and vice versa. The finding that the sampling patterns of participants did not significantly change over the course of the experiments was surprising, and may well be due to a unique aspect of the experiments, the constant setting of the ecological validity at the 0.50 level.

Setting the ecological validity at 0.50 was likely a key determinant of all of the results in the three experiments. As noted previously, some of the participants in each of the experiments did not manifest patterns of attention allocation that were well-described by functions that incorporated AF as a principal component of the dials. In many of these cases, these participants exhibited patterns of attention that were well-described by functions that incorporated BW instead of AF. Hence, these participants, which generally scored poorly on the task, were unable to disassociate the proximal cue of BW from the distal criterion of AF. Of course, in the case where BW and AF are perfectly correlated, when EV is equal to 1.0, it would not matter whether participants were keyed in on AF or BW. Miller and Kirlik (2004) manipulated the EV of the four dial set across a series of experiments and showed that participants differentially weight AF and BW in their sampling strategies as a function of the EV of the four dial set. Generally speaking, their results indicate that when the EV is set at a very low level (0.25 in their experiments), it is easier for participants to disassociate BW from AF. This makes intuitive sense, since at such a low EV, it may be much more obvious that AF and BW are unrelated. Contrastingly, at an EV of 0.75, it was more difficult for participants to disassociate AF

from BW. Presumably, the higher the correlation between AF and BW, the more difficult it is for participants to disassociate them. A correlation of 0.50 may not be low enough for observers to easily recognize that BW is not correlated with AF. Of those participants whose sampling strategy was well-described by one of the models, about half of them were described by a model incorporating AF and about half by a model incorporating BW.

Although all of the participants noted that the bandwidth and the alarm frequencies of the dials were not perfectly correlated, I did not ask them about the degree to which they believed them to be associated. It possible that some participants, those who scored poorly at the task, did not believe the two to be greatly divergent. And this result may be function of the EV used. Future work will involve examining the attention allocation patterns of the participants in the Miller and Kirlik (2004) studies. It is hypothesized that at lower levels of EV, we will find fewer participants who use BW as a primary determinant of their sampling strategy.

It should be noted that the EV level of 0.5 was not chosen arbitrarily. It is possible, and even probable, that there are information sources in the environment whose rate of change is not perfectly correlated with significant events on that information source. For instance, it is easy to imagine a dial that has a high rate of movement, yet only rarely enters the extreme ranges. The altitude gauge in an airplane cockpit is constantly fluctuating, yet only very rarely does it enter into a range that would constitute an alarm (hopefully). If observers are unable to decouple the rate of movement and the frequency of significant events of an information source, then they may exhibit sampling patterns which are less than optimal. The results of this series of experiments

incorporated with those of Miller and Kirlik (2004) suggest that the likelihood that participants will be able to decouple these aspects of information sources is a function of the correlation between them. Further, the results of this series of studies suggests that at a correlation of 0.5, not only will some participants not show sampling patterns that indicate they have detected the disassociation between elements, but that their less-than-optimal sampling patterns may persist over time. The participants in these experiments whose sampling patterns were well described by functions incorporating BW did not change their sampling patterns much over the course of the 3.5 hours of practice at the task. However, although 3.5 hours constitutes a lengthy psychology experiment, it is not a long time with respect to training on real-world systems, such as the training an airplane pilot goes through. Perhaps with additional practice at the task, the participants would have learned that BW was a less than optimal cue to drive sampling behavior. Further, the participants in the task were not given any training at the task nor were they ever told that AF and BW were not perfectly correlated and that AF was the cue on which they should base their sampling strategies. Perhaps additional training and some explicit knowledge regarding the properties of the display would be all that it takes to get people to begin to abandon less than optimally informative environmental cues.

On the whole, the results of this set of experiments are in accord with previous work in the domain of visual sampling in supervisory control tasks. Except for the "expert" participants, whose data were not well described by any models due to the experiment environment, the sampling patterns of all participants were well-described by one of the models, either those incorporating BW or those incorporating AF. Of course, in previous work, other than that of Miller and Kirlik (2004), BW and AF have been

perfectly correlated, and the ability of a model using only BW as variable to describe sampling behavior would produce good approximations to human data. This series of experiments extends previous work in the domain in two ways. First, we demonstrated the capability of Information Foraging Theory models to be useful tools to describe visual sampling behavior. The IFT class of models showed improvement over the other sampling models when the behavior of participants was constrained by the experimental environment, specifically when the effort to sample an information source became more costly, and hence an important factor in to allocating attention efficiently. The work also extended the realm of visual sampling models by demonstrating the capability of simple models to describe sampling behavior in a wider variety of contexts. The models may be used to describe people in less than optimal information environments, when there is variability in the visual salience of information sources and events and in the effort required to sample different sources.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409-429.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96, 703-719.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396-408.

Bacon, W. F., & Egeth, H. E. (1997). Goal directed guidance of attention: Evidence from conjunctive visual search. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 948-961.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.). (1992). *The adapted mind*. Oxford: Oxford University Press.

Baron, S., Kleinman, K.L., & Levison, W.H. (1970). An optimal control model of the human response, Part II: Prediction of human performance in a complex task. *Automatica*, 6, 371-383.

Baron, S., & Levinson, W.H. (1973). A display evaluation methodology applied to vertical situation displays. *Proceedings of the Ninth NASA Annual Conference on Manual Control*. (NTIS No.N75-19126/2. Cambridge, Mass.: M.I.T.

Baron, S., & Levinson, W.H. (1975). An optimal control methodology for analyzing the effects of display parameters on performance and work load in manual flight control. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-5, 423-431.

Baron, S., & Levinson, W.H. (1977). Display analysis with the optimal control model of the human operator. *Human Factors,* 19, 437-451.

Brogan, D., Gale, A., & Carr, K. (1993). *Visual search 2*. Bristol, PA: Taylor & Francis.

Byrne, M. D. (2001). ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies, 55,* 41-84.

Carbonell, J. R. (1966). A queueing model for many-instrument visual sampling. *IEEE Transactions on Human Factors in Electronics*, HFE-7, 157-164.

Carbonnell, J. F., Ward, J. L., & Senders, J. W. (1968). A queuing model of visual sampling: Experimental validation. *IEEE Transactions on Man Machine Systems*, MMS-9, 82-87.

Cashdan, E. (1992). Spatial organization and habitat use. In E. A. Smith & B. Winterhalder (Eds.), *Evolutionary ecology and human behavior* (pp. 237-266). New York: de Gruyer.

Chou, C., Madhavan, D., & Funk, K. (1996). Studies of cockpit task management errors. *The International Journal of Aviation Psychology*, 6(4), 307-320.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind* (pp. 163-228). Oxford: Oxford University Press.

Curry, R.E., Kleinman, D.L., & Hoffman, W.L. (1977). A design procedure for control display systems. *Human Factors*, 19, 421-436.

Deco, G. & Lee, T.S. (2002). A Unified Model of Spatial and Object Attention Based on Inter-Cortical Biased Competition. *Neurocomputing,* 19, 421-436.

Dismukes, K. (2001). The challenge of managing interruptions, distractions, and deferred tasks. *Proceedings of the 11th International Symposium on Aviation Psychology*. Columbus, OH: The Ohio State University.

Edwards W. (1961), Probability Learning in 1000 trials, *Journal of Experimental Psychology*, 4, 385-394.

Ellis, S. R., & Stark, L. (1986). Statistical dependency in visual scanning. *Human Factors*, 28, 4, 421-438.

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

Fitts, P.M., Jones, R.E., & Milton, J.L. (1950a). Eye fixations of aircraft pilots III. Frequency, duration and sequence of fixations when flying Air Force Gropund Controlled Approach system (GCA). United States Air Force Tech. Rep. #5967. Dayton, Ohio: Wright Patterson Air Force Base.

Fitts, P.M., Jones, R.E., & Milton, J.L. (1950b). Eye movements of aircraft pilots during instrument landing approaches. *Aeroautical Engineering Review*, 9, 1-5.

Fleetwood, M. D. & Byrne, M. D. (2002) Modeling icon search in ACT-R/PM. *Cognitive Systems Research, 3,* 25-33.

Fogel, L.A. (1956). A note on the sampling theorem. *Transactions of IRE Professional Group on Information Theory, IT-12,* 47-48.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 1030-1044.

Gillan, D. J., & Lewis, R. (1994). A componential model of human interaction with graphs: I. Linear regression modeling. *Human Factors, 36*, 419-440.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds Matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experiment Psychology: Applied, 6(4),* 322-33

Hames, R. (1992). Time allocation. In E. A. Smith & B. Winterhalder (Eds.), *Evolutionary ecology and human behavior* (pp. 203-235). New York: de Gruyter.

Helleberg, J., & Wickens, C. D. (2001). Effects of data link modality and display redundancy on pilot performance: An attentional perspective. *International Journal of Aviation Psychology*.

Holling, C. S. (1959). Some characteristics of simple types of predation and parasitism. *Canadian Entomology*, *91*, 385-3998.

Hornof, A. J. & Kieras, D. E. (1997). Cognitive modeling reveals menu search is both random and systematic. *Human Factors in Computing Systems: Proceedings of CHI 97*, pp. 107}114. New York: ACM Press.

Jones, R.E., Milton, J.L., & Fitts, P.M. (1949). Eye fixations of aircraft pilots I. A review of prior eye movement studies and a description of eye movements during instrument flight. United States Air Force Tech. Rep. #5837. Dayton, Ohio: Wright Patterson Air Force Base.

Jones, R.E., Milton, J.L., & Fitts, P.M. (1950). Eye fixations of aircraft pilots IV. Frequency, duration and sequence of fixations during routine instrument flight. United States Air Force Tech. Rep. #5975. Dayton, Ohio: Wright Patterson Air Force Base.

Kahneman, D, Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty heuristics and biases*. *Cambridge*, Mass.: Harvard University Press.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.

Kaplan, H., & Hill, K. (1992). The evolutionary ecology of food acquisition. In E. A. Smith & B. Winterhalder (Eds.), *Evolutionary ecology and human behavior* (pp. 167-201). New York: de Gruyter.

Kleinman, K.L., Baron, S., & Levison, W.H. (1970). An optimal control model of the human response, Part I: Theory and validation. *Automatica*, 6, 357-369.

Kok, J. & Wijk, R. van. (1978). *Evaluation of models describing human operator control of slowly responding complex systems*. Delft, Netherlands: Delft University Press.

Kvalseth, T. (1977). Human information processing in visual sampling. Ergonomics, 21, 439454.

Liu, Y. & Wickens, C. D. (1992). Visual scanning with or without spatial uncertainty and divided and selective attention. *Acta Psychologica*, 79, 131-153.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *HumanComputer Interaction*, 8, 353-388.

Miller, S., Kirlik, A., Kosorukoff, A., Byrne, M. D. (2004). Ecological Validity as a Mediator of Visual Attention Allocation in Human-Machine Systems. University of Illinois Human Factors Division Technical Report AHFD-04-17/NASA-04-6.

Milton, J.L., Jones, R.E., & Fitts, P.M. (1949) Eye fixations of aircraft pilots II. Frequency, duration and sequence of fixation when flying the USAF instrument low approach system (ILAS). United States Air Force Tech. Rep. #5839. Dayton, Ohio: Wright Patterson Air Force Base.

Milton, J.L., Jones, R.E., & Fitts, P.M. (1950) Eye fixations of aircraft pilots V. Frequency, duration and sequence of fixations when flying selected maneuvers during instrument and visual flight conditions. United States Air Force Tech. Rep. #6018. Dayton, Ohio: Wright Patterson Air Force Base.

Moray, N. (1981). The role of attention in the detection of errors and the diagnosis of failures in man-machine systems. In J. Rasmussen & W.B. Rouse (Eds.), H*uman detection and diagnosis of system failures*. New York: Plenum, 1981.

Moray, N. (1986). Monitoring behavior and supervisory control. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and performance*, Vol II (pp. 40-1-40-51). New York: Wiley & Sons.

Moray, N., Richards, M., & Low, J. (1980). The behaviour of fighter controllers. Technical Report. London: Ministry of Defense.

Neisser, U., Novick, R., & Lazar, R. (1964). Searching for novel targets. *Perceptual and Motor Skills, 19*, 427-432.

Norman, K. L. (1991). *The Psychology of Menu Selection: Designing Cognitive Control of the Human/Computer Interface*. Norwood, NJ: Ablex.

Pashler, H., Johnston, J. C., & Ruthruff, E. (2001). Attention and performance. *Annual Review of Psychology*, 52, 629-651.

Payne, J.W., Bettman, J.R. and Johnson, E.J. (1993). *The adaptive decision maker*. Cambridge University Press, Cambridge.

Pinker, S., & Bloom, P. (1992). Natural language and natural selection. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind* (pp. 451-493). Oxford: Oxford University Press.

Pirolli, P. and Card, S. K. (1999). Information Foraging. *Psychological Review* 106(4): 643-675.

Pirolli, P. and Fu, W.-T. (2003). SNIF-ACT: A model of information foraging on the world wide web. *Ninth International Conference on User Modeling*, Johnstown, PA.

Poisson, M. E., & Wilkinson, F. (1992). Distractor ratio and grouping processes in visual conjunction search. *Perception, 21*, 21-38.

Raby, M., & Wickens, C.D. (1994). Strategic workload management and decision biases in aviation. *International Journal of Aviation Psychology, 4(3),* 211-240.

Rasmussen, J. (1983). Skills, rules, and knowledge: Signals, signs, and symbols and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-13, 257-266.

Senders, J.W. (1955). Man's capacity to use information from complex displays. In H. Quastler (Ed.), *Information theory in psychology*. Glencoe, Il.: Free Press.

Senders, J. W. (1964). The human operator as a monitor and controller of multidegree of freedom systems. *IEEE Transactions on Human Factors in Electronics*, HFE-5, 2-6.

Senders, J.W. (1966). A re-analysis of the pilot eye-movement data. *IEEE Transactions on Human Factors in Electronics, HFE-7*, 103-106.

Senders, J. W. (1983). Visual scanning processes. Netherlands: University of Tilburg Press.

Senders, J.W., Elkind, J.E., Grignetti, M.C., & Smallwood, R.P. (1964). An investigation of the visual sampling behavior of human observers. NASA-CR-434. Cambridge, Mass.: Bolt, Beranek, & Newman.

Shen, J., Reingold, E. M., & Pomplun, M. (2000). Distractor ratio influences patterns of eye movements during visual search. *Perception, 29*, 241-250.

Sheridan, T. (1970). On how often the supervisor should sample. *IEEE Transactions on Systems Science and Cybernetics*, SSC-6(2), 140-145.

Sheridan, T. B., & Rouse, W. B. (1971). Supervisory sampling and control: Sources of suboptimality in a prediction task. Seventh NASA Annual Conference on Manual Control. University of Southern California.

Smith, E. A. (1981). The application of optimal foraging theory to the analysis of hunter-gatherer group size. In B. Winterhalder & E. A. Smith (Eds.), *Hunter-gatherer foraging strategies* (pp. 36-65). Chicago: University of Chicago.

Smith, E. A., & Winterhalder, B. (Eds.). (1992*). Evolutionary ecology and human behavior*. New York: de Gruyter.

Stassen, H.G. (1978). Man-machine systems: Manual and supervisory control. In J. F. Michon (Ed.), *Handbook for psychonomy*. Netherlands: Deventer.

Stephens, D. W., & Charnov, E. L. (1982). Optimal foraging: Some simple stochastic models. *Behavioral Ecology and Sociolobiology*, 10, 251-263.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind* (pp. 19-136). Oxford: Oxford University Press.

Teichner, W. H. & Mocharnuk, J. B. (1979). Visual search for complex targets. *Human Factors, 21(3)*, 259-275.

Theeuwes, J. (1994). Visual attention and driving behavior. In C. Santos (Ed.), *Human factors in road traffic*. Lisbon Portugal: Escher.

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

Tversky, A. & Kahneman, D. (1971). The belief in the "law of small numbers." *Psychological Bulletin*, 76, 105-110.

Vaughan, W.S., & Mavor, A.S. (1972). Behavioural characteristics of men in the performance of some decision-making task components. *Ergonomics*, 15, 267-277.

Wickens, C. D. (2001). A model of attention allocation for aviation. *Association for Aviation Psychology Newsletter, 23(1)*, 4-9.

Wickens, C.D., Goh, J., Helleberg, J., Horrey, W.J., Talleur, D. A. (2003). Attentional models of multitask pilot performance using advanced display technology. *Human Factors*, 45, 3, 360-380.

Wickens, C.D., Helleberg, J., Goh, J., Xu, X., &. Horrey, W.J. (2001). Pilot Task Management: Testing an Attentional Expected Value Model of Visual Scanning. (Technical Report ARL-01-14/NASA-01-7). Savoy, IL: University of Illinois, Aviation Research Lab.

Wickens, C.D., Kroft, P., & Yeh, M. (2000). Data base overlay in electronic map design: Testing a computational model. Proceedings of the IEA 2000/HFES 2000 Congress (pp. 3-451-3-454). Santa Monica, CA: Human Factors & Ergonomics Society.

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 & Performance, 15*, 419-433.

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

Wolfe, J. M., & Gancarz, G. (1996). Guided Search 3.0. In *Basic and Clinical Applications of Vision Science* (pp. 189-192). Dordrecht, Netherlands: Kluwer Academic.

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

Appendix A: Instructions for Experiment 1

Experiment 221: Eye Tracking a Visual Monitoring Task
Instructions

The Task:

Your task in this experiment is to detect alarms on a group of four dials (Figure 1, next page). An alarm occurs whenever the needle of one of the dials enters the dark gray region of the dial. For instance, in Figure 1, an alarm has occurred on the lower-left dial. Whenever an alarm occurs on a dial, your task is to press a key corresponding to that dial. The keys corresponding to the upper-left and the lower-left dials are the "A" and "Z" keys, respectively. The keys corresponding to the upper-right and the lower-right dials are the apostrophe (') and slash (/) keys, respectively. Note that the location of these keys on the keyboard correspond to the location of the dials on the screen, i.e., the "A" key is the upper-left of the four keys and corresponds to the upper-left of the four dials. You'll need to keep your fingers on these four keys throughout the course of the experiment.

Scoring:

You will be awarded points based on the correct detection of alarms. Each dial has a number of points associated with it, and you will be given the number of points for a dial whenever you correctly detect an alarm on that dial. The number of points associated with each dial is shown below each dial in Figure 1, such that pairings are upper-left: 10 points, lower-left: 6 points, upper-right: 2 points, lower-right: 4 points. Also, in the alarm range of each gauge, there are bars of different lengths. These bars are used to indicate the number of points each gauge is worth. Hence, the longest bar is worth 10 points, the

bigger medium-sized bar is worth 6 points, the smaller bar is worth 4, and the smallest bar is worth 2 points. Each time you correctly detect that a gauge has gone out of bounds you will receive the number of points indicated by the length of the bar. Conversely, you will receive negative points for misses (any time that an alarm occurs on a dial and you do not detect it) and false alarms (when you push the button associated with a dial and there is no alarm on that dial). The number of negative points you will receive is the same as the number of positive points you would receive for the correct detection of an alarm on that dial.

Your overall score will be the positive points you earn for each time you correctly detect that a gauge has gone out of bounds minus the points you loose for both misses and false alarms. Since there are many ways to loose points, it is possible (or even likely) that you will have negative points – especially at the beginning.

Your total score will be given to you at the end of the each 5-minute trial. There are eight trials per one-hour experimental session.

Eye Tracking:

The task itself is a very easy one, monitoring the dials and pressing a button when an alarm occurs. However, we are not only interested in measuring your performance at the task but also in monitoring your eye movements. We are investigating how when and where you look changes as you get better at the task. The more difficult (and often more frustrating) part of the task involves working with the eye tracking equipment so that we

may record your eye movements. You will only be working at the task for 30 to 40 minutes each one-hour experiment session. The other 20 minutes or so of each hour will be spent working with the eye tracking equipment. We just thought you should be warned.

Awards:

There are five participants in this experiment. The participant with the highest average score over the course of the five experimental sessions will receive a $15 gift certificate to the web retailer of their choice (e.g. Amazon.com). The second and third highest average scorers will receive gift certificates for $10 and $5, respectively. Additionally, on each session on which your average score is above zero (0), a couple pieces of candy will be your just reward.

If you have any questions, please don't hesitate to ask the experimenter at any time. Thank you for your participation.

Figure 1. The experiment screen. Reduced in size for presentation and including the point values of the dials.

Appendix B: Experiment 2 Instructions

Experiment 221: Eye Tracking a Visual Monitoring Task
Instructions

The Task:

Your task in this experiment is to detect alarms on a group of four dials, and you will be asked to do detect alarms in two separate conditions. In the first condition, (Figure 1, next page), an alarm occurs whenever the needle of one of the dials enters the dark gray region of the dial. For instance, in Figure 1 (page 3) an alarm has occurred on the lower-left dial. Whenever an alarm occurs on a dial, your task is to press a key corresponding to that dial. The keys corresponding to the upper-left and the lower-left dials are the "A" and "Z" keys, respectively. The keys corresponding to the upper-right and the lower-right dials are the semicolon (;) and period (.) keys, respectively. Note that the location of these keys on the keyboard correspond to the location of the dials on the screen, i.e., the "A" key is the upper-left of the four keys and corresponds to the upper-left of the four dials. You'll need to keep your fingers on these four keys throughout the course of the experiment.

In the second condition, your basic task will remain the same, detecting alarms. However, it will become a bit more difficult. In this version of the task, only one of the four dials will be visible at any time (Figure 2, page 3). You select which dial you would like to be visible by pressing the key corresponding to a particular dial. For instance, when you would like to view the upper-left dial, you would press the "A" key. In order to indicate that an alarm has occurred on a dial, you must press the key corresponding to the visible dial. For instance, when the upper-left dial is visible, pressing the "A" key indicates that

you have detected an alarm on that dial. Hence, you may only indicate alarms on the dial that is currently visible. Pressing the "Z" key when the upper-left dial is visible tells that system that you would like to view the lower-right dial, and the view will change such that the lower-right dial is visible and the upper-left dial is occluded from view.

For the first hour and a half of the experiment, your task will be of the easier type, where all four dials are visible. For the remaining hours of the experiment, your task will be the more difficult version, where only one dial is visible at a time.

Scoring:

You will be awarded points based on the correct detection of alarms. Each dial has a number of points associated with it, and you will be given the number of points for a dial whenever you correctly detect an alarm on that dial. The number of points associated with each dial is shown below each dial in Figure 1, such that pairings are upper-left: 10 points, lower-left: 6 points, upper-right: 2 points, lower-right: 4 points. Also, in the alarm range of each gauge, there are bars of different lengths. These bars are used to indicate the number of points each gauge is worth. Hence, the longest bar is worth 10 points, the bigger medium-sized bar is worth 6 points, the smaller bar is worth 4, and the smallest bar is worth 2 points. Each time you correctly detect that a gauge has gone out of bounds you will receive the number of points indicated by the length of the bar. Conversely, you will receive negative points for misses (any time that an alarm occurs on a dial and you do not detect it) and false alarms (when you push the button associated with a dial and there is no alarm on that dial). The number of negative points you will receive is the same

as the number of positive points you would receive for the correct detection of an alarm on that dial.

Your overall score will be the positive points you earn for each time you correctly detect that a gauge has gone out of bounds minus the points you loose for both misses and false alarms. Since there are many ways to loose points, it is possible (or even likely) that you will have negative points – especially at the beginning.

Your total score will be given to you at the end of the each 5-minute trial. There are eight trials per one-hour experimental session.

Eye Tracking:

The task itself is a very easy one, monitoring the dials and pressing a button when an alarm occurs. However, we are not only interested in measuring your performance at the task but also in monitoring your eye movements. We are investigating how when and where you look changes as you get better at the task. The more difficult (and often more frustrating) part of the task involves working with the eye tracking equipment so that we may record your eye movements. You will only be working at the task for 30 to 40 minutes each one-hour experiment session. The other 20 minutes or so of each hour will be spent working with the eye tracking equipment. We just thought you should be warned.

Awards:

There are five participants in this experiment. The participant with the highest average score over the course of the five experimental sessions will receive a $15 gift certificate to the web retailer of their choice (e.g. Amazon.com). The second and third highest average scorers will receive gift certificates for $10 and $5, respectively. Additionally, on each session on which your average score is above zero (0), a couple pieces of candy will be your just reward.

If you have any questions, please don't hesitate to ask the experimenter at any time. Thank you for your participation.

Figure 1. The experiment screen, condition 1. For instruction purposes, the point values of each of the dials are shown just below each dial. Note that an alarm has occurred on the lower-right dial.

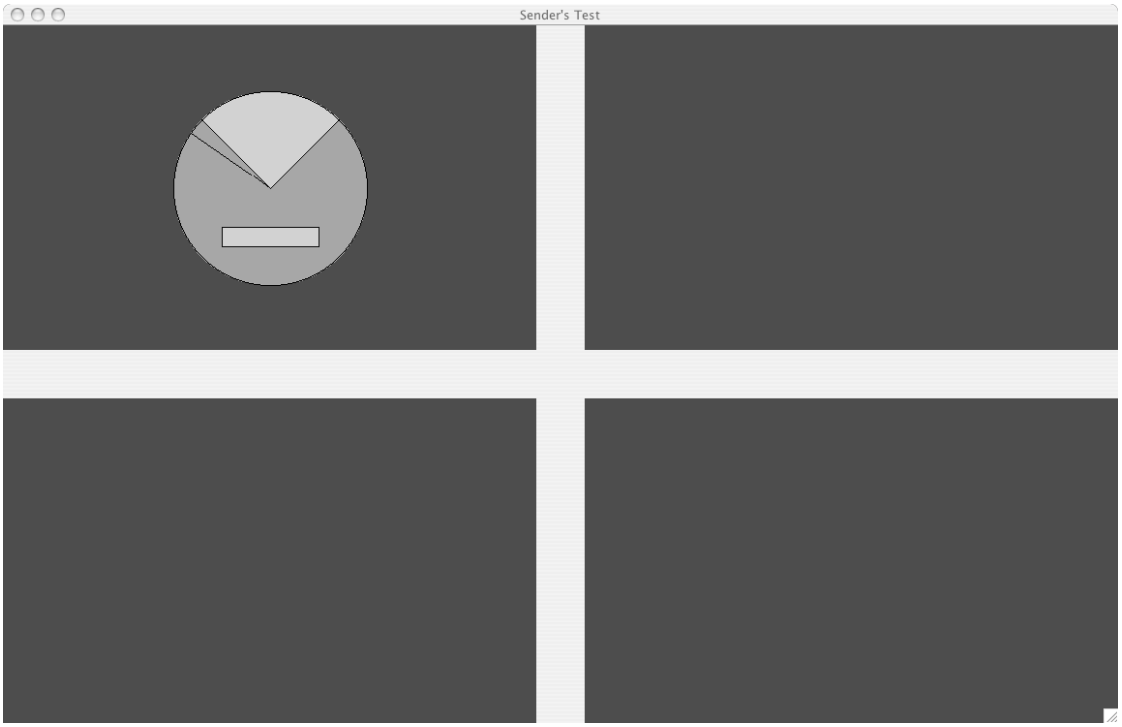


Figure 2. The experiment screen, condition 2, one-dial-visible condition. An alarm has occurred on the visible dial.

Appendix C: Experiment 3 Instructions

Experiment 221: Eye Tracking a Visual Monitoring Task
Instructions

The Task:

Your task in this experiment is to detect alarms on a group of four dials (Figure 1, page 3). An alarm occurs whenever the needle of one of the dials enters the dark gray region of a dial. For instance, in Figure 1, an alarm has occurred on the upper-left dial. Whenever an alarm occurs on a dial, your task is to press a key corresponding to that dial. The keys corresponding to the upper-left and the lower-left dials are the "A" and "Z" keys, respectively. The keys corresponding to the upper-right and the lower-right dials are the semicolon (;) and period (.) keys, respectively. Note that the location of these keys on the keyboard correspond to the location of the dials on the screen, i.e., the "A" key is the upper-left of the four keys and corresponds to the upper-left of the four dials. You'll need to keep your fingers on these four keys throughout the course of the experiment.

Scoring:

You will be awarded points based on the correct detection of alarms. Each dial has a number of points associated with it, and you will be given the number of points for a dial whenever you correctly detect an alarm on that dial. The number of points associated with each dial are: upper-left, 10 points; lower-left, 2 points; upper-right, 8 points; lower-right, 6 points. To help you remember the relative values of the dials, in the alarm range of each gauge there are bars of different lengths. These bars are used to indicate the number of points each gauge is worth. Hence, the longest bar is worth 10 points, the

bigger medium-sized bar is worth 8 points, the smaller bar is worth 6, and the smallest bar is worth 2 points. Each time you correctly detect that a gauge has gone out of bounds you will receive the number of points indicated by the length of the bar. Conversely, you will receive negative points for misses (any time that an alarm occurs on a dial and you do not detect it) and false alarms (when you push the button associated with a dial and there is no alarm on that dial). The number of negative points you will receive is the same as the number of positive points you would receive for the correct detection of an alarm on that dial.

In order to receive points for the detection of an alarm, you must detect the alarm within one second from the time it entered the alarm region. If you press the button for an alarm after one second has elapsed, you will receive negative points, i.e., detecting an alarm after one second has passed will be considered a false alarm. (In fact you will receive double negative points in this case—negative points for originally missing the alarm and negative points for a false alarm.) Also note that the needle must be completely within the alarm region to be considered an alarm—right on the edge of the region is not an alarm.

Your overall score will be the positive points you earn for each time you correctly detect that a gauge has gone out of bounds minus the points you loose for both misses and false alarms. Since there are many ways to loose points, it is possible (or even likely) that you will have negative points – especially at the beginning.

Your total score will be given to you at the end of the each 5-minute trial. There are eight trials per one-hour experimental session.

The Flashing Dial:

In order to make the task slightly easier, one of the dials will flash (or blink) every time an alarm occurs on that dial. It will continue to flash as long as the needle remains in the alarm region. You will quickly learn which dial has this property. However, when an alarm occurs, you still must press the button associated with that dial in order to receive points for detecting the alarm.

Eye Tracking:

The task itself is a very easy one, monitoring the dials and pressing a button when an alarm occurs. However, we are not only interested in measuring your performance at the task but also in monitoring your eye movements. We are investigating how when and where you look changes as you get better at the task. The more difficult (and often more frustrating) part of the task involves working with the eye tracking equipment so that we may record your eye movements. You will only be working at the task for 30 to 40 minutes each one-hour experiment session. The other 20 minutes or so of each hour will be spent working with the eye tracking equipment. We just thought you should be warned.

Awards:

There are five participants in this experiment. The participant with the highest average score over the course of the five experimental sessions will receive a $15 gift certificate to the web retailer of their choice (e.g. Amazon.com). The second and third highest average scorers will receive gift certificates for $10 and $5, respectively. Additionally, on each session on which your average score is above zero (0), a couple pieces of candy will be your just reward.

If you have any questions, please don't hesitate to ask the experimenter at any time. Thank you for your participation.

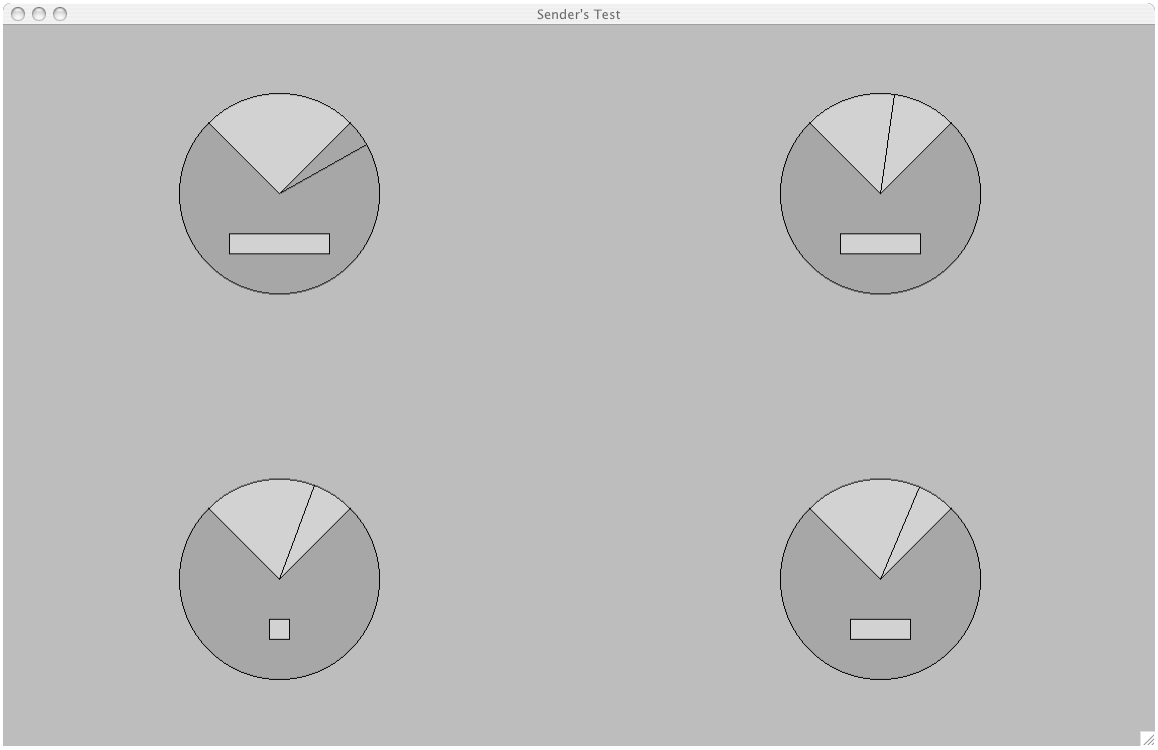


Figure 1. The experiment screen. An alarm has occurred on the upper-left dial. The length of the bars on each dial correspond to the points associated with each dial, such that the point values of the dials are (clockwise from upper-left) 10, 8, 6, and 2.