SIGNAL DETECTION THEORY IN THE ANALYSIS AND OPTIMISATION OF INDUSTRIAL INSPECTION TASKS

A Thesis submitted for the degree of

Doctor of Philosophy

by David Edward Embrey BSc.

Department of Applied Psychology

University of Aston in Birmingham

September 1976

SUMMARY

The overall aim of this study was to investigate the applicability of Signal Detection Theory (SDT) to a number of problems in the area of industrial inspection, including training and selection. Two industrial studies comprising three experiments are presented, together with four laboratory experiments and a correlational study.

The first three chapters of the thesis comprised a comprehensive review of SDT and the literature of inspection.

Chapter 2 described an industrial case study, the inspection of photographs of nuclear particles, designed to test the applicability of SDT in an applied setting. The variables of auditory noise, defect complexity and time on task were also considered. SDT in its unequal variance form was found to fit the data. The next case study attempted to apply SDT to the inspection of photographic film. A two stage decision making model was proposed to describe performance in this task.

The first two laboratory studies investigated the effect on performance of within and between session changes in defect probability. It was found that the subject could adjust his criterion appropriately to between session changes if feedback was provided, and to within session changes if he received prior warning of the change.

The final laboratory studies were concerned with training the inspector's ability to modify his criterion, and the enhancement of his sensitivity. The first experiment replicated previous work in

perceptual training, and the second utilized a wide range of differing training techniques. It was found that certain combinations of conditions were significantly superior in achieving the training goals.

A correlational study was conducted utilizing the results of the previous two experiments and tests of various cognitive skills. Significant correlations were found between certain groups of test scores and performance on the task. These tests were proposed as potential selection techniques for inspectors.

"Quality is shapeless, formless, indescribable. To see shapes and forms is to intellectualize. Quality is independent of any such shapes and forms. The names, the shapes and forms we give Quality depend only partly on the Quality. They also depend partly on the a priori images we have accumulated in our memory. We constantly seek to find, in the Quality event, analogues to our previous experiences. If we didn't we'd be unable to act. We build up our language in terms of these analogues. We build up our whole culture in terms of these analogues."

Pirsig R.M., ZEN and the ART of MOTORCYCLE Maintenance.

ACKNOWLEDGEMENTS

I would like to express my gratitude to everybody who has helped and encouraged me during the long gestation period of this thesis. David Whitfield, my supervisor, was always ready to offer assistance, despite his heavy timetable. Mrs. Christine Maddison deserves my thanks for both deciphering my handwriting and for producing accurate typescripts at remarkable speed.

Acknowledgements are due to Professor W. T. Singleton for providing Departmental facilities and to the technical staff including, Les Bagnall, Paul Bernard and Colin Mason, who assembled the experimental equipment. The case studies were made possible by the generosity of Professor D. C. Colley of the Physics Department, University of Birmingham, and Mr. R. F. Salmon of Ilford Ltd. All the subjects deserve my thanks for the often arduous experimental sessions they endured.

Last, but not least, I must thank all my close friends, whose apparently unlimited willingness to work hard on my behalf made everything possible.

Contents

CHAPTER 1 INTRODUCTION

1.0 INTRODUCTION

In preparing this study, an attempt has been made to satisfy a need that has become increasingly apparent to ergonomics practitioners: the provision of data from experiments that have direct relevance to situations encountered in real world tasks. Although ergonomics is essentially an applied science, the orientation of much research has been towards purely laboratory based studies that provide little in the way of information which is readily applicable in an industrial context. Chapanis (1967) discusses this problem in detail. In an emerging science this bias is perhaps understandable. However, ergonomics and human factors have now been in existence for nearly thirty years and should by this time be producing such data on a large scale. In order to achieve this aim, it seems clear that research work needs to change its orientation from the formalized consideration of the effects of a few selected variables in highly specialized laboratory studies, to a more pragmatic approach yielding results which can be utilized more readily in real-life applications. This approach does not mean that compromises necessarily have to be made in standards of experimentation. It is a question of broadening the experimental focus rather than of producing work that has no theoretical significance.

An attempt has been made to make this study sensitive to these needs in a number of ways. Rather than considering a narrow research topic in some depth and then attempting to show that it also has practical relevance, the starting point of this study is a whole area of application within the industrial sector, that of quality control. Since this study is primarily concerned with human factors, the main consideration will be given to the inspection aspect of quality control.

The approach has been to show how data from a number of research areas contribute towards the goal of optimizing those aspects of a quality control system in which human beings play a predominant role. The utility of this approach is that it brings together data from a number of disparate areas in such a way that their effect on the inspection function can be clearly seen and some insights gained into the ways they may interact. The aim is to provide a usable body of information for the practitioner on the factors known to affect inspection. Of course, such a review also serves the more usual function of a literature survey in that it suggests potential areas of further research.

From the practitioner's standpoint, the data available in such a review will be most useful if it is classified in a manner clearly related to attributes common to a large proportion of real inspection tasks.

Another way in which this study attempts to maintain relevance to real world problems is in the experimental phase. The impetus for the laboratory based work is provided by field studies which clarify the nature of the important variables which need further consideration in a more controlled environment. The laboratory studies then seek to simulate the critical features of the real life task to provide directly relevant data. At the same time the experimental studies provide a vehicle in which a number of more theoretical ideas can be investigated.

In addition to the areas outlined above, the study attempts to assess the utility of considering the area of inspection from a particular theoretical standpoint - that of Signal Detection Theory, which serves as a unifying concept for a wide range of experimental work relevant to inspection.

The general aim then, can be summed up as an attempt to advance knowledge on a broad front, in a manner related to practical needs.

1.1 The importance of industrial inspection as an area of study

As the increasing application of technology to manufacturing industries reduces the importance of purely manual, motor skills, attention is being increasingly focussed on areas of industry in which higher level perceptual abilities of the operator, such as pattern recognition and decision making, are employed. Industrial inspection is such an area.

The reason why the human operator continues to be important in the quality control area is that these higher level functions cannot readily be performed by machines. Although automated pattern recognition is possible with sufficiently simple patterns, and computer aided decision making is being utilized in certain situations, (see Whitfield (1975) for a review of this area), the extremely high cost of such devices, and their inflexibility relative to a human operator, means that they are unlikely to find wide application in the industrial inspection area. The attribute of quality has a multidimensional nature which can encompass both simple parameters such as specification of size, as well as complex aesthetic judgements. The possibility of building machines to handle the whole range of quality judgements seems remote, although it seems feasible, and indeed desirable, to use automated techniques where very simple discriminations such as size and weight are required between acceptable and non-acceptable articles.

It is found in industry that even when simple judgements of this type are required, it is common to inspect manually. In fact Fox (1973)

states that 90% of all inspection in the United Kingdom is dependent on the unaided human operator. The reasons for this are usually straightforward cost-effectiveness considerations. The volume of the product to be inspected may not justify the design and manufacture of an automatic device to monitor quality. Alternatively, quality specifications may change frequently and the human operator is more readily 'reprogrammed' than his machine counterpart.

It seems clear that inspection is likely to remain a labour intensive area, and as such, ergonomics and human factors will continue to be able to make important contributions in optimizing the performance of the human operator.

Another important reason for studying inspection is that research in this area is not restricted in its usefulness purely to the industrial sector. Many other important tasks have characteristics very similar to those found in industrial inspection. For example the surveillance of radar screens in both military and civil applications can be seen to contain many elements common to inspection tasks. The operator is continuously monitoring an information source which provides signals which are complex in nature, infrequently occurring and unpredictable both in time and space. In the medical sphere, the examination of medical X-ray plates and cervical cancer smears are examples of almost classical inspection tasks which employ vast resources of trained manpower. A similar inspection problem occurs in high energy nuclear physics research, in which millions of bubble chamber photographs are scanned for patterns of tracks. This task will be considered in detail in chapter 4.

A final important reason for applying human factors and ergonomics principles to the optimization of inspection systems can be found from considerations of system reliability. The inspection phase of system development is a vital link between the manufacture of components and their incorporation in a total system. Many disastrous system failures can be traced to inadequate inspection procedures. Meister (1971) discusses the importance of inspection from a system reliability standpoint.

From the above discussion we can conclude that although inspection as an area of study has received relatively little attention compared with tasks in industry involving mainly motor skills, it is an area of considerable intrinsic interest and importance. With growing automation it is likely to grow in importance, by comparison with traditional manufacturing tasks. Additionally, inspection-like tasks are found in many spheres distinct from the industrial sector, and hence any research findings are likely to be widely applicable.

1.2 Characteristics of inspection tasks

One of the most obvious characteristics of inspection tasks is their diversity. Virtually any item which is manufactured is likely to be inspected at some stage of manufacture or assembly. Although the visual modality is the usual one employed in inspection, other sensory inputs are sometimes employed. Thomas (1962) describes how vacuum cleaners were inspected for mechanical faults by the tester actually listening to the sound the cleaner made in operation. Several categories of fault could be distinguished by this method. Needles are inspected for straightness by gently rolling them under the palms on a flat table. Frequently combinations of sensory modalities are employed, as when an

article is inspected for surface finish both by its appearance and by its tactile characteristics.

Perhaps the most common form of inspection is when the inspector sequentially examines a series of items to determine whether the characteristics of the items fall within the quality specification. Even in this case a more detailed task description rapidly leads to complications. For example the items may be on a conveyor belt, moving at a range of speeds or the task may be completely self paced. They may be large items, in which case some form of scanning may be required, or they may be small enough to be examined in a single fixation. The defects may be visible with the naked eye or may require enhancement by means such as magnification or lighting. They may be distinguishable from non-defects by an extremely large range of variables including shape, size, colour, texture, weight and any combination of these and other parameters. This is an additional reason why automated inspection is often impracticable.

Another common type of inspection is the examination of materials in a continuous form. Examples of this type of task are the examination of cloth, steel and glass strip and cine film, which will be discussed as a detailed study in chapter 5 of this thesis.

In chapter 3 the various characteristics of inspection tasks will be considered in a more systematic manner. For the moment it is useful simply to be aware of their diversity and to attempt to describe a common behavioural structure which applies to the majority of inspection tasks. This will be considered in the next section.

1.3 An informal model of the inspection task

As was implied in the last section, the types of inspection task that occur show such diversity that any specific task cannot be representative of the whole range of situations encountered. The majority of inspection tasks do, however, share certain common characteristics, and in this section an informal structure will be described which facilitates the consideration of the tasks from a psychological and ergonomics standpoint.

The model to be described is not predictive in nature, although it could provide the basis for a simulation approach to the predictive modelling similar to that described by Siegel and Wolf (1969) in the context of man-machine systems. For our present purposes, the model will serve to indicate the areas of knowledge relevant to inspection situations in general.

It is convenient to consider inspection as consisting of four broad phases (Figure 1).

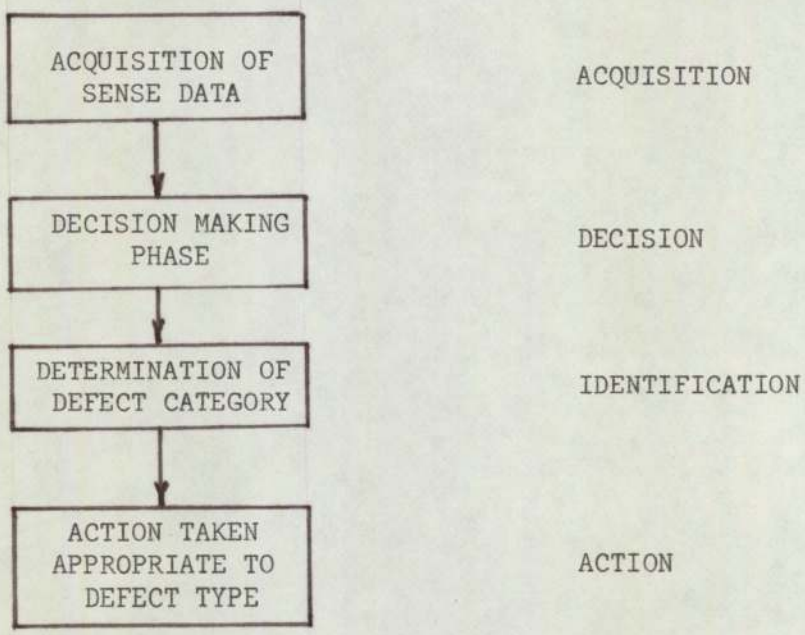


Figure 1. Phases of inspection.

a. Acquisition of sense data. In order that the inspector can decide on the presence or otherwise of a defect, he requires sensory evidence from the item being examined to provide inputs for the decision making phase.

b. Decision making. This aspect of the model refers to the process whereby the inspector assigns the item to the general categories of defect or non-defect, without further sub-categorization.

c. Identification. This is clearly also a decision making process which can be regarded as one stage higher in the decision hierarchy. It involves the classification of the defect into one of a number of sub-categories, if more than one exists

d. Action. Once the nature of the defect has been ascertained the inspector performs the action appropriate to that class of defect, e.g. rejects the item, returns it for reworking etc.

Each of the phases considered can be analysed in terms of the psychological and physiological factors which affect performance at each phase. Examples of these are set out in Figure 2.

1.3.1 Acquisition of the data

The acquisition phase has been considered as being affected by situational, physiological and psychological factors, although many of these could be included under more than one heading.

Situational factors are those which are external to the inspector and include environmental variables as well as the specific attributes of the task itself. The physiological factors affecting target acquisition are primarily visual, reflecting the preponderance of this modality in inspection tasks. Environmental conditions are included under

Figure 2. Examples of psychological and physiological factors affecting various phases of inspection.

1. ACQUISITION FACTORS

(a) Situational

I Paced or unpaced presentation, rate of pacing.
II Enhancement of discriminability of defect, e.g. X-rays, ultrasonics, magnification, lighting.
III Inherent discriminability of defect.

(b) Physiological

I Visual acuity, static or dynamic.
II Visual skills in general, e.g. colour vision.
III Environmental conditions affecting inspection performance e.g. heat, noise, lighting levels.
IV Visual fatigue

(c) Psychological

I Perceptual 'set', i.e. ability to recognize cues characteristic of defects as opposed to other configurations occurring in both defective and perfect product.
II Visual search strategies.
III Vigilance and attentional variables.
IV Organismic variables such as pattern recognition skills, field dependence and distractability.

2. DECISION MAKING FACTORS

I Expected incidence of defects.
II Costs associated with missing defects and rejecting good products.
III Social factors

3. IDENTIFICATION FACTORS

I Training and experience.
II Provision of reference standards.
III Expectancies concerning type of defect likely to occur.
IV Number of categories of defect.

4. ACTION FACTORS

I Existence of clearly defined actions to be taken for various types of defects.
II Consequences of action.
III Social factors.

physiological factors even though they may affect functioning via decrements in psychological skills. The psychological factors are intended to represent the intra-subject variables which influence target acquisition. The question of the perceptual skills which influence the probability of target detection will be discussed in some detail in this study. In including perceptual skills within the phase of acquisition we are referring to the inspector's sensitivity for cues which identify the sample being examined as belonging to the categories defect or non-defect. These cues are often indirect and the inspector may have to infer the existence of a hidden defect from the external evidence available. It is important to note that we are referring to a concept of sensitivity which is independent of the inspector's tendency to respond 'defect' or 'non-defect' as a result of prior knowledge of the probability of a particular item being defective, or because he will be heavily penalized if a good item is incorrectly rejected. These latter factors are considered to be decision making variables.

The concept of intrinsic sensitivity refers to the ability of the operator to effect categorization utilizing the evidence available in the sample. As has been noted earlier, we regard this facility as being uncontaminated by a bias to respond defect or non-defect due to data other than that available from the sample. Whether it is possible to separate sensitivity from 'response bias' due to other factors is a question we shall pursue at length in chapter 2.

It is clear that the inspector's ability to distinguish good from bad products will depend partly on his knowledge of the cues which indicate defectiveness and partly on the amount of data that he can acquire from the sample. A third, slightly more controversial factor, concerns his intrinsic ability, independent of training, to isolate a particular

configuration of cues embedded in a confusing background.

The utilization of cues present in the sample is clearly a function of training and also the provision of reference aids defining defective items. This topic will also be explored more fully later. The inspector's ability to acquire information from the sample is strongly influenced by physiological factors such as visual acuity discussed earlier. The area of visual search has been included in the psychological factors although it has a strong physiological element. Visual search skills consider the ability of the inspector to search an area exhaustively, efficiently and rapidly. Many of the studies of visual search which will be considered later have been concerned with a subject's ability to economically scan large areas for a target. To this extent the prime interest in search strategies is in the area of the inspection of sheet materials for defects. Clearly there is a high degree of interaction between some of the factors being considered. For example if the situational factor of pacing is very high the question of the time taken by the inspector to scan an item becomes highly important.

The relevance of research on vigilance tasks to industrial inspection is a function of the degree to which the task under consideration approaches that of the classical vigilance decrement situation. If an inspection task is conducted for prolonged periods in an unstimulating environment, with low probability, irregularly occurring defects, then vigilance effects might be expected. As will become apparent in the literature survey, there is some controversy as to the applicability of much vigilance research to industrial situations.

The possibility that an inspector's detection skill is related to innate abilities such as field dependence and pattern recognition is an

intriguing one that has not yet been explored by workers in the inspection field. The whole area of individual differences is one which has been neglected, particularly in relation to the selection of inspectors. It is hoped that this study will provide an impetus for research in this area.

1.3.2 Decision making in inspection

Many of the issues relevant to the decision making aspects of inspection skills cannot be discussed at this point because they are an integral part of the Signal Detection Theory orientation to the area which will be considered in detail later. At this stage it is sufficient to point out that one would expect an inspector's decision about whether or not a sample is defective to be influenced not only by the evidence available from the sample but by the expected incidence of defects in the whole series of samples. Similarly it is reasonable to expect the consequences of a particular decision to influence the inspector's judgement. For example, if the item being inspected is a critical part for a spacecraft, then the inspector will be far more likely to reject it if there is even a suspicion that it is defective, than if it were a non-critical item. These commonsense notions will be related to a theoretical structure in a later section.

Social factors can be seen to affect decision making in a situation where a worker is able to exert social pressure on an inspector if too large a proportion of his work is being rejected. Such pressures can be quite subtle and overt threats are not necessary to influence the judgement of the inspector.

1.3.3 Identification factors

Identification is distinguishable from acquisition in that the inspector is attempting to decide between different categories of defect after having decided that a particular signal is in fact a defect. Most of the factors operative at the decision making phase are also important here, although it is essentially a multiple categorization problem rather than a binary decision making process, at least where there are several types of defect. As before, training and experience will be important in allowing accurate differentation between types of defect. Of particular importance is the provision of reference standards in order to provide examples of the distinguishing characteristics between defects. As in the decision making phase, expectancies concerning the type of defect likely to occur will influence the categorization process.

1.3.4 Action factors

The absence of clearly defined actions to be taken in the event of the various types of defect occurring can lead to a considerable degradation in the efficiency of the inspection system. Some defects may indicate the presence of certain manufacturing malfunctions and hence necessitate a rapid feedback to the production section of the factory. Other defects may require reworking rather than rejection. Certain types of defective items may be acceptable to some customers who may be selling at discount to a less discriminating market. The consequences of certain actions may affect the inspector's behaviour in the same way as during the decision making phase of inspection.

Social factors need to be considered again in this context. The prevailing employment situation in a factory might, for example, influence

an inspector in deciding whether a particular item was reworkable or should be scrapped. The action phase of inspection has received little attention in the literature, and should perhaps be considered more explicitly in the analysis of inspection systems.

1.4 Conclusion

The consideration of an informal model of the inspection task has provided an overview of many of the topics which will be considered in more detail in subsequent chapters. If quantitative estimates were available for the effects of the variables considered on inspection performance, then it would be possible to use a model of this type for predictive purposes. As will become apparent during the subsequent review chapters, however, we have a considerable way to go before we can realistically assess the combined effects of some of the variables considered in the model, on inspection performance in general. Nevertheless the predictive modelling of the performance of human operators in an inspection system should be regarded as a desirable long term objective.

The validity of industrial inspection as an area of study for the behavioural sciences was also established in this chapter. It was pointed out that inspection requires a wide range of cognitive, decision making and pattern recognition skills which cannot readily be automated, and hence quality control is likely to remain a labour intensive area of industry. Data from inspection studies can be readily generalized to a number of other important areas such as radar screen surveillance and any situation where prolonged monitoring or repeated perceptual decision making takes place. Finally the quality of an inspection system has a considerable effect on the overall

reliability of a system, which is becoming increasingly important as larger and more complex systems are produced.

CHAPTER 2 SIGNAL DETECTION THEORY AND ITS APPLICATION TO INSPECTION

2.0 INTRODUCTION

In the development which follows, Signal Detection Theory, (hereinafter referred to as SDT) will first be briefly considered from an historical standpoint, emphasizing the psychophysical ideas from which it emerged. Next the important concept of response bias will be considered, with particular reference to inspection and other industrial applications.

The theory will then be developed in its simplest form, the equal variance model, together with the experimental evidence for this position, and then the more general form will be considered. Methods of analysing situations in which the simplified assumptions do not apply will be discussed in detail, particularly from the point of view of applying SDT to real as opposed to laboratory tasks.

The existing studies in which SDT has been applied to inspection tasks will be reviewed in detail and general rules for applying SDT to this area will be set out. Finally potentially rewarding areas of applicability of the theory to inspection will be summarized and directions for further research proposed.

In line with the applied nature of this study, no attempt will be made to produce a mathematically sophisticated analysis of SDT, and therefore detailed mathematical justifications of various points will be referred to standard texts.

2.0.1 Historical background

SDT was developed in its modern form by two groups of researchers working independently at Michigan and Harvard Universities on the

problems of detecting signals in noisy channels, particularly in the context of radar (Swets, 1963). In fact its origins can be traced back as far as Fechner (1801-1887). Fechner observed "the great variability of sensitivity due to individual differences, time, and innumerable internal and external conditions" (Fechner, 1860 p.44), that was found with human detection performance. He hypothesized the existence of a physiological threshold above which a stimulus of a particular intensity could be perceived and below which detection could not occur.

However, evidence began to accumulate that the position of the threshold was influenced by external factors such as the probability of occurrence of the signal over a series of trials, assumed by the subject. Subjects who had been trained to expect a high incidence of stimuli invariably had a lower threshold than those who expected a low probability of stimuli occurrence. In classical psychophysical terms, the subjects were committing the 'stimulus error' by basing their reports on external characteristics of the stimulus rather than their perceived sensations. Another problem was the question of false alarms. If a fixed threshold existed, the observer must either be in a 'detect' or 'non-detect' state. For this reason the occurrence of a 'stimulus' response on a null trial was difficult to account for. In early experiments subjects producing false alarms were simply admonished by the experimenter to take more care. This of course had the effect of biasing them to respond negatively if uncertain.

Later experimenters such as Thurstone sidestepped the problem of response bias by using paired comparison or forced choice techniques. With these methods, a series of pairs of trials are presented, one of each pair containing the stimulus, and the subject simply has to

indicate which one of each pair is the signal trial. However, in many real detection situations the forced choice paradigm is inappropriate.

Blackwell (1952) was concerned with the large scale determination of absolute visual thresholds. By this time it had been realized that even during a trial in which no stimulus was presented, the nervous system of the observer was continually active. Random firing of neurons occurs, exercising a tonic effect on brain functioning, (Pinneo, 1966). It seems likely therefore, that in the case where near threshold signals were to be detected, then sometimes the magnitude of this 'noise' distribution, even on non-signal trials, might be sufficiently great to be construed as a signal. Blackwell's assumption (known as 'high threshold theory') was that the observer's threshold was sufficiently high such that the magnitude of the background noise would never exceed it and give rise to a spurious signal report ('false alarm'). How then, could the undoubted occurrence of false alarms in this situation be accounted for? Blackwell assumed that on a certain proportion of the non-signal trials the subjects simply guessed, incorrectly, that a signal had been presented. A 'guessing correction' was therefore applied to the data to take into account this tendency of subjects.

This assumption, in retrospect, seems unsatisfactory on a number of counts. It is rather arbitrary and still provides no satisfactory account of the effects of, and the reasons for, judgemental bias on the part of the observer.

It seems then, that by the beginning of the fifties the problems of the satisfactory analysis of detection experiments had not been solved and the time was ripe for the introduction of a new theory. Before considering this theory however, it is useful to consider the concept of

observer bias in more detail.

2.0.2 The nature of response bias

Judgemental bias is an all-pervasive aspect of any human choice or decision making situation and is not confined to the rather narrow area of psychophysical judgements that we have just been considering.

A doctor looking, listening or feeling for signs of a disease may far prefer a 'false alarm' to a missed signal, particularly if the disease is serious. In a control room of a nuclear power station the operator may be required to shut down the reactor in the event of certain evidence from the instrumentation, which may be ambiguous, that a dangerous condition has occurred. Shutting down the reactor is usually an expensive business and may cost tens of thousands of pounds in terms of lost output. On the other hand not shutting down the reactor may be even more expensive in terms of damage to plant or even loss of life. The operator's decision will clearly be influenced by these cost factors and also by the relative probability that a dangerous condition is really likely to occur. If it is during the commissioning phase, and similar incidents have been quite common, then the operator will have no hesitation in shutting down the reactor. If on the other hand the plant has been in operation for some years and such an incident has never before occurred, the operator may well defer his decision for some time before he makes the critical control action.

In the area of inspection we can see that the inspector's criterion is going to be influenced by the on-going level of defects present. If it is known that a particular batch has been produced by a 'rogue machine', then the inspector will be far more likely to reject an article that is

a borderline case. A similar effect could occur where cost factors were involved. If the manufacture of an article involved an extremely expensive series of processes, the inspector would be biased to reject the article only if he was absolutely certain that it was defective. Similarly, if the manufacturing process were very cheap but the product was destined for a highly discriminating market, then the inspector would reject if there were any doubt at all that the item was not perfect.

It is important to note that we are referring to subjective probabilities in this discussion, and the costs and values of the observer's decisions are in fact personal utilities. These quantities may or may not be the same as the objective probabilities and payoffs which occur in a real situation. Viewed in these terms we can see that many factors which are known to affect the judgemental process can be mapped on to the dimensions of subjective probability or perceived utility. For example social pressures from co-workers in an industrial situation could be seen to affect judgement via the mechanism of altering the personal payoffs of a particular decision. If an inspector is examining the work of a colleague and is aware that too many rejections could lead to his dismissal, then he may be influenced, consciously or otherwise, to accept a higher proportion of borderline items than would be the case if they arrived from an anonymous source. The subjective estimates of the probability of a defect occurring would clearly be highly dependent on the degree of feedback given to the inspector.

Up to this point we have demonstrated that in detection situations, unless forced choice procedures are adopted, any measure of the observers sensitivity is inevitably contaminated by the factors of response bias discussed in this section. One factor that has not been mentioned up to

this point is the question of quantifying the degree of observer bias. The attitude of classical psychophysics has been that observer bias is a 'nuisance variable' that one should attempt to eliminate as far as possible. However, it is clear that the degree of response bias that occurs in a situation is in itself a quantity of some interest. Given that judgemental bias is a fact of life in any real discrimination task, it is clearly of interest to quantitatively measure the amount of bias, in order to answer such questions as the degree to which some of the factors discussed up to this point affects it.

In subsequent sections we will see that SDT provides the means for quantifying aspects of response bias that have been discussed in this section.

2.1 SDT - general considerations and the basic model

A number of general reviews of SDT now exist, e.g. Pastore and Scheirer (1974), Swets et al. (1961), Egan and Clarke (1966), Coombs et al. (1970), Lee (1971) and Swets (1973). Two textbooks and a book of collected papers have also been produced: Swets and Green (1966), McNicol (1972) and Swets (1964).

Signal Detection Theory has applications in a very wide range of situations in which an observer has to make a discrimination or choice on the basis of equivocal or 'noisy' evidence.

Noise is a central concept in SDT and can be regarded as any random process which tends to interfere with discrimination. We can consider two analogous situations, the detection of a very faint signal in a background which may tend to degrade the signal, and discrimination between

two very similar signals which may also be embedded in a background which tends to obscure the differences between them.

Another source of noise might be the inevitable slight variations in the physical nature of the stimulus, particularly if its precise characteristics are ill-defined. These sources of noise are all external to the observer, but noise is also added internally due to the random firing of neurons in the nervous system (Pinneo, op. cit.).

Whenever a subject makes an observation, the sensory effect that occurs can be represented as a point y in an n-dimensional space, the n dimensions representing the n possible characteristics of the response of the sensory system. For example an inspector may be examining coins for defects by their visual appearance, weight and the sound they produce when dropped. In this case we might expect the sensory evidence x to consist of visual parameters such as size, shape, colour etc., tactile evidence including weight and texture, and auditory information such as the slight difference in sound of a 'dud' coin. It is clear that each coin examined will produce a slightly different point y due both to slight variations within the categories 'perfect' and defective coins and also to the other internal and external sources of noise discussed earlier. It is obvious that the distribution of y values due to good coins, $f_N(y)$, will have a mean less than that due to faulty ones, $f_{SN}(y)$, and that the two distributions will overlap. (The subscript N refers to the fact that the former distribution is regarded as being due to noise only, and SN as the latter being due to signal + noise, the signal in this case being the attributes of a defective coin). If we regard $f_N(y)$ and $f_{SN}(y)$ as being the sensory evidence on which the inspector bases his decision, the question arises as to how he decides whether a perfect or defective item has been presented on a particular

trial.

Signal Detection Theory assumes that as a result of experience on the discrimination task the observer is able to compare the probability that a particular sensory effect has arisen, given that a perfect coin was presented, compared with the probability that it arose, given that a defect was presented. To do this, he forms a likelihood ratio from the ordinates of the two density functions $f_N(y)$ and $f_{SN}(y)$ corresponding to the particular value of the sensory evidence on that trial, i.e.:

$$\frac{f_{SN}(y)}{f_N(y)} = \lambda(y), \text{ the likelihood ratio}$$

The likelihood ratio, then, represents the likelihood that the point y arose from the defect (SN) distribution relative to the likelihood it arose from the perfect (N) distribution. Since any point in the space, i.e. any sensory datum, may thus be represented as a real, non-zero number, all sensory data can be regarded as lying along a single axis. Any observation x can therefore be identified with a particular value of the likelihood ratio $\lambda(y)$. It is convenient if x is identified with a transformation of $\lambda(y)$, i.e. $\lambda(x)$, such that Gaussian density functions of the sensory effects of signal and noise, $f_N(x)$ and $f_{SN}(x)$, result. This is produced by using a logarithmic transform of $\lambda(y)$. A further assumption will be made, although this will be modified later, that the distributions are of equal variance. This gives rise to the familiar SDT diagram, Figure (2.1). The normality assumption can be justified on the basis of the Central Limit Theorem, in that if observations are independent, then the distribution of the sums of the noise and signal plus noise distributions each approach normality for reasonably sized samples.

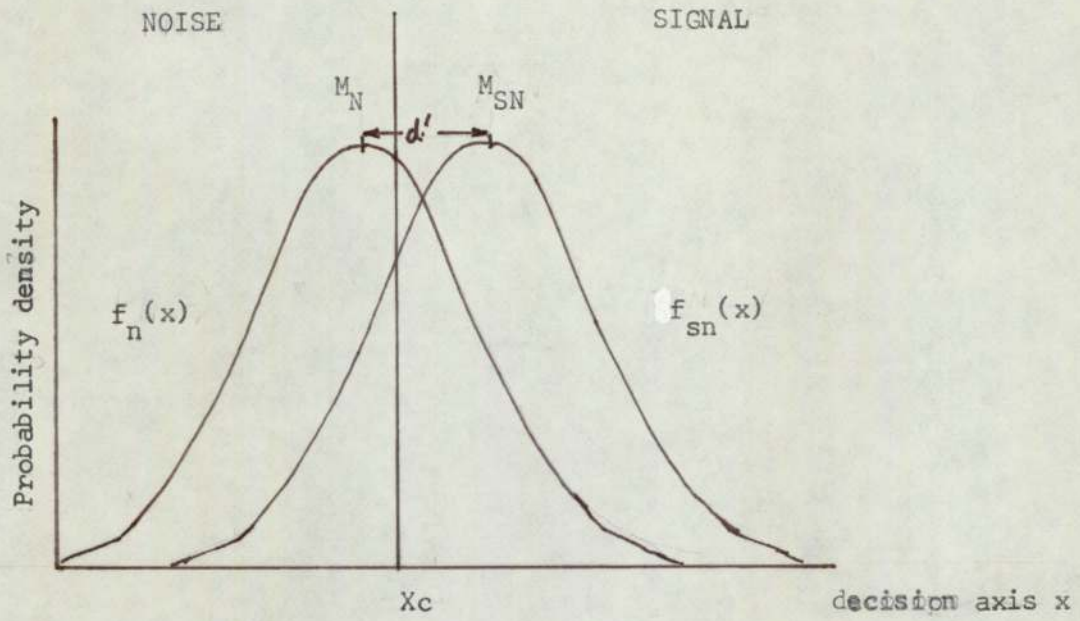


Figure 2.1 Equal variance probability density functions from log transform of likelihood ratio

2.1.1 Measurement of sensitivity and bias

An intuitively reasonable measure of sensitivity is immediately apparent from Figure 2.1 . This is d', the distance apart of the two distributions, scaled in terms of their common standard deviation σ.

$$\text{i.e. } d' = \frac{M_{SN} - M_N}{\sigma}$$

Clearly the further apart the distributions, the more discriminable are the signal and noise. If the two distributions overlapped completely, discrimination would be impossible, and if they were a considerable distance apart, the overlap, and hence the ambiguity involved, would be vanishingly small.

Having mapped the effects of the noise and signal on to a one dimensional axis, the question arises as to how the decision is made. SDT assumes that the likelihood ratio axis is a decision axis, and that the observer establishes a cutoff value of x, x_C, (Fig.2.1), such that if the transformed likelihood ratio exceeds x_C he responds signal, and if less than x_C he responds noise, i.e. non-signal. The position of x_C is defined by the ratio of the ordinates of $f_{SN}(x)$ to $f_N(x)$ at x_C, known as the criterion β (beta).

$$\text{i.e. } \beta = \frac{f_{SN}(x)}{f_N(x)} \Bigg|_{x = x_C}$$

The quantities d' and beta can readily be calculated from an experiment in which the probabilities of correct detections and false alarms can be estimated, using the equal variance assumptions. The results of such an experiment are shown below.

Observer's Decision

state of the world	NOISE N	SIGNAL S
NOISE n	correct 'no signal' response probability p(N/n)	false alarm probability p(S/n)
SIGNAL s	missed signal prob. p(N/s)	correct detection probability p(S/s)

Figure 2.2 Result of hypothetical detection experiment.

The probabilities are estimated from the relative frequencies of the various types of response. From Figure 2.1 it can be seen that a knowledge of P(S/n), the false alarm probability, given by the part of $f_N(x)$ above x_C, and the missed signal probability, given by the portion

of $f_{SN}(x)$ below x_c, together with a table of the areas under the normal curve, will enable d', the distance between the distribution means to be calculated. Similarly a table of the ordinates of the normal distribution will enable $f_N(x)$ and $f_{SN}(x)$ at x_c to be obtained and hence beta.

2.2.2 Some characteristics of beta

SDT assumes that the observer is able to place his criterion at any point on the decision axis x. If the criterion is placed far to the right this will lead to performance characteristic of the cautious observer. Only sensory data falling in the extreme right hand tail of the signal plus noise distribution will elicit a signal response. Performance will be characterized by a low false alarm probability but also by a low signal detection rate. Conversely a criterion placed far to the left (a 'lax' criterion) will produce a high detection rate accompanied by a high false alarm rate. A criterion near the centre of the decision axis will give an intermediate level of both false alarms and correct detections.

We can see that SDT takes a quite different approach to the classical psychophysical theories. Instead of a fixed threshold the criterion is infinitely variable and is independent of the sensitivity index d'.

2.2.3 The position of the criterion as a decision rule

We have not yet considered the factors which influence the position of the criterion. Some insights into this question were gained in the earlier discussion as to the nature of judgemental bias. SDT assumes that the position of the criterion is determined by two factors: the a

priori probability of a signal occurring relative to that of noise, and the costs and values associated with various decision alternatives, e.g. false alarms, correct detections.

The basic assumption is that the position of the criterion chosen by the observer represents the application of a decision rule designed to maximize the payoff of the series of decisions made by the observer during the signal detection tasks. It can be shown (Green and Swets (op.cit.) p.20) that the application of a likelihood ratio decision rule to maximize the expected value (in a decision theory sense) of the observers decisions will also maximize the payoff for a variety of other criteria. It can be shown (Coombs et al., 1970) that the expected value will be maximized if the decision rule is taken such that:

$$\beta_{OPT} \text{ (optimal criterion)} = \frac{P(n)}{P(s)} \cdot \frac{(V_nN+C_nS)}{(V_sS+CsN)}$$

where P(n) = a priori probability of noise

P(s) = " " " " signal

V_sS = value of making a correct detection

VnN = " " " " " rejection

CsN = cost of missing a signal

C_nS = " " making a false alarm

2.3 The ROC curve

The ROC curve, (Receiver Operating Characteristic from the electrical engineering origins of SDT) is a very useful way of representing performance using the SDT approach.

The ROC curve consists of a plot of the probability of correct detections against the probability of false alarms in a detection experiment. As implied earlier, these quantities are estimated from the frequencies of 'hits' and 'false alarms' observed. The pair of probability estimates obtained from an experiment in which the criterion remains fixed and the signal discriminability remains constant produces a single point on the ROC curve. In order to generate a complete curve, the subject has to be induced to alter his criterion to produce a series of points of differing x_c but constant sensitivity. This can be done either by varying the pay-offs for the various decision alternatives or by altering the a priori probabilities. Another method is to employ a rating scale technique, (Egan et al., 1959), whereby the subject has to make ratings as to his degree of confidence that a signal is present on a particular trial, e.g. definitely signal, possibly signal, possibly non-signal etc. This is equivalent to conducting several experiments simultaneously using different criteria, and is regarded as the most efficient way to generate an ROC curve, Green and Swets (op.cit.), McNicol (op.cit.). This technique will be employed extensively in the experimental part of this study. The ROC curves generated by these procedures are shown in Figure 2.3.

Each of the curves corresponds to a signal of different discriminability, and where the equal variance assumptions hold true the curves are symmetrical with respect to the negative diagonal. The constant d' curves are sometimes referred to as isosensitivity curves. A ROC can be regarded as being generated from right to left when the criterion is swept from left to right ('lax' to 'strict') across the decision axis. The positive diagonal represents 'chance' performance and an ROC curve can only be produced below the diagonal by 'malingering', i.e. by

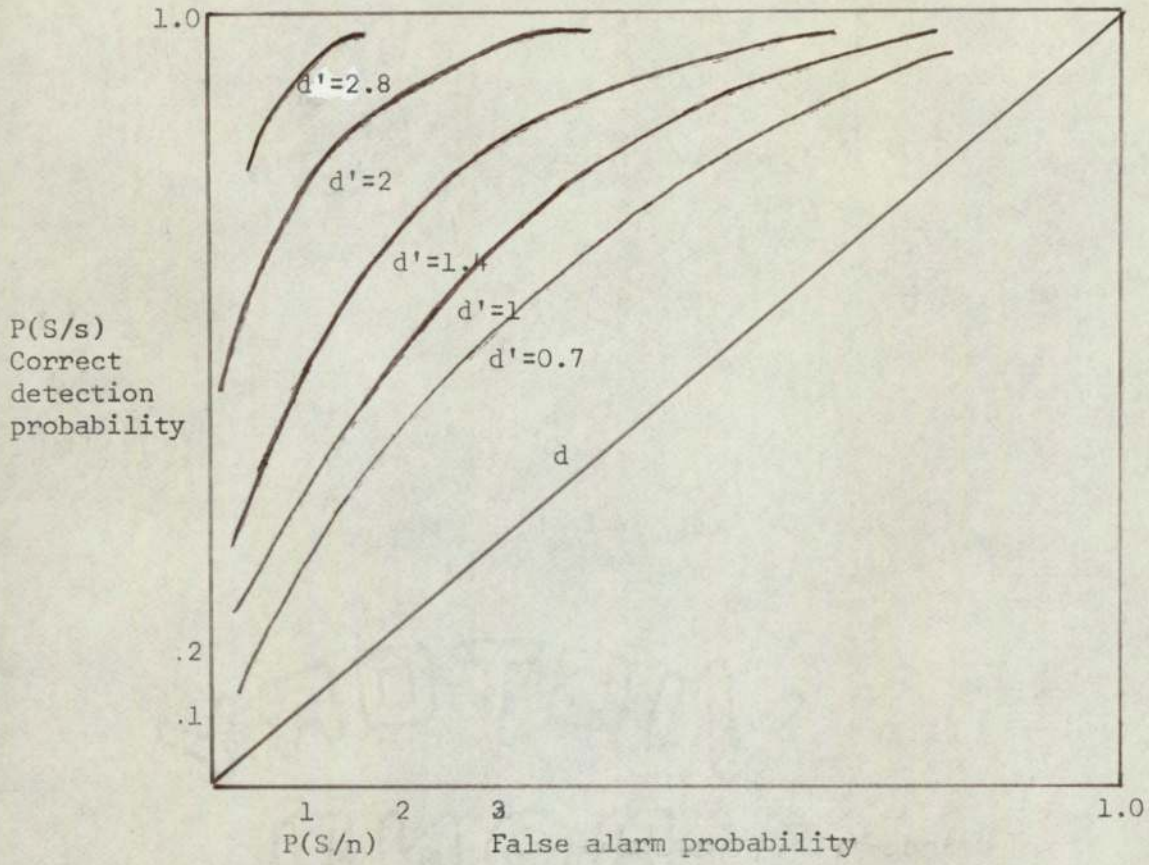


Figure 2.3 ROC curves generated with d' as the parameter

deliberately trying to perform badly. Considerable use is made of normalized ROC curves, sometimes referred to as Z-ROC curves. These are produced by plotting the Z-Scores corresponding to the probability data or by using double-probability paper with axes scaled in terms of the normal deviate. Both of these methods produce straight lines corresponding to the ROC curves of constant d'. In the case of equality of the variances of the underlying distributions these lines are parallel to the positive diagonal, i.e. have slope equal to 1.0. A further property is that the difference between the normalized co-ordinates of any point along the ROC curve is equal to the sensitivity d'.

2.4 Unequal variance model

2.4.1 Experimental consequences

Although many of the earlier experiments involving the detection of

auditory signals in white noise were satisfactorily fitted by the equal variance model described above, it soon became clear that in general an assumption of inequality in the underlying variances provided a better fit to the experimental facts. In many cases no checks were made on the data to find out the appropriate model in a particular experimental situation. A large proportion of the early experiments which employed SDT as a means of separating sensitivity from bias, simply used single estimates of correct detections and false alarms to generate d' and beta measures. These can be obtained without the effort of resorting to normal probability tables, by using tabulated values of d' and beta, e.g. Freeman (1973). As will be discussed subsequently, such a procedure can lead to large errors of estimation in the SDT parameters if the assumption of equal variance underlying distributions is incorrect.

Taylor (1967) has discussed why there is likely to be an asymmetry between the underlying distributions. A signal is normally thought of as something added to the non-signal. The subject normally knows quite well what the non-signal would be, were it not obscured by noise. He does not know so well, however, what the signal would be without the noise. There is an essential asymmetry between the signal event class and the non-signal event class, in that the subject usually knows less about what is a valid example of the signal class than the corresponding non-signal class. It is of interest to note that unequal variance distributions are much more frequently observed with visual signals, which often have complex attributes which cannot be specified exactly, than with auditory signals which can be specified precisely in terms of phase, duration and amplitude.

On purely practical grounds one would expect the variances of the noise and signal + noise distributions to be asymmetrical. In most detection

experiments there are usually far fewer false alarms than correct detections. The variance associated with estimating false alarm probability from a low false alarm frequency is intrinsically greater than with correct detections, where a greater sample size is available.

Whenever a subject knows less about a signal than a non-signal, the same effect will be produced. The ROC curve will cling to the left hand edge of the ROC space longer than it does to the top. The less the observer knows about the exact characteristics of the signal, the more skewed the curve will be.

2.4.2 ROC curve analysis of the unequal variance model

The unequal variance model assumes underlying noise and signal variances of σ_n^2 and σ_s^2. This means that an additional parameter has to be added to the basic SDT model in order to specify the shape of the ROC curve. This is the ratio of the respective standard deviations of the noise and signal plus noise distributions, i.e. $\sigma n/\sigma s$. It can be shown, Green and Swets (op.cit.) p.64 that the slope of the resulting Z-ROC is given by $\sigma n/\sigma s$. As $\sigma s/\sigma n$ increases the slope of the line decreases. A comparison of the Z-ROC's for the equal and unequal variance case is given overleaf. (Figure 2.4).

2.4.3 Measures of sensitivity

One important consequence of the unequal variance situation is that the measure of sensitivity d' is correlated with the criterion position chosen and that this correlation increases as the Z-ROC line becomes less parallel to the positive diagonal. If we consider line A, the

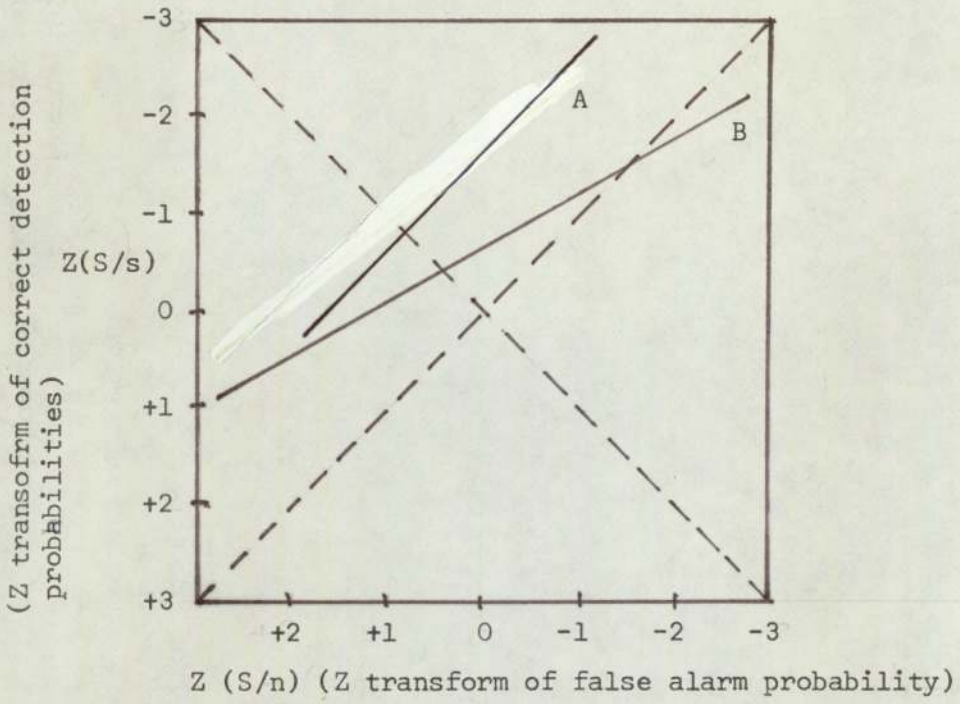


Figure 2.4 Normalized ROC curves representing equal (A) and unequal (B) underlying variances.

equal variance situation, it is obvious that d', (Figure 2.4) the difference between the corresponding Z-scores will remain constant. On the other hand different points on B, the unequal variance Z-ROC line, will indicate different values of the corresponding Z-Scores. It will be recalled that different points on the Z-ROC lines represent different degrees of bias and hence the non-independence of d' is clear. In this situation some decision has to be made about where on the ROC curve the sensitivity measure is to be read. Two sensitivity measures are commonly used. The first of these, Δ m, is the distance between the means of the signal and noise distributions measured in standard deviation units of the noise distribution. It is equal to Z(S/n) at the point on the ROC curve where Z(S/s) = 0.

Another measure of sensitivity is d'e, also called d_s, due to Egan and Clarke, (op.cit.). d'e is defined as twice the value of Z(S/s) or

Z(S/n), ignoring signs, at the point where the Z-ROC curve intersects the negative diagonal. One reason for using d'e is that Z(S/s) and Z(S/n) are equal where the Z-ROC line meets the negative diagonal and hence it gives equal weight to the signal and noise distributions. Since Δ m is scaled in units of the noise distribution, it is the appropriate measure if the noise variance is expected to remain constant over a series of experimental treatments, but the signal variance may change. If both variances are likely to change then d'e would be a more stable measure. Also if we expect signal variance to remain constant and noise variance to change, the most appropriate sensitivity measure would be the value of Z(S/s) at the point on the Z-ROC line where Z(S/n) = 0. This analogous measure to Δ m is scaled in units of the signal distribution and is employed in Thurstonian Category Scaling, (Lee, 1969).

Two final arguments have been advanced in favour of d'e. The first is that the point that it is read from the ROC curve generally falls within the range of responses made by observers. Hence extrapolation is not necessary. Secondly, Egan and Clarke (op.cit.) report that the changes in slope of the ROC curve observed from session to session within the same observer tend to alter the value of Δ m more than d'e, which thus appears to represent a more stable measure. In any event one measure can readily be converted into the other by means of the conversion formula provided by Green and Swets (op.cit.).

$$\text{i.e.} \quad d'e = 2 \Delta m \left(\frac{S}{1+S}\right)$$

where S is the slope of the Z-ROC curve

Theodor (1972) illustrates how incorrect assumptions can lead to erroneous conclusions. The table below represents data for a single subject under three conditions of an experiment with d' calculated

under the assumption that $\sigma s/\sigma n = 1$ and $\sigma s/\sigma n = 2$.

condition	P(correct detections)	P(false alarms)	Z (C.D)	Z (F.A)	d' $\frac{\sigma s}{\sigma_n} = 1$	d' $\frac{\sigma s}{\sigma_n} = 2$
A	.5000	.0668	0	-1.5	1.5	1.5
B	.6915	.3085	.5	-0.5	1.0	1.5
C	.8413	.6915	1.0	0.5	0.5	1.5

The assumption of equal variances ($\frac{\sigma s}{\sigma n} = 1$) leads to the interpretation that the points are on three different ROC curves of different sensitivity, whereas $\frac{\sigma s}{\sigma n} = 2$ gives the impression that all the points are on the same ROC curve. Unless $\frac{\sigma s}{\sigma n}$ is known, there is no way of telling which hypothesis is true.

2.4.4 Measures of bias

A number of problems arise when measurements of response bias are considered in the unequal variance case. Unfortunately in this situation there is no longer a simple monotonic relationship between the likelihood ratio scale and the underlying evidence variable. For example, in the case where the signal variance exceeds the noise variance, up to the first intersection of the distributions β will be greater than one, after the intersection it will be less than one, and after the second intersection it will be greater than one again. Therefore, there will in general be two values of the likelihood ratio which maximize the expected value of the observer's decision. In actual practice, the second cross-over point will generally occur in the extreme tails of the distributions and in general the observer behaves as if he places his cutoff at a particular value of beta. Even in the equal variance case

there are problems in comparing changes in beta between experimental conditions. If d' is not constant, then apparent changes in beta may be due to changes in d'. Baker (1975) points out that the likelihood ratio can vary between 0.0 and 10.0 when d' = 1.0 and can be as high as 100 when d' = 3.00.

One measure of bias which does overcome this difficulty has been proposed by Banks (1970). This is C, the distance along the likelihood axis from the noise distribution mean to the criterion scaled in Z-units of σ_n. The range of C is not a function of the separation of the distributions and it is always monotonic with the likelihood ratio axis. C for any point on the Z-ROC curve can be determined from the Z score on the false alarm axis. The other bias measurement often used is log β, this being a monotonic function of the evidence variable, i.e. the magnitude of the sensory evidence which the observer uses as input to the decision making mechanism.

Some of the problems of assessing response bias using beta are discussed in McNicol (1972) p. 119. One of the difficulties in using a single index of bias is that in a multiple criteria situation such as occurs in rating experiments, there is no indication if the observer has moved all his criteria up or down the x-axis, or has spaced them closer or wider apart, as found in Broadbent and Gregory (1963) for example. It is of course open to question whether the observer actually uses a likelihood ratio criterion in setting his cutoff. Although such a criterion achieves the broadest range of objectives, there is no guarantee that the observer will use the most rational criterion. It is not impossible that the response criterion be based purely on the evidence variable, the sensory effect produced by the stimulus. In this

case the observer will use a value of y as his criterion, and if the sensory evidence is less than y he will respond 'noise' and if greater than y, 'signal'. In the equal variance case this would produce no anomalies because the likelihood ratio is monotonic with the evidence variable. In the unequal variance case discrepancies would occur. Ingleby (1974) presents persuasive evidence that the decision maker actually does employ a likelihood ratio criterion, however, in auditory detection experiments in which the observer's criterion was systematically varied.

Dusoir (1975) critically reviews the attempts that have been made to measure observer bias using a wide variety of models in addition to SDT. He concludes that none of the existing indices of bias account for all the experimental evidence, and that it may be futile to search for such an index which remains invariant under all types of task, subject and experimental conditions. He also points out that the form of the iso-bias (ROC) curves needs to be established before experimentally manipulating such factors as the a priori signal probability or the payoffs in an attempt to modify the observers bias over experimental conditions. Finally he suggests that inferences about the degree of bias change in groups of subjects should not be used until it is clear that the iso-bias curves are all of the same family for different subjects. However, in spite of these criticisms he does not suggest a usable alternative bias measure, which somewhat weakens the validity of his criticism.

It must be emphasized that this study is not concerned primarily with the theoretical issues considered by Dusoir. As will be discussed in detail subsequently, the intention is rather to investigate the utility of detection theory ideas as a tool in a practical setting. The

intention is to adopt a more rigorous approach that has hitherto been used in real world applications of SDT in order to establish the validity and usefulness of the model in this context. From this standpoint we shall assume that the theoretical validity of the SDT model has been sufficiently well established by the body of evidence in existence. The data provided in this study will serve to provide further confirmatory or otherwise evidence for the theory in the context of real-life tasks.

2.5 Non-parametric indices of sensitivity and bias

As will have been gathered from the preceding sections, the question of the nature of the underlying distributions creates problems when using SDT to measure changes in sensitivity and bias. Fortunately there exist several 'non parametric' measures of sensitivity and bias that require fewer assumptions concerning the nature of the underlying distributions than the parametric indexes d' and beta.

The first of the sensitivity indices to be considered is P(A), the area under the ROC curve. As shown in Figure 2.3, as the index d' increases, the ROC curve becomes closer to the top left-hand corner of the unit square (within which the ROC curve is drawn). Green and Swets (op.cit.) p.45 show that the area under the curve is a measure of sensitivity independent of the shape of the underlying distributions. P(A) lies between 0.5 and 1.0. If an ROC curve is generated using the rating procedure mentioned earlier, a numerical integration technique such as the trapezoidal rule can be used to estimate the area under the curve. McNicol (op.cit.) p.114 gives an example of the technique. Pollack and Hsieh (1969) investigated the sampling distributions of both d'_e

(discussed earlier) and P(A). The expression they obtained for the standard deviation of P(A) is useful in statistical analysis using these indices. Where only a single point on the ROC curve is available another measure of sensitivity is available, A', due to Pollack, Norman and Galanter (1964) and Pollack and Norman (1964). This index is an approximation based on the measure P(A) discussed earlier. It is derived by considering the maximum and minimum values that P(A) can take. Grier (1971) provides a convenient computing formula for this measure and also for P(I), another sensitivity index suggested in Pollack and Hsieh (op. cit.) which is related to A'. Another index which utilizes rating scale data is AH, due to Hammerton and Altham (1971) and Altham (1973). This index has been criticized by Dusoir (op.cit.) because although it makes no assumptions about the underlying variances, it does assume that the observer employs a likelihood ratio criterion. **Navon (1975) has produced a sensitivity index derived from response latency measures.** It is clear that there is a considerable choice of indices available to measure sensitivity. The same cannot, however, be said about bias. Hodos (1970) developed a non-parametric measure of bias B, based on the fact that the negative diagonal of the unit square represents the locus of points where the subject would be equally likely to respond signal or noise given ambiguity. The measure reflects the degree to which a data point deviates from the negative diagonal relative to the maximum possible deviation. A computational formula for B is given by Grier (op.cit.). Hodos' measure is however criticized by Dusoir (op.cit.) as not being 'non-parametric' according to his definition and as being simply an arbitrary parameter that makes no specific reference to any sensitivity parameter.

Apart from B the only other measures of bias available within the SDT model are beta, log beta and C, as discussed earlier, although

McNicol (op.cit.) p.123 presents a procedure for deriving a non-parametric measure of bias from a rating experiment where there is insufficient data to obtain a beta value for each criterion. Only a single overall measure of bias is produced by this technique and it is, at best, a somewhat crude estimate.

2.6 Signal detection theory and inspection

It will be recalled that the basic ideas of the SDT model were developed using an example from the inspection area. The advantages of using the SDT approach in examining inspection tasks are numerous.

One of the most important applications of the SDT model is to provide an index of inspection performance. This is an important issue,because if a variety of different methods are employed to measure inspection performance,it is extremely difficult to compare inspection studies to assess the effects of differing factors on performance, as we shall see during the literature survey. The most common performance index employed in industry is,of course,the percentage of defects detected. McCornack (1961) discusses a number of other indices that have been employed and suggests that the suitability of a particular performance index depends on the objectives of the inspection system. For example different indices are appropriate if the object is to maximize correct detections regardless of false alarms, or whether false alarms should be minimized. Sosnowy (1967) also considers some of the performance indices available and shows that the ranking of inspectors in terms of efficiency can vary drastically depending on which of the performance measures is employed.

Signal Detection Theory, with its separation between sensitivity and bias, offers unique advantages as an inspection performance index. We can identify some of the requirements of an ideal index as below:

1. Should provide insights into why performance is good or bad in a particular case.
2. It should allow quantitative costs and values to be assigned to the various types of errors and correct decisions possible.
3. The index should separate the aspects of performance due to the inspector's sensitivity, from his response bias.
4. The performance index should enable inspection performance to be related to general theories of human performance.

It is clear that the signal detection indices of sensitivity and bias fulfill these requirements.

The likelihood ratio concept is particularly useful, in that it suggests an index of performance that has a theoretical optimum. We can thus compare the actual performance of the inspector with the optimum to see how it needs to be changed to produce a more efficient inspection system. The separation of performance into sensitivity and bias variables also provides insights into the way in which performance can be improved. Sensitivity, for example, can be improved by training the inspector to recognize the whole range of attributes that characterize good and bad products. The ability of observers to alter their response criterion on the basis of instructions suggests that inspectors can be induced to alter their response strategies towards the optimum. This area will be discussed in more detail subsequently.

When using SDT parameters as indices of inspection performance a certain amount of confusion is possible due to the differing definitions of what constitutes a 'signal' in detection experiments and in inspection situations. The two matrices in Figure 2.5 contrasting a detection and an inspection experiment, will make the distinction clear.

observer's decision

state of the world	NOISE N	SIGNAL S
NOISE n	correct 'no signal' response P(N/n)	false alarm probability P(S/n)
SIGNAL s	missed signal probability P(N/s)	correct detection probability P(S/s)

inspector's decision

state of product	perfect N	defect S
perfect n	correct 'product perfect' decision P(N/n)	false alarm or good product called bad P(S/n)
defective s	missed defect P(N/s)	defect correctly detected P(S/s)

Figure 2.5 Comparison of signal detection and inspection situations.

2.7 Relation between the SDT model and acceptance sampling

In industrial quality control, considerable use is made of various types of statistical sampling plans. It is useful to clarify the relationship between SDT concepts and those of statistical quality control (S.Q.C.).

SQC adopts the standard statistical usage of referring to false alarms and misses as Type I and Type II errors respectively. The missed signal probability $P(N/s)$ corresponds to failure to reject a false null hypothesis and is referred to as β in SQC (not to be confused with the entirely different usage of β in SDT). It is also known as the consumer's risk in acceptance sampling inspection, i.e. the risk of accepting a bad

lot. Similarly the probability of false alarms, P(S/n), is analogous to the quantity α in SQC, known as the producer's risk, the risk of rejecting a good lot.

The SDT criterion x_c is analogous to C in sampling inspection where C is the number of sampled defective items which must be exceeded in order to reject the entire lot as defective.

Thus the SDT conceptualization of the inspector as an inferential decision maker is very similar to the theories of inferential acceptance sampling as practised by quality control engineers and statisticians. This adds further weight to its adoption as a conceptual paradigm in inspection.

2.8 Applications of SDT in inspection studies

Considering the very obvious advantages of the SDT approach, it is surprising how infrequently it has actually been applied in the inspection area. This is probably a reflection of the fact that inspection in general has been an under-researched area as far as human factors is concerned. Even where SDT has been applied to inspection this has often been in the context of laboratory studies rather than real life tasks. The methodological issues of applying SDT to inspection situations are dealt with explicitly in most of the papers to be considered in this section, mainly because SDT is still relatively unfamiliar to ergonomists with an applied orientation. A paper by Baker (op.cit.) comprehensively reviews the whole area of SDT applicable to inspection tasks, without being orientated towards a specific study. Other review papers are Drury (1975), which also considers Information Theory and the application of Bayes' theorem to inspection, and Adams (1975). The

growing interest in the application of SDT to inspection was very evident at the 1974 symposium on human reliability in quality control held in Buffalo, U.S.A., where virtually every paper contained some reference to the theory.

If we exclude studies in such areas as sonar detection, where SDT was applied as early as 1967 (Colquhoun, 1967), the first published paper employing SDT in an inspection context was Wallack and Adams (1969), although it reported earlier unpublished work by Wallack (1967). The Wallack and Adams study will be considered in some detail, because its methodological shortcomings serve to exemplify the problems which ensue when SDT is used without due regard for the underlying assumptions of the model. In this study, inspectors were required to examine samples of 260 electrical cables for examples where the conductors had been nicked or abraded in a wire stripping operation. Four sample lots were inspected containing 5, 15, 25 and 35 percent defectives. The inspectors were trained with a scheme which presented them with samples containing 80, 60, 15 and 35 percent of defects respectively, the higher incidence of defects sample being used as a teaching aid in which each wire inspected was discussed with the trainee to provide complete feedback. A payoff matrix was then assigned such that correct acceptances and rejections had a value of 1 unit and missed defects and false alarms were associated with a cost of 3 units. These values were purely abstract - the inspector did not receive any concrete rewards or payoffs for his performance. The inspectors were given further practice until they were familiar with the payoffs, the training terminating when they were able to achieve a particular level of payoff. The experiment proper was conducted such that the different incidence of defects samples were inspected at random times during the week. The

SDT parameters d' and beta were calculated from the incidence of correct detections and false alarms, as outlined earlier. Also calculated was the distance of the criterion from the noise distribution. The results were interpreted by the authors as follows. The obtained mean beta values for each of the different probability of defects samples were not in general equal to the theoretically optimal values predicted by SDT, and the discrepancy was greatest with the lowest defect probability. The obtained betas did not even give the same rank order as the optimal ones. Two reasons were suggested for this. One was that the inspectors did not employ the payoffs assigned, and the other that they were not a homogeneous group, and could be divided into two sub-groups with differing degrees of bias.

The major general criticism that can be levelled at this study, was that no attempt was made to examine ROC curves for the experiment, to check if the data did conform to the SDT model. Without such evidence, any conclusions drawn from the SDT parameters must remain highly questionable. The failure to draw ROC curves is particularly strange, in that data to do this are published in the paper in the form of correct detections and false alarms for each incidence of defects for each inspector. According to the SDT paradigm, the differing a priori probability of defects should induce a series of different criteria in the inspector and hence generate an ROC curve. The ROC curves, when plotted, show that although two of the seven inspectors appear to produce a straight line Z-ROC, the slope does not support the equal variance assumption and hence d' and an uncorrected value of beta do not provide meaningful measures of bias and sensitivity. The fact that some of the results are not fitted by the SDT model does not mean that it is necessarily inappropriate. The fact that the inspectors did not receive any training with some of the defect levels employed could

account for the results for example. In view of these considerations, attempts to compare obtained values of beta with the theoretical optima are obviously misplaced. Another methodological point concerns the instances in the data where the false alarm probability is zero or the hit probability is unity. In these situations, the corresponding z-scores tend to plus or minus infinity and hence neither d' nor beta can be calculated. In spite of this, values of d' do appear in the results table at these places. This is because the authors have used one of the approximations that can be employed in these situations and which will be discussed in detail later. If such approximations are used it is essential that they be made explicit. In the data under consideration they produce inflated values of d' that are incorporated in the group means. It is obvious that SDT cannot be applied in such a casual manner if meaningful results are to be obtained. In SDT terms, the only concrete facts to emerge from this experiment are that some inspectors' performance can be described by the unequal variance model, and that inspectors do not appear to change their criteria according to the a priori probabilities.

The next published paper to use SDT in an inspection context was Embrey (1970). This concerned the inspection of bubble chamber photographs, produced in high energy physics investigations, for the occurrence of particular configurations of tracks. SDT was employed to ascertain whether differences in detection efficiency were due to the differing discriminability of the two configurations employed in the experiment or to differences in response bias on the part of the inspectors. The effects of differing levels of ambient noise, and time on the task were also considered. Although the results were of interest, the study was again a naive application of SDT in that ROC curves could not be plotted and the basic assumptions remained unverified. Wallack and Adams (1970)

reanalyse the data of their earlier paper using the measures of McCornack (1961), discussed earlier. A Bayesian measure of performance is also calculated, and a measure due to Freeman et al. (1948), which is related to statistical quality control considerations. It is pointed out that SDT measures of inspection efficiency are the only ones available that consider the effects of the costs of the various decision alternatives.

Lusted (1971) used SDT to analyse the performance of radiologists evaluating X-ray photographs. He found that the lack of agreement between radiologists on the diagnosis of the photographs could be explained in terms of differing criteria rather than differing sensitivities. ROC curves plotted for this task indicated that the unequal variance SDT model gave an excellent fit to the data. Lusted used the sensitivity parameter d'e to compare the effects of alternative presentation methods. ROC curves were also used to show that paramedical personnel had a lower sensitivity than radiologists in this task. Interestingly enough, the ROC plot for this experiment also indicates that the less experienced paramedical group showed a greater σ_s/σ_n ratio than the radiologists, presumably because they were less familiar with the characteristics of the signal. The data from these studies are not presented in detail but the use of d'e, an appropriate sensitivity model for the unequal variance case, gives one far more confidence in the author's conclusions. This study is one of the very few practical applications of SDT in which such a sophisticated approach has been adopted. Another study in the radiology area, Sheft et al. (1970), used the measure d'e to show that the detection performance of X-ray technicians who had received five months training was indistinguishable from that of senior consultants.

The next industrial application of SDT to appear was Sheehan and Drury (1971), the detailed results of which were published earlier, Drury and Sheehan (1969), but not analysed in SDT terms. Inspectors with a wide spread of age and visual acuity, but judged to be of equal competence by the company, inspected 6 batches of steel hooks containing 20 percent of items with a single defect and 5 batches each containing a similar proportion of defective items, but each item contained two defects. From the point of view of SDT, one of the main results of interest was that when the usual probabilities were plotted on an ROC curve, most of the inspectors results seemed to be moderately well fitted. However, one inspector's data seemed completely discrepant. Subsequently it was found out that she was rejecting a large number of acceptable hooks because of a surface blemish which should actually have been ignored. The authors suggested that a particular advantage of the SDT approach was that it enabled such an error to be more readily detected than using other indices of detection performance. The ROC curve also indicated that all of the inspectors were employing a stringent criterion, leading to a low level of false alarms but a relatively high incidence of missed defects. In this particular inspection situation, this was an inappropriate strategy, since missed defects were 'expensive'. Using tables from Swets (1964), (which assume equal variance underlying distributions), values of d' were calculated and the results used to show a negative correlation between age and sensitivity. d' was also used in a second experiment to show that prior knowledge as to which type of defect was going to occur, enhanced sensitivity. This study was considerably more satisfactory than the first one considered, in that an ROC curve was drawn, and no attempt was made to calculate beta with inadequate information. However, there were still several methodologically suspect aspects. Rather than attempting the difficult task of fitting ROC

curves to the data points by eye it is far more sensible to convert them to z-scores or use double probability paper to produce a straight line z-ROC plot, thus facilitating an objective test of the fit of the data to the equal variance assumptions. The usual criticism of the use of d' without testing the assumptions applies. One of the nonparametric measures would have been more appropriate. The paper discusses the ways in which the inspector might be induced to vary his criterion by rapid feedback of the results of his inspection.

The next paper, Drury and Addison (1973) represents the best and most extensive available application of SDT to inspection tasks. There are a number of reasons for this. The study used data from an on-going industrial task rather than a simulation or a laboratory study. Techniques were presented for obtaining estimates of SDT parameters from the data normally available in inspection situations. The data was tested carefully for its conformity with the SDT assumptions, and the theory was used to gain new insights into the way in which the inspectors performed their task. In the experiment, inspectors in a glassworks, 100% examined certain unnamed expensive glass products (actually colour television tubes). There was a sample inspection by special examiners of the items classified as good and faulty by the 100% inspectors. During the period of measurement the special examiners were moved to a point closely following the 100% inspection and their results made available to the inspectors much more rapidly than before. Twelve consecutive weeks performance before and including the change were measured and eleven weeks after the change. Each week's data consisted of the percentage of items classified as good by the inspectors together with the special examiners reports of the percentage of the rejected items that were faulty, and the percentage of the rejected items

that were good. The latter two quantities are subject to variable degrees of sampling error, due to the variable size of the sample and the fault percentages involved. It also seems likely that the standards of the sample examiners were probably subject to the same limitations as the 100% inspectors, e.g. pressures due to the varying nature of customer's requirements.

The authors point out that although, a priori, the data seem unlikely to conform to the SDT model, the situation does have some resemblence to the classical SDT experiment where some uncontrolled variation still exists. They emphasize that they are attempting to analyse group performance in SDT terms, and that the situation is analogous to the usual SDT experiment.

By using a decision tree approach some extremely useful expressions are derived for obtaining estimates of the quantities P(S/s), P(S/n) and the a priori probability of a defect occurring, from the percentage of items classified as good by the inspectors and the special examiners estimates of the percentage of Type II and Type I errors. These expressions are of considerable general utility, in that they are often available in real life inspection situations, where the actual a priori probability of defects is not usually known of course. The expressions are given below:

P_1 = probability that a good item will be accepted

$$= \frac{(1 - y)\, P_A}{x(1 - P_A) + (1 - y)\, P_A}$$

P_2 = probability that a faulty item will be rejected

$$= \frac{(1 - x)\,(1 - P_A)}{yP_A + (1 - x)\,(1 - P_A)}$$

P_G = a priori probability of a defect = $x (1 - P_A) + (1 - y) P_A$

where P_A = proportion of items inspected which are accepted as good.

x = proportion of items rejected by 100 % inspector which are in fact good.

y = proportion of items accepted by 100% inspector which are in fact faulty.

x and y are both obtained from the special examiners' sample inspection.

The results were first fitted by Z-ROC curves and it was found that the pre and post feedback data (subsequently referred to as 'before' and 'after' data) could be fitted by two straight lines of slope not significantly different from one. The good fit of the data to the equal variance SDT model greatly facilitated subsequent analysis. The fitting of straight lines to this type of data using least squares techniques cannot normally be recommended since there are errors in both variables. Least squares techniques assume that one of the variables is independent. Maximum likelihood fitting techniques have been advocated in this situation by Dorfmann and Alf (1968, 1969), Ogilvie and Creelman (1968) and Grey and Morgan (1972). In view of the very high correlation coefficients obtained for the two lines (0.803 and 0.802), perhaps this criticism is unnecessarily purist, but it could become important in situations where the scatter of the points were greater.

It was shown that the detectability of the defects, as measured by d' increased significantly after the introduction of feedback. The reasons for this are not entirely clear unless the examiners gave the inspectors insights into the nature of the critical defects. A plot of fault density against probability of detecting a defect showed the decreasing relationship that has often been observed in inspection studies, e.g. Fox and Haslegrave (1969).

As pointed out in section 2.2.3, SDT suggests that if an inspector uses a likelihood ratio criterion to maximize the expected value of his decisions, then (using the terminology of this paper):

$$\beta \text{ optimal} = \frac{P_G}{1 - P_G} \times \text{(relative cost factor)}$$

Therefore a plot of beta against $P_G/(1 - P_G)$ will be a straight line through the origin. Drury and Addison's data confirm this prediction only for the data from the feedback conditions. In other words, the inspectors were only able to adjust their criteria optimally to the incoming glass quality when feedback was provided. This result might be expected from the assumption that the inspector requires some means of adjusting his subjective estimate of the defect incidence in order to adjust his criterion. This is provided more effectively by the direct feedback from the special examiners than by the intrinsic feedback present in the task.

The paper also suggests that the inspectors change their criteria to keep their outgoing type II error (y) constant, since this is one of the important factors on which their performance is judged. The data presented seem to lend support to this hypothesis.

This paper shows how SDT measures can be used to gain insight into a wide range of inspector behaviour. There are certain criticisms that can be made. One is the use of least-squares techniques in fitting ROC curves, as already mentioned. Another concerns the use of the aggregated measures of the number of inspectors. Although the overall performance of the inspection system is well described by this approach, it masks the considerable inter subject variation that must be present. In order to generate an ROC curve at all, the inspections must be utilizing a very wide of criteria. In view of the fact that there was

no systematic large change in the aspects of the task likely to affect the criterion, the wide variation in this parameter is surprising. Also the use of grouped data conceals the idiosyncrasies of individual performance which would need to be investigated in order to improve the overall efficiency of the inspection system. For example it is possible that a small proportion of inspectors with low sensitivity and inappropriate criteria may be adversely affecting the system. In spite of these criticisms, this study illustrates the insights that can be gained into an inspection system by the use of SDT.

Other studies employing SDT have paid less attention to the underlying assumptions of the theory. Smith (1975) used d' in a study on the optimal magnification levels for microminiature inspection. Buck (1975) discusses the applications of SDT in the dynamic visual inspection area. He concludes that the theory is only likely to be useful in situations where the relationships between task parameters and their effects on SDT variables can be established. A study by Zunzanyika and Drury (1975) used the rating scale technique, where the inspectors were required to place inspected items in boxes labelled 'definitely accept', 'probably accept', 'probably reject' and 'definitely reject'. The experiment was designed to investigate the effects of various types of information on inspection performance. The three conditions considered were feedforward, where the inspectors were given prior knowledge of the type of defect likely to occur, feedback, where knowledge of results were provided, and a combination of these two information sources. Batches used in the study consisted of 10, 20 or 30% defects. The control group inspected the same batches as the experimental group but without the information conditions. For some reason, the rating part of the experiment was not used to generate an ROC curve, but was instead used to

provide three estimates of d' from the boundaries of the different categories employed. The hypothesis was that if SDT applied then d' would remain constant for the different criteria generated by the rating procedure. The results suggested that SDT did apply to this study, since there were no significant differences in d' across criteria. The differences between the various information conditions was no greater than that due purely to learning in the control group. Surprisingly, the authors state that criterion information is lost with studies of this type. In fact it is simple to calculate the position of the series of criteria that result from rating scale experiments. The applicability of SDT was further tested in a similar manner to the Drury and Addison study, by plotting the beta values obtained from the different a priori defect levels, against the ratio of the probability of a defect occurring to that of the probability of perfect product. The experimental group was fitted by such a straight line, but not the control group. In spite of this partial confirmation that SDT was applicable, both the sensitivity and criterion were affected by the changes in the a priori probability of a defect, an effect that was statistically significant. It is not easy to draw any definite conclusions from this study regarding the applicability of SDT, partly because of the equivocal nature of the results and partly because no ROC curve was drawn and hence we cannot make any inferences about the nature of the underlying distributions.

Most of the other applications of SDT have used the theory merely to provide the quantities d' and beta, using the equal variance assumptions and without any particular theoretical discussion of the applicability of the theory. Examples of such approaches have been Moraal (1975), in the inspection of steel sheets and Chapman and Sinclair (1975),

concerning the inspection of food products.

2.8.1 General discussion of the literature

At first sight SDT appears to be ideally suited for application in the area of industrial inspection. The separation of sensitivity and response bias, the existence of normative standards for the criterion, and the possibility of incorporating the costs and values of the inspector's decision making into the model are powerful arguments for its use. The analysis of the preceding papers has suggested that up to now, however, SDT has been used in a somewhat naive manner in the inspection area. As has been suggested throughout this section, unless the basic assumptions of the theory are shown to be applicable to the data under consideration, then any conclusions drawn from the use of d' and beta must be viewed with considerable reservations. Some workers in the applied area, primarily Drury and Lusted,have taken the precaution of checking the SDT assumptions before using the theory. Also there has been some recognition by these workers of the difficulty of applying the parametric measure of bias, beta, in situations where the SDT equal variance assumptions do not hold. Even where the ROC curve has been drawn, there has been a tendency to assume that an approximate fit of the data to freehand curves is sufficient evidence that SDT can be applied. A far more accurate procedure is to transform the probabilities to Z-Scores to give a Z-ROC plot which can be readily fitted by a straight line either by eye or by the maximum likelihood techniques outlined in Grey and Morgan (1972). Such a procedure allows the slope of the line to be evaluated such that decisions can be made as to which form of the SDT model is appropriate.

One aspect of industrial experiments that is of interest concerns the way that the ROC curve is generated. In all industrial experiments to date, no explicit attempt has been made to generate the curves by the procedures used in laboratory experiments on SDT, i.e. by manipulating the a priori probability of signals or by varying the payoffs. The general practice has been to simply plot the probabilities of correct detections and false alarms (or their Z-Scores) and by good fortune there has been sufficient random variability in the response bias to generate an ROC curve. It is certainly of interest to see the wide variability of criteria that exist in data taken from industrial situations. From the Z-ROC curves in Drury and Addison (1973) it can be calculated that the criteria vary between 1.02 to 5.03 in terms of beta. These data were, of course, produced by a wide variety of personnel working in conditions where there might well be frequent changes in criterion due to alterations in customer standards. In other industrial situations, it is possible to foresee problems if the inspectors' criteria were to be very homogeneous and stable. This might lead to an inadequate spread of points to establish the ROC curve. In this situation it might be necessary to perform supplementary experiments using a rating scale approach, in order to provide the range of criteria necessary. When one is attempting to apply SDT in industrial situations it is essential that the subjects are thoroughly experienced in the task. Otherwise they cannot be expected to have acquired the necessary knowledge of the conditional probability distributions to make decisions in accord with the SDT model.

There is so little data available that it is difficult to say whether all the predictions of the model have been verified in the area of quality control. Certainly, the Drury and Sheehan, Drury and Addison and Lusted data, lend support to the basic precepts of the model

concerning the underlying normality of the sensory distributions. In all cases the ROC curve provided an acceptable fit, although only Drury and Addison provided a test of statistical significance for this, and their fit was obtained using an inappropriate least squares procedure. As far as the other aspects of the model are concerned, which predict that the inspector will attempt to maximize the expected value of his decisions by modifying his criterion in terms of the equation given in section 2.2.3 i.e. beta = (prob. good/prob. defect) x (relative/cost factor), the evidence is more equivocal.

Drury and Addison plotted beta against (probability of good product/ probability of bad). In terms of the above equation, this should produce a straight line through the origin. This was found to be the case for the sessions where feedback had been provided but not before. This can be accounted for by assuming that one of the results of feedback was to give the inspectors a more accurate subjective estimate of the proportion of defects present, thus facilitating the optimization of their criteria. An alternative explanation is that the effect of feedback was to stabilize the payoff matrix by the special examiners providing fixed standards for the relative costs of the different types of error. The authors show that yet another possibility is that the inspectors modified their criteria to maintain a constant output proportion of defectives, this being one of the factors on which their performance is judged and which affects their relationships with the customers. In fact, the data suggest that this hypothesis is also reasonable. In terms of the SDT model, the inspectors could be regarded as altering their payoff matrix to keep the output proportion of defects constant, via a modification in the criterion.

It seems then, that on the modest amount of evidence available, SDT does provide useful insights into inspection performance. However, the situation is complex, and needs to be investigated further, particularly from the standpoint of the effects on the criterion of various aspects of the task.

Very little work has been done on the inspector's ability to inspect to a specific set of payoffs.

In summary the following requirements are important in order to meaningfully apply SDT in inspection situations:

1. The inspectors should be well practised.
2. The data should be subjected to an ROC curve analysis to ascertain if the underlying assumptions are correct.
3. If the spread of the criteria are insufficient to define an ROC curve, then some form of off-line experiment may be necessary employing the rating technique or some other means of generating a range of criteria.
4. Straight line Z-ROC plots are to be preferred to attempting to fit complex ROC curves by hand to the data. An accurate Z-ROC enables the ratio of the variances of the underlying distributions to be established. If data of the right form is available it is best to fit the data points using a maximum likelihood ratio technique.
5. Data for groups of subjects should not be combined unless it has been established that they all conform to the SDT model.

2.9 Directions for research into inspection using Signal Detection Theory

Two main areas of research work can be identified in which further SDT

orientated studies would be of interest. The first of these is concerned with the application of the theory to on-going inspection situations. As has been repeatedly emphasized in the last section, the amount of data available from real life studies which has been examined in a rigorous manner, using SDT techniques, is very small. Further studies are necessary to establish the range of application of the theory. Additionally, it would be useful to investigate the usefulness of the various non-parametric measures of sensitivity and bias in applied situations. If it could be established that at least some of these measures were relatively insensitive to variations in parameters such as the slope of the ROC curve, then they might be more readily used in real-life inspection situations than the corresponding parametric measures.

The other main area of research work concerns the use of SDT to investigate a number of more specific aspects of the effects of various taks parameters on inspectors' performance. Perhaps the most interesting of these, and one which is particularly suited to an SDT approach, is the question of the effects of changes in fault density in the incoming items to the inspector. We have seen from the Drury and Addison study that it is not at all clear whether the inspector actually employs the optimal criterion suggested by SDT or whether he bases his criterion on other considerations such as the outgoing percentage of defectives. Further studies are needed to establish in greater detail how the inspector responds to the a priori probability of a defect. Another aspect of this question concerns the inspectors response to change in defect probability within the task itself. Studies in both the signal detection area and inspection have tended to concentrate on static tasks. If one considers an inspector examining products in a continuous flow system, for example, it is possible that one of the

manufacturing units may develop a fault which suddenly increases the incidence of defects. An analogous situation would be if an inspector was transferred from a situation where a low incidence of defects were the norm to an inspection line where a more inherently faulty product were being inspected. Equally the fault density might change from high to low in the latter situation. It would be of interest to consider the inspectors reaction to change in these cases. Would they, for example, move their criteria in the optimal direction predicted by SDT?

In terms of the SDT model, where the defect probability increases, the inspector should lower his criterion to one appropriate to the new defect probability and make a greater number of 'defect' responses. But such an adjustment presupposes that the inspector has a perfect knowledge of the new defect probability and that he is able to act on this knowledge by adjusting his criterion. It is clear that the inspector will only have a limited sample on which to base his estimate of the new a priori defect probability. An accurate estimate would only be available if the inspector were able to discriminate perfectly between defects and non-defects, or if complete knowledge of results were available. In the absence of this, his subjective estimate of the degree of change in fault density will be a function of his intrinsic sensitivity for defects. We know that human beings are very conservative interpreters of evidence as far as the revision of subjective probabilities is concerned (Edwards, 1962, Slovic et al., 1975). Bayes' theorem indicates the optimal degree to which subjective probability estimates should be revised on the basis of evidence, but the experimental evidence suggest that subjects do not generally do this to the optimal degree. These two factors, the limited evidence available due to the finite sensitivity of the operator, and his innate

conservativism, mean that his subjective estimate of the defect probability will lag behind the actual probability. The ability of the inspector to adjust his response bias to the new probability can be regarded as a separate attribute to his skill at estimating this probability. This question has been investigated by Sims (1972) in a laboratory simulation of on-inspection task, with equivocal results. Signal detection theory, with its separation of sensitivity and bias parameters, and its specification of an optimal criterion, is clearly an ideal model with which to investigate this problem. The question will be considered again when the detailed programme of experimentation is set out.

2.10 Summary

The theoretical foundations of SDT have been reviewed in detail. The necessary conditions for applying the theory have been set out, and the available inspection studies in which SDT has been used have been considered from the point of view of their adherence to these conditions. Very few studies were seen to have tested the assumptions underlying SDT prior to employing the sensitivity and bias parameters beta and d'. Those studies that have employed SDT more rigorously have suggested that the theory is applicable in this area, particularly in the context of on-going inspection tasks. Further research is seen as necessary in establishing this validity in a wider variety of inspection situations, and also as a tool in investigating a number of important practical problems. In particular, the ability of the inspector to adjust his criterion appropriately, if the incidence of defects changes, is seen as being highly amenable to a SDT approach.

CHAPTER 3 A REVIEW OF THE LITERATURE OF INDUSTRIAL INSPECTION AND RELATED THEORETICAL AREAS

3.0 INTRODUCTION

Producing a comprehensive classification scheme for inspection literature presents a number of difficulties. Inspection is an activity carried out in a very wide range of industries and one which utilizes many differing skills. Some of the taxonomic difficulties present in this area will be discussed subsequently. The literature review which follows will be divided into two broad sections. In order to establish the context for the research objectives of this study, the more important areas in the inspection literature will first be surveyed. Although the emphasis of this review will be on studies which can be directly applied to practical situations, it will also be necessary to consider some of the theoretical areas which underlie the applied studies.

In the second part of the review, the topics which have been selected as part of the experimental investigations will be treated in greater depth. From a detailed consideration of these areas, together with insights gained from the SDT literature reviewed in Chapter 2, the broad outlines of the experimental investigations will emerge.

3.0.1 Some taxonomic considerations

It is not proposed in this study to develop a formal task taxonomy for inspection tasks, although there is certainly a need for such an endeavour. In applied research particularly, as the body of data on human performance grows, it is increasingly necessary to be able to generalize research findings from laboratory studies to operational settings and from one operational setting to another. (Levine et al., 1971). An extensive research programme concerning the problems of behaviour taxonomies has been in progress at the American Institute for

Research, e.g. Fleishman et al. (1970), Miller (1971). Several different lines of approach have been considered in this area. One of the earlier attempts by Miller (1962) was motivated by the desire to obtain data for design decisions in man-machine systems. The main characteristics of Miller's scheme resemble those of the informal model for the inspection process presented in Chapter 1. The behavioural task structure suggested by Miller proposes scan, identification, decision making and effector phases which can be readily equated with the acquisition, decision, identification and action phases proposed in Chapter 1.

Unfortunately, both schemes share another characteristic - they do not provide a very satisfactory means of organizing the available research and applied literature in a particular area. One of the difficulties is that the model is in terms of separate psychological processes, whereas in any real situation the importance of the various hypothesized stages in the inspection procedure is determined very largely by the characteristics of the task. This problem is of course common to any taxonomy. Another problem that occurs when attempting to classify inspection studies according to the scheme set out in Chapter 1, is that many of the important variables are global in nature and could affect performance through a number of the stages postulated. Examples of such variables are individual differences and social factors. It seems then, that although the earlier descriptive scheme is useful in specifying the sequential stages in the inspection task, and the variables which need to be considered at each stage, a different approach is necessary from the standpoint of structuring the literature.

The scheme eventually decided upon follows the task characteristic approach proposed by Farina and Wheaton (1971). Four major sets of variables are seen to determine inspection performance: the

characteristics of the inspection tasks themselves, the physical environment in which the tasks are performed, the organizational and social structure of which the inspection function is a part, and individual, operator centred variables. Of course, although these categories and the variables they comprise are described as independent factors, performance in a real inspection task will be a complicated interaction of these variables. In general, inspection studies have tended to concentrate on single or at the most two interacting variables, although there have been some exceptions, e.g. McFarling (1974), who examined the interactive effects of noise, sex and pacing variables.

Many theoretical areas impinge on the inspection situation, but the emphasis in the first part of the review will be on studies that are either applied, or attempt to simulate at least one aspect of real life inspection tasks. However there will be a preliminary discussion of the major research areas that have implications for a large number of inspection problems. The inspection model described in Chapter 1 will be utilized during the review where appropriate.

3.1 General inspection literature survey

3.1.1 Some relevant theoretical areas

Three theoretical areas of psychology have considerable relevance for inspection tasks. These are decision theory, vigilance and visual search. Decision theory has already been discussed in detail from the orientation of signal detection theory in Chapter 2. Vigilance is an important area because many inspection tasks are largely perceptual in nature and involve prolonged periods of attention with a low probability of signal occurrence. Visual search considerations provide

insights into many of the task characteristics which will be considered in this review. For this reason we will begin with a brief consideration of the latter two theoretical areas.

3.1.1.1 Vigilance and its relevance to inspection

Vigilance has been an important area of psychological research since Mackworth (1950) was able to demonstrate in the laboratory some of the performance decrements occurring during prolonged watchkeeping which had first been observed in radar operators during the war. The apparent relevance of vigilance research to many applied problems gave rise to a voluminous literature. A review by Halcomb and Blackwell (1969) referenced 700 articles. No attempt will be made here to review this literature. Several comprehensive reviews are available, e.g. Davies and Tune (1970), Mackworth (1969 and 1970), Broadbent (1971). Mackworth identified a number of factors which contributed to the performance decline over time, such as a low signal rate, adverse environmental conditions, and unfamiliarity with the work. Later researchers added sleep deprivation, (Wilkinson, 1960) inappropriate signal expectancies (Colquhoun and Baddeley, 1964, 1967) and poor motivation of experimental subjects (Mackworth, 1970).

The similarity between vigilance tasks and inspection tasks is clear. Both involve prolonged attention by the subject for signals (or defects) which may occur infrequently, randomly in time and space and be difficult to readily discriminate.

Many of the findings in vigilance experiments parallel those found in inspection. For example, one of the most consistent results found

in vigilance is the importance of signal rate in determining efficiency. Experiments by Colquhoun and Baddeley (op.cit.) demonstrated the importance of signal rate in influencing the overall level of performance in vigilance tasks. Increases in signal probability produce increases in both the detection rate and the false alarm rate, (Baddeley and Colqhoun, 1969), a finding which could have been predicted from SDT considerations (Chapter 2). Closely analogous results are found in the inspection literature (see later section) and there does seem to be a close affinity between the classical vigilance task and many inspection situations. Many writers, e.g. Poulton (1973), make the assumption that most vigilance data can be readily applied to inspection tasks as long as the subjects have received sufficient practice.

However, other workers have had reservations about the applicability of much vigilance research to real life industrial problems. Kibler (1965) made the following comments when comparing the basic task dynamics of typical vigilance research with those of contemporary monitoring tasks.

1. The weak, brief duration signals typically employed in laboratory vigilance studies are rarely encountered in applied monitoring tasks.

2. The human monitor is typically required to keep watch over multiple information sources, and frequently more than one type of target or information class is the object of his vigil.

3. The signals are often complex and multidimensional rather than the simple, unidimensional events usually employed in laboratory studies.

4. In most monitoring tasks, determining the appropriate response to a

signal event entails a decision process much more complex than those required in vigilance studies.

Elliott (1960) suggests that the classical vigilance decrement has never been observed in any closely simulated radar task and that the social isolation usually found in vigilance studies is not typical of military situations. Smith and Luccaccini (1969) maintain that a vigilance decrement has never been demonstrated in an industrial situation. They suggest that this is due to the greater complexity of the industrial task and that the vigilance decrement can be explained by the lack of motivation of laboratory subjects who are insufficiently aroused to continuously perform what is an essentially meaningless task. Harris (1969) also cautioned in directly applying laboratory results to industrial inspection situations. Belt (1971) attempted to clarify this question by comparing performance by the same subjects on a laboratory vigilance task with an authentically simulated industrial inspection task. He found that the usual vigilance decrement occurred with the laboratory task but that a constant level of performance was maintained with the inspection experiment. The author suggested on the basis of subjects' comments that this was a result of the greater motivation present on the inspection task.

The general conclusion that emerges from these considerations is that we cannot blindly use all the results from vigilance experiments to predict performance in industrial situations, particularly when the laboratory studies have task characteristics unrepresentative of inspection tasks. On the other hand it would be foolish to ignore the considerable body of knowledge that has been gained on human performance in monitoring situations, particularly when results obtained from

vigilance studies are paralleled by data from tasks more representative of the inspection situation. As usual a process of discrimination is necessary when generalizing from research findings to the real world. It is necessary to examine the characteristics of any specific inspection task in detail in order to decide whether or not vigilance research is applicable. The more infrequent the defects, the less arousing the task conditions and the more prolonged the inspection period, the greater the likelihood of vigilance data being applicable, particularly if the defects are simple in nature.

3.1.1.2 Visual search considerations

This topic is surveyed in depth in Bloomfield (1970) and its applications to inspection described in Bloomfield (1975).

During visual search the eye makes a series of fixations and it is assumed that the probability of detection of a defect decreases as its distance from the fixation point increases. This leads to the concept of a visual lobe, which is a hypothetical area about the fixation point within which there is some arbitrary probability, e.g. 0.5, of detection. During search, the saccadic movements of the eyes give rise to a series of overlapping visual lobes which cover the area to be searched. Search is efficient if the area is covered completely with a minimum overlap of the visual lobes. The larger the visual lobe the more efficiently the area to be searched will be covered in the minimum time. The size of the visual lobe will be influenced by a number of factors. Individual differences in peripheral visual acuity, background luminance, length of exposure time, and discriminability of the target are all relevant variables. In addition to the size of the visual lobe, search efficiency will also be influenced by the fixation strategy adopted by

the observer. Both random and systematic sampling will cover the whole area to be searched and will ultimately detect any defect which is discriminable if it falls within the visual lobe. A systematic, regular strategy which optimally covers the entire search area with minimum overlap between the lobe areas will always be quicker on average. Bloomfield (1970) showed however, that even with well practised subjects, their search strategy was better fitted by a random rather than a systematic scanning model. In general the time taken to detect a target consists of two components, the search time itself and the time taken to respond when the target falls within the visual lobe i.e. there is no search involved. With readily discriminable targets this approximates to simple reaction time, but with near threshold targets more complex considerations using SDT or other models may become important (Pike, 1971).

One tends to assume that visual search considerations only become important in situations where large areas have to be scanned for defects. In reality, even if the total area to be searched is relatively small, if the targets, i.e. defects, are similarly small, the visual lobe is effectively reduced in size and so search is still necessary. Bloomfield (1975) considers three categories of visual inspection in which search is important. Where displays contain a number of small items some of which may be defective, the mean search time has been shown to be inversely proportional to the square of the discriminability of the defects. Discriminability can refer to differences in dimensions such as size, area or shape between good and defective items. This is known as a competition search situation. In multipart inspection a single complex object is shown to the inspector and he has to examine many features of the item which may be faulty. An example of this type of inspection is given in Harris (1966) in which ten items of equipment

were rated in terms of complexity, where this was largely in terms of the number of major parts each item contained. A high negative correlation was found between the number of defects found and the rated complexity. In the final type of inspection considered by Bloomfield, the inspection of sheet materials, a fault may be difficult to detect for several reasons. It may fail to emerge perceptually from its immediate background because of patterning effects, it may have a very low contrast difference with respect to the background, or it may simply be very small relative to the total area that has to be inspected.

The efficiency of visual search will be influenced by several factors in addition to those already discussed. Operator centred factors such as experience, eyesight and age will clearly be important as will task characteristics such as presentation rate (Perry, 1968) display size and shape (Baker et al., 1960) and the provision or otherwise of visual aids (Schoonard et al., 1973).

3.1.2 Task characteristics

3.1.2.1 Pacing and movement of the item being inspected

These variables have been considered in a number of studies because of their importance to industrial engineers in determining the maximum throughput of an inspection station possible without degrading the efficiency of defect detection. In terms of the inspection model of Chapter 1, the variables can be regarded as operating at the acquisition phase of the process.

Drury (1973) discusses two major aspects, the effect of the rate of movement of the item being inspected, and the effect of pacing per se.

These two aspects are not necessarily identical. It is well known, e.g. Blackwell (1959) that dynamic visual acuity is inferior to static visual acuity, and Sury (1964) has shown that pacing can degrade performance even if the subject is paced at the same rate as his unpaced performance.

With regard to the first variable, Williams and Borrow (1963) and Eriksen (1964) have shown that the rate of movement does not produce degradation in performance unless the angular velocity exceeds $7°$-$8°$ at the eye.

Drury (1973) considers a number of studies and attempts to deduce a general relationship between the time available per item being inspected and the probability of a correct decision being made. The two types of correct inspection decision concerned are P_1, the probability that a good item is accepted, and P_2, the probability that a faulty item is rejected. The studies considered covered a wide variety of inspection tasks, e.g. Fox (1964) concerning coin inspection, Perry (1968), glass bottle inspection, Sinclair (1971), food products, and unpublished studies on the inspection of sheet glass. In all of these studies the general effect is observed that as more time is allowed to inspect each item, the probability of rejecting a faulty item increases whilst the probability of accepting a good item decreases. Drury accounts for this finding by postulating that the inspection process consists of two phases, a visual search process until a potential defect is found or time runs out followed by a decision process. Thus for very short time intervals, corresponding to a high rate of pacing, relatively few of the potential faults are seen. This leads to a low value of P_2, the correct rejection probability, and also a high value of P_1, the correct acceptance probability, because the opportunity for making an error in

the other direction, i.e. false alarms, is lessened. At very long time intervals, or low pacing rates, search considerations become unimportant and the limiting values of P_1 and P_2 are largely a function of the SDT variables considered in Chapter 2, i.e. the intrinsic sensitivity of the inspector and his criterion.

A paper by Buck (1975) presents a highly detailed analysis of what he refers to as dynamic visual inspection, DVI. DVI occurs in a conveyor belt situation where the item being inspected moves past the inspection station at various speeds. Some of the factors affecting DVI can be identified as the direction of movement, the speed of the conveyor, the object interspacing distance and variability, and the lateral variability of the object on the belt. Buck discusses in detail the effect of various viewing constraints such as the 'viewing window' past which the items on the conveyor flow. DVI involves first visually tracking the moving inspection item. Crawford (1960) shows that up to an angular velocity of about 25^{o}-30^{o}/second the eyes can acquire a moving object in a single saccade. At greater speeds, additional eye movements are required and hence belt velocities of this order mean that the observer will spend more time in making visual corrections and therefore less time will be available for visual search within the object.

A number of studies, e.g. Ludvigh and Miller (1958) show that a complex set of factors affect dynamic visual acuity, e.g. the angular velocity of the item, and the exposure time available.

The belt velocity in DVI therefore plays a complex role both through its affects on visual acuity and the exposure time available for visual search (this variable will be considered in more detail subsequently).

Recent data from a number of studies (Rizzi et al., (1974), Nelson and Barany (1969), Smith and Barany (1971), Purswell et al., (1972) and Lion et al., (1975)) suggest that the following generalizations can be made about pacing and the more general area of dynamic visual inspection. DVI is improved with (a) increased exposure time for an object, (b) lower belt velocities, (c) greater interspacing between successive objects. Self paced inspection appears to be superior to externally paced work, (Williges and Streeter (1971), McFarling and Heimstra (1975)). Individual differences in peripheral visual acuity influence efficiency of DVI. The random or ordered arrangement of the items on the conveyor belt is another factor found to be important. The fact that the latter two variables are also important in static visual search, Bloomfield (1970) lends weight to the idea that DVI consists of detecting the on-coming item, visually acquiring it, searching it for faults and then making an accept reject decision, Buck (1975). Cochran et al., (1973) present a model for predicting the combined effects on DVI of change in visual angle, angular velocity, time to view, illumination and contrast.

3.1.2.2 Magnification, lighting and other aids to enhance defect discriminability

Techniques such as X-rays, ultrasonics, gamma rays and dye penetrants are all used in non-destructive testing in industry to render visible defects which could not be detected with the unaided senses. Often the resulting display or trace constitutes an inspection problem in itself, where the operator may have to continuously monitor the output for some subtle change which indicates a defect. These problems have been considered in Embrey (1975) (ultrasonic testing) and Lusted (1971) (X-ray photographs). The most common methods of enhancing defect

discriminability are lighting and magnification.

Lion (1964) compared the effects of fluorescent with tungsten lighting on a simulated inspection task involving grading ball bearings for size, and found that fluorescent lighting produced a significantly higher rate of work without any concomitant increase in errors. Lion et al., (1968) repeated the above comparison in a conveyor belt situation. The items used in one experiment were plastic discs with a link pattern drawn on their faces. 'Defective' items had a break in the link. The other experiment used plastic buttons as the test items, some of which had off-centre holes. Detection efficiency was higher for the link inspection task under the fluorescent lighting than with the tungsten lighting. However, there was no significant difference between lighting conditions with the button inspection task. The authors accounted for this latter result by suggesting that the link task was primarily a test of visual acuity whereas the button task involved more perceptual elements. It was proposed that the differential effect of the different light sources was due to the fact that the tungsten lamps, being point sources, were more readily obscured by the subject at the workbench, producing an effectively lower lighting level. This would readily account for the results in view of the relationship between visual acuity and ambient illumination. It seems strange that the authors did not verify this hypothesis by adjusting the different types of lighting to provide equal illumination with the subject in situ. The question of the very different spectral composition of the different light sources would seem to be another uncontrolled variable.

Further insights into this area are given by Sakguchi and Nagai (1973). In a comparison between various types of lighting in a Landolt ring

recognition task, it was found that eye fatigue, as defined by subjective reports, seemed to result from the narrow bandwidth of coloured lighting sources such as sodium lamps. This variable clearly needs to be considered when lighting for prolonged inspection periods is being specified.

Lighting considerations for inspection are far broader than merely specifying the optimum intensity and type of lighting to be used. Faulkner and Murphy (1973) illustrate a number of ingenious ways in which special purpose lighting can be used to enhance the discriminability of defects. They point out that the simple expedient of increasing lighting levels is not necessarily the most effective way of increasing task performance. They describe a number of lighting techniques including grazing illumination, polarized light, spotlighting, dark field illumination etc. which can be used in various situations. Case studies from the glass industry in which lighting is used to enhance defect discriminability are given in Gillies (1975).

Inspection using microscopes and other magnification aids has become increasingly important with the growth of the microminiature integrated circuit industry. Smith and Adams (1971) and Smith (1975) propose that over a wide range of conditions, optimum performance can be expected when the visual angle subtended by the defect, as a result of magnification, is between 9 and 12 minutes of arc. Froot and Dunkel (1975) point out, however, that a number of other parameters of the microscope system need to be taken into account, including resolving power, aperture and depth of field. They cite a case study where two groups of inspectors could not agree over the incidence of defects in a batch of products. It transpired that although they were using microscopes of identical magnification, they had differing resolving power for defects.

A detailed case study involving the use of magnification in the inspection of rubber seals is given in Astley and Fox (1975).

3.1.2.3 Complexity

The variable of complexity is an extremely difficult one to quantify. Firstly there is the question of defining an adequate index of complexity. In inspection studies complexity has usually been described in terms of the number of items which would be potentially defective on each unit inspected. For example Harris (1966) found that the index of complexity assigned to circuit modules by a panel of experienced judges, correlated highly with the number of major parts making up the item. It seems likely however, that variables such as the arrangement of the parts constituting the item are also important, either because of the Gestalt consideration suggested by Fox (op. cit.) or through the facilitation or otherwise of an efficient search strategy. At least two other aspects of complexity can be considered, the complexity of the defect itself and the complexity of the background in which it is embedded. The conspicuity of the defect could well be regarded as the degree to which it shares common attributes with the field within which it is embedded. One might expect this variable to be partly dependent on the number of these shared attributes which occur within the defect and its background. Possibly a quantitative estimate of detectability as a function of this concept of complexity could be derived.

The clearest effect of complexity on inspection efficiency comes from the Harris study (op. cit.). Inspector performance in terms of defects detected showed a strong negative correlation with the complexity of

the module. A DVI task used by Purswell et al. showed a similar relationship, where subjects had to remove grids containing geometrical patterns from a conveyor belt. McFarling and Heimstra (1975) used printed circuit boards containing varying amounts of circuitry as constituting differing complexity levels. Decision time was found to increase with increasing circuit complexity whilst defect detection performance declined. Another variable investigated was pacing, and it was found that as circuit complexity increased, there were larger increases in decision time for self paced subjects than for their machine paced counterparts.

It seems therefore, that increases in complexity can be regarded as generally reducing defect detection probability. This can partly be accounted for by visual search considerations, because in paced situations there will be many occasions when the inspector will not have had time to examine each potentially defective attribute of the item in the viewing period allowed. The persistence of the effect even in self paced situations suggests however that other, perceptual variables may also be important. Studies of image interpreters, e.g. Powers et al., (1973) have shown that in multiple defect situations, interpreters often have a far lower detection efficiency for subsequent defects after an initial one has been found.

3.1.2.4 Display organization

This variable is another which one might expect to be influenced both by visual search and perceptual considerations. Williges and Streeter (1971) found no significant differences in performance as measured by correct defect detections and false alarms between an ordered and a random display consisting of 600 transparent discs containing

occasional pin hole defects. This result is in conflict with visual search studies and with other inspection studies which have considered the same variable. The authors suggest that this may have been due to the close proximity of the inspection items in both the ordered and random arrangement. In fact many of the subjects after completion of the experiment stated that they had been unaware that the discs had been displayed in both a random and ordered fashion.

In a laboratory visual search situation Bloomfield (1970) found that an irregular display took longer to search than a regular one. One would therefore expect this variable to be of importance in paced situations, and in fact in the Williges and Streeter study a significant interaction was obtained between paced and unpaced presentation and the ordering or otherwise of the items. Detection performance was significantly better with the regular display under the paced condition than the irregular display.

The effects of display arrangement are not necessarily due to visual search considerations alone. An industrial study by Fox (1964) considered the effects of the random or regular arrangement of coins on a conveyor belt at the Royal Mint. The regular display proved considerably superior to the random arrangement and Fox proposed that the result could be explained in terms of Gestalt theory, i.e. the defective coins emerged more readily from the regular, 'good gestalt' background than from the random, and therefore difficult to perceive arrangement. An alternative explanation is simply that the search time was longer in the irregular display case. Since the inspection was paced, a longer search time would produce a lower rate of detection of defects independent of any effects on the intrinsic perceptibility of the defects. It

is not possible to decide which of these explanations is appropriate from the Fox paper. Indeed it is difficult to see how the two variables could be experimentally disentangled. Nevertheless, the possibility of effects at a perceptual level in addition to the peripheral consideration of visual search cannot be ruled out. Scott Blair and Coppen (1942) suggested that 'learnt gestalten' seemed to develop in certain skilled operators and Thomas (1962) quotes several industrial examples where inspectors seemed to detect defects on a wholistic, figure/ground basis, rather than by a process of search. Eye movement studies in situations of this type would presumably clarify the issue as to whether search was taking place.

Lion et al., (1975) compared performance on a single line conveyor belt system with a three line system. The test items consisted of plastic discs which contained either a broken link pattern (defects) or a complete pattern. Detection performance was superior on the three belt system. However, in order to provide the same amount of material per unit time to the inspector, the single belt was run at 18cm per sec., i.e. three times the speed of the three line belt. Since this exceeds the limits proposed by Williams and Borow (1963) (i.e. 2cm per sec.) at which performance decrement occurs, it seems likely that the obtained results were due to this variable rather than the organization of the display.

3.1.2.5 Signal rate

Signal rate is an important variable which has received considerable attention in both vigilance and inspection research. Early work in vigilance tasks, Jenkins (1959), Kappauf and Powe (1959), suggested that lower signal rates produced reduced detections by causing an increased vigilance

decrement. Colquhoun and Baddeley (1964, 1967) showed however that this was at least partly an artefact due to the subjects having an inappropriately high expectancy of the signal rate. It was further shown by Colquhoun (1961) that it was the conditional probability of a signal given an event occurred, that determined detection efficiency rather than the actual frequency of signals in time. In general it has been shown that an increase in signal probability produces an increase in both the detection rate and the false alarm rate, e.g. Baddeley and Colquhoun (1969), as would be predicted from the SDT considerations discussed in Chapter 2.

Rather few inspection studies have explicitly varied defect rate or probability as an explicit experimental variable. Fox and Haslegrave's study (1969) had the virtue that it was conducted in an industrial environment. They attempted to verify Colquhoun's finding of the importance of signal probability as a determinant of defect detection efficiency as opposed to stimulus frequency. They investigated both a static and a conveyor paced situation where screws were inspected for a variety of faults. No significant differences in correct detection probability were found in the paced condition, but the static condition showed the expected straight line relationship between detection probability and defect probability. It is not clear why the effect was not observed in the paced condition, but it seems likely that other variables such as the conveyor speed (52 feet/minute) swamped the probability effect.

It is clear however that data from experiments of this type have to be interpreted with care. A well known study by Harris (1968) employed four different defect rates on a scanning type inspection task using

four groups of inspectors. Using as measure of inspection accuracy the proportion of defective items that were detected, Harris stated that the accuracy of inspection declined with decreasing defect probabilities. Baker and Schuck (1975), however, took the Harris data, and using the equal variance SDT model, calculated that d' was not significantly different for any of the probability conditions. They therefore stated that Harris was in error and that inspector accuracy did not change as a result of signal probability. This is a good illustration of the confusion that ensues when differing measures of inspector performance are used. In reality both author s were correct within the descriptive terminology they were employing. If anything, Harris's conclusions were to be preferred to those of Baker and Schuck because the latter authors did not bother to check the appropriateness of the equal variance assumptions implicit in their use of the SDT model. The results are of course explicable in terms of the inspectors employing a criterion appropriate to the ongoing defect probability.

The use of artificial signals in an inspection task to increase the apparent defect rate and thereby enhance the detection rate (and false alarm rate) has been proposed by Wilkinson (1964). Although his 'inspection' task is actually a laboratory vigilance task, there seems to be no reason why the technique should not be applicable in a real life situation as long as the artificial 'defects' could be readily separated from the real ones and the situation was such that false alarms were not 'expensive'.

The only dissonant study on the general conclusion that an increased defect rate enhances detections is one by Sosnowy (1967). This study, which was a simulated inspection of ball bearings, only showed the

usual relationship at a high rate of pacing, i.e. 240 items per minute. Unfortunately the original study could not be obtained by the reviewer (it is reported in Badalamente and Ayoub, (1969)) and hence the index of inspector accuracy used and the precise experimental conditions could not be ascertained.

One study exists in which the variable discussed in Chapter 2, a within session change in the defect rate is considered (Sims (1972)). The study used printed circuit board inspection as a simulated industrial task, and showed that although inspectors generally accurately perceived the quality level of the incoming product, there was considerable inter-subject variability in their ability to adjust to changes in defect rate.

The remaining studies in which defect rate is an important but not explicit variable have been discussed in the SDT review of Chapter 2.

3.1.2.6 Number of inspectors

Schlegel et al., (1973) compared the performance of a single versus a dual inspector system in a Landolt ring, conveyor based inspection task. The task used an inspection period of 45 minutes in order to investigate the effects of the two systems on a possible vigilance decrement. In terms of probability of detecting a defect, the two inspector system was significantly superior to the single inspector. The false alarm probability showed no significant difference. These results were in line with what would be predicted from the simple statistical combination of the two inspectors efficiences. The two inspector system had the additional bonus that there was significantly less vigilance decrement in performance. The reasons for this are not obvious, since the

inspectors were separated by screens and could not monitor one anothers performance. It seems possible that the element of competition in the two inspector case provided an increased level of arousal and hence reduced any vigilance decrement. This study has obvious implications for real life inspection systems. It should be noted, however, that the stimuli used were considerably easier than those normally found. A slightly different arrangement was used by Lion (1975). In an experiment described in detail earlier (Section 3.1.2.5), the performance of inspectors on a three line arrangement of items on a conveyor belt was compared with the performance of two inspectors seated on opposite sides of a six line conveyor. Performance in the two inspector arrangement was significantly better than using a single conveyor both in terms of correct detections and fewer false alarms. The authors attributed the result to the stimulus of doing a job in unison, the feeling of competition and the reduction in boredom due to the opportunity to talk. Since the length of each session was only 12 minutes, the first two factors seem more likely to be important than the last.

A study by Morrissette et al., (1975) showed a similar improvement in performance in team monitoring using a laboratory task. The evidence suggested that performance was improved by social facilitation, i.e. the monitors performed better when working together than separated.

The evidence strongly suggests therefore that inspection efficiency will be improved by operator redundancy. In practice, such redundancy does occur in industrial situations when particularly critical products are being inspected.

In using more than one inspector, a decision has to be made whether the

economic advantages accruing from the higher detection efficiency expected exceeds the cost of additional inspectors.

3.1.2.7 Repeated inspection

Batches of product are sometimes inspected repeatedly in order to raise the probability of defect detection. This may be done by the same inspectors or an independent inspection may be utilized. This latter arrangement is clearly preferable since re-inspection by the same inspector means that the same defects are likely to be missed particularly if there are systematic errors in inspection strategy. Belbin (1957) gives an example of the repeated inspection of ball bearings, although he does not state whether or not it was independent. After the first inspection 63% of the defects were found, and a further 16% were found after the second pass through the system. Harris and Chaney (1969) p.78 describe the repeated inspection by ten independent inspectors of an electronic module. Performance in detecting critical defects continued to improve (at a slightly increasing rate) with the addition of independent inspections up to a total of six, after which performance levelled off. Performance in detecting non-critical defects, on the other hand, continued to improve as additional inspections were added. These results tend to confirm that different inspectors do not necessarily find the same defects. Eilon (1961) presents an operational research type model which specifies situations, depending on the cost of inspection, when it is economically worthwhile to recycle products through an inspection station. The question of the efficiency of independent repeated inspections versus a team inspection approach is a complicated one. One way of looking at the situation is to assume that an interacting inspection team combines the advantages of social

facilitation noted in the last section with the theoretically higher probability of detection due to the possible 'blind spots' of the individual inspectors being eliminated by the overlap of abilities. On the other hand it is possible that the team consensus as to what constitutes an acceptable product may be incorrect. The effect of this could mean that the findings of more accurate inspectors within the groups could be rejected, or not reported, due to group pressures. Further work is clearly needed in this area.

3.1.3 Environmental factors

Traditionally, these factors include heat, lighting, noise and workplace design. Lighting has already been considered in an earlier section. There have been no studies in which heat has been considered as an explicit variable. The evidence from vigilance tasks (J.T. Mackworth (1969), p.167) suggests that whereas cold may slow reactions and interfere with detections particularly at the beginning of a session, heat tends to increase missed signals at the end of a prolonged vigil. These results may be extended to inspection tasks with the caution suggested earlier.

McFarling (1974) considered performance under noise conditions in a simulated printed circuit board inspection task which investigated the effects of a number of interacting variables. Inspector performance in a quiet condition was significantly better, measured by defect detection probability, than under a 90dB white noise condition. False alarm scores were not significantly different in each case. This latter result is unexpected since in vigilance tasks the decrement in detection performance under noise conditions is generally a result of an increase in beta (Broadbent 1970). The result obtained suggests a

lower sensitivity in noise. The author was unable to account for the effects but J. Mackworth (1969) suggests that similar results in vigilance experiments are due to distraction effects. A very similar study by Ehlers (1972) also considered the effects of noise on the inspection of printed circuit boards. Three white noise levels, 50, 70 and 90 dB were employed, and defect detection performance was significantly worse in the 90 dB condition. A deterioration in performance over the 70 minute inspection task occurred under the 50 dB noise condition but not under the 70 or 90 dB condition. False alarms were not analysed because only 10 in the 19,000 good circuits inspected were called bad. These results suggest that in common with other monitoring tasks noise levels of 90 dB and above should be avoided in inspection situations, although a moderate level of noise serves to reduce vigilance effects, presumably via its arousing qualities.

Workplace design is clearly a variable which needs to be considered in the design of inspection stations, where a worker may be carrying out a visually demanding job for long periods. Astley and Fox (1975) present a case study which shows how anthropometric considerations are taken into account in this situation.

3.1.4 Organizational factors

Organizational factors are important determinants of the effectiveness of any real inspection system. If there is disharmony within the inspection team, or conflict between production and quality control, then even the most perfectly designed inspection system will either fail to function effectively, or its findings will go unheeded. We will first consider the more general socially orientated factors that

influence the effectiveness of inspection and then some of the specific organizational factors which directly influence inspection performance.

3.1.4.1 Managerial and social aspects

McKenzie and Pugh (1957) consider the effects of the relationship between the production and inspection departments in industry. Where lack of communication exists, production departments will generally be critical of attempts by the quality control section to assess their work. Social pressures by their workmates will often persuade inspectors to modify their judgements even when there has been no objective change in quality. The authors comment on the very high degree of individual variation in inspectors and even inconsistency within their own judgements. The general point made is that it is this inconsistency that leads to the deterioration of relationships between inspection and production departments. Recommendations are made to regularly calibrate inspectors with reference standards.

Belbin (1957) discusses the issue of the differences between the customer's quality standards and those adopted by inspection departments. He presents a real life example which shows that many of the complaints from customers of a particular firm were due to defects for which the inspector had not been told to look for. Similarly many of the faults for which inspectors did reject items were of no consequence to the customer. The effect of the continuing complaints from the customer was to make the inspector reject more and more products for the wrong reasons (presumably via a change in criterion). Belbin points out that the quality standards required may fluctuate due to variations in demand, and suggests a scheme for defining quality levels so that

inspection criteria can be readily modified. The effects of social pressures on inspectors is illustrated by an example from a hosiery factory where inspectors had to feed back repairable defects to their workmates for mending. The defects could be classified as rejects or mendable, the latter producing a much higher rate of pay for the menders. A combination of social pressure from the operatives on the inspectors, and a lack of clearly defined quality standards meant that virtually all of the defects were classified as being mendable.

McKenzie (1958) regards inspection accuracy as being determined by basic individual abilities, environment and formal organization, and interpersonal and social relations. One of the most important organizational factors is the provision of reference standards for the inspector. Thomas (1962) emphasizes the importance of clear, unambiguous examples of both rejects and perfect product being provided because of the tendency of perceptual judgements to drift with time. Equally important is the provision of inspection instruction. Raphael (1942) describes how some viewers inspecting fabric were rejecting 53 per cent of the product whilst others in the same group were rejecting only 13 per cent. It transpired that the specification allowed a tolerance of 3 millimetres but some of the viewers had not been informed of this. The question of the drift of perceptual standards and the possibility of ameliorating this by a weekly 'calibration meeting' is discussed in McKenzie (op. cit.). The same paper cites further examples of the effects of social factors on inspection standards. A group of two operators and an inspector worked as an isolated group apart from occasional visits by a supervisor. There was obviously a tendency for the groups to identify itself as a cohesive whole and hence the inspector could not be expected to make decisions that might disrupt

the group, unless they were based on unambiguous evidence. Mitchell (1935) provides evidence of poor social relationships between operator and inspector biasing the latter's inspection standards, as does Roethlisberger and Dickson (1939).

McKenzie (op. cit.) points out that inspection sections tend to have a tightly knit social structure, partly because of their small numbers and partly because of their control function vis a vis production, which tends to lead to strained relations between the two functions. A survey conducted amongst patrol inspectors suggested that they felt that they were viewed unfavourably by production operatives.

Thomas and Seaborne (1961) criticize many experimental studies of inspection because they remove the individual task from its socio-technical context and examine it purely in terms of psychophysical performance. They point out that the laboratory study lacks much of the concomitant information which serves as a frame of reference in the industrial task. Often the industrial inspector ulitizes sources of information, such as a knowledge of the supplier, which enables him to use an appropriate criterion, in the SDT sense. Of course if such information is unreliable, the inspector's accuracy may be reduced. There is very little opportunity for inspectors to develop consistent standards in a situation where the range of quality of the input items varies widely and where there is little feedback as to the quality of the final product required. Again, in SDT terms, such feedback is known to be necessary to develop a stable criterion. In many real inspection systems the inspector is required to continuously modify his standards to take into account factors such as market conditions, level of output and demand. Because of the emphasis on inspection

studies in critical areas where quality specifications are clearly defined, this variable quality aspect of much 'bread and butter' inspection has tended to be neglected. In an analysis of a particular inspection task, Thomas and Seaborne (op.cit.), showed that the inspector's function could be regarded as satisfying the sometimes conflicting needs of the sales organization, the production department and the raw material purchasing department. The inspector utilized his knowledge of the manufacturing process to inform production operatives of deterioration in the process. His inspection of raw materials provided information influencing his judgement of the final product. Finally he was frequently approached by the sales manager, and on the basis of information on the state of the market would raise or lower his standards with regard to certain faults.

This analysis suggests that in real life situations the sources of information utilized by the inspector in arriving at his accept/reject decision are far more complicated than in the laboratory situation. He receives feedback from sales and from any check inspection that may be carried out. He receives feedforward information about the state of the raw materials and from his knowledge of the manufacturing process, and finally he utilizes the sensory data present in the actual item being inspected.

This situation can readily be analysed in SDT terms. We can regard the training and any reference standards that are provided as influencing primarily the sensitivity of an inspector, his ability to recognize the cues which indicate good or bad product. The instructions which apply to an inspection task at a specific time will provide the base data by which the inspector sets his criterion level, i.e. decides on what standards of acceptability shall apply at that time. This information

is modified by feedback from sales as to the importance of particular defects. This can be regarded as changing the inspector's implicit pay-off matrix. Feedforward from production modifies the criterion by affecting the subjective probability of a defect occurring. We can see that in real inspection systems, the ability of the inspector to modify his criterion on the basis of additional information, is, as pointed out in Chapter 2, an important quality.

3.1.4.2 Motivational variables

As Weiner (1975) points out, it is surprising that there has been little scientific investigation of the effect of motivational factors on inspection, particularly in view of the popularity of this approach in industry. Many propaganda-style exercises such as the 'zero defects' programme (Swain 1972) have been tried and varying degrees of success reported. Unfortunately the quality control journals which report such research do not employ indices of performance which are sufficiently precise to make unambiguous conclusions possible.

The use of financial incentives is the obvious way to influence motivation, and has no doubt been employed in many companies for this purpose. There is, however, little hard evidence as to its efficacy or otherwise. Ergonomists working in inspection have tended to reject financial incentives in this area. However as Weiner (op. cit.) points out these assumptions may not be true for other forms of incentive such as knowledge of results. Mitten (1957) describes how female roller bearing inspectors were allowed to go home as soon as they had achieved their inspection quota. Although this incentive seemed to be effective in this case, social disharmony could result in inspection systems such as

that investigated by Embrey (1970) in which there were a wide range of ages and abilities.

Vigilance studies have not shown financial incentives to be very effective in maintaining performance, possibly because the financial rewards offered are unrealistically small (Wiener 1969). In many SDT studies attempts have been made to influence performance by manipulating the payoff matrix. As shown in Chapter 2, this has not proved very successful, largely because of the difficulty of relating subjective utilities to external rewards. Perhaps this is again a case of insufficient quantities of money being offered. It is possible that financial rewards may influence performance through indirect means such as modifying an inspector's visual strategy. Bloomfield (1970) describes experiments in which he produced extremely large increases in visual search performance by offering monetary rewards.

The use of Knowledge of Results (KR) in training for perceptual skills will be considered in detail in a subsequent section. It has been suggested by some workers that KR exerts a motivational effect quite distinct from its informational content. For example in vigilance tasks, where there is very little informational content, KR is known to enhance performance. Even though KR may be difficult to provide in a real inspection task, the evidence suggests that the benefits of KR extend to sessions where it is withdrawn, Wiener (1963), Annett (1966). Drury and Addison's (1973) industrial inspection study clearly demonstrated the enhancement of performance due to KR, but it was still difficult to say whether this was due to its informational content, motivational effect, or a combination of both. Despite Annett's (1969) position that motivation is an unnecessary construct in explaining the

effects of KR, the evidence overall seems to suggest that at least some motivational effect must be present whenever information feedback is given.

3.1.5 Individual factors

In virtually all studies of inspector proficiency the largest contribution to the variance of the results is individual differences between the inspectors. In view of this, it is surprising that so little work has been done on identifying the source of this variability. Usually individual differences are regarded as a 'nuisance variable' which experimenters seek to eliminate from their designs. A detailed consideration of the nature of the individual differences which affect inspection performance would seem to be a most effective way of enhancing performance in inspection systems where this task and environmental variables have already been optimized. The major ways in which individual differences affect the performance of an inspection system are through selection and training. These will be considered in detail in subsequent sections.

3.1.5.1 Selection

As implied by the inspection model in Chapter 2, two major groups of variables will affect a person's ability to perform inspection, peripheral factors such as eyesight which affect the acquisition of the necessary sense data, and cognitive factors which determine how the sense data will be interpreted. Although differences in cognitive skills can be reduced by training, it would not be surprising if there were intrinsic differences in certain cognitive abilities necessary for

inspection which selection procedures could identify. As we shall see, up to now, research on selection methods appropriate to inspection has proceeded in a very ad hoc manner. Very little attempt has been made to analyse the non-intellectual skills which may be necessary for inspection. Nearly all of the studies which exist on selection for inspection have attempted to correlate some performance index with a more or less arbitrary group of standard industrial tests.

In the sections which follow, we will first consider the individual variables that more directly affect the target acquisition phase and then consider the work on the intellectual and cognitive factors that affect inspection performance.

3.1.5.1.1 Visual abilities

It is clear that the visual skills of an inspector are an important determinant of his overall efficiency. Depending on the nature of the task, static or dynamic visual acuity will be important. Ayers (1942) and Tanalski (1956) both found strong relationships between various static visual measures and indices of inspection performance. Nelson and Barany (1969) have described a dynamic visual acuity selection test suitable for conveyor based inspection. Standard visual tests have been developed for static visual skills required for various industrial tasks including inspection. The best known of these, the Orthorater, is described in Trimby (1959). In situations where colour is an important cue in identifying defects one of the standard colour blindness tests needs to be incorporated in the routine testing schedule.

Virsu (1972) presents a sophisticated review of the visual factors affecting the inspection of radiographs. The use of the Modulation

Transfer Function (MTF) (Cornsweet 1970) approach to measuring visual performance is proposed as being far more efficient than traditional measures of visual acuity. He proposes that in the case of radiographs, the translation of monochrome photographs into coloured slides using MTF techniques would optimize the visual performance of the interpreter.

3.1.5.1.2 Age

The effects of ageing are to produce a gradual loss in sensory capabilities such as visual acuity. This is however compensated for by the perceptual skills acquired through long experience. Few studies are available in which age and inspection accuracy have been considered. Sheehan and Drury (1971) and Drury and Sheehan (1969) report a gradual decline in d' with age of about 0.2 units per ten years of age. Since their results were based on only five inspectors however, they may not be readily generalizable. Jamieson (1966) showed increasing performance with age in electronics inspectors, Jacobsen (1953) found accuracy increased up to age 34 and then declined to age 55. Evans (1951) found no effect at all.

The evidence from vigilance tasks shows only slight effects of ageing, unless the stimulus presentation rate is high, e.g. Thompson et al., (1963). This is consistent with the general finding that short term memory capacity declines with age. This will clearly have a deleterions effect on rapid conveyor paced inspection, where the older inspector may have difficulty in retaining the information from the display in his short term memory before the next item has to be inspected.

In general it seems likely that older persons will make good inspectors as long as they are employed in tasks in which their perceptual skills

are utilized but which do not place heavy demands on their sensory or information processing capabilities.

3.1.5.1.3 Personality variables

Virtually no work has been done on the use of personality tests on selecting for inspectors. Colquhoun (1959, 1960) used the Heron personality inventory to investigate time of day effects on detection performance in extreme personality groups but the task was a laboratory experiment. In vigilance tasks of this type, the Eysenck Personality Inventory (EPI) has been extensively employed to investigate the importance of the extraversion-intraversion personality variable in influencing performance. Although it is often stated that the intrinsically more highly self-aroused (according to Eysenck's theory) intraverts do better at vigilance tasks, the results are somewhat more equivocal (Mackworth 1969). A number of other studies using personality variables in vigilance experiments, have found very few significant correlations with performance. In spite of the lack of success in utilizing these variables in vigilance tasks, there does seem to be a case for investigating the use of some of the more easily administered tests, such as the EPI, in an inspection context, particularly if prolonged monitoring is involved.

3.1.5.1.4 Sex

Although women are more often found in the inspection departments of manufacturing industry than in most other sections, there is no solid evidence that there are substantial differences between the sexes in ability for inspection work. Only one inspection study, McFarland (1972),

seems to have included sex as a major variable, and the only difference found was in the greater variability of response times for women. Of the vigilance studies that have considered this variable, six found no significant differences at all (Waag et al., (1973), R. Smith et al., (1966), Gale et al., (1972), Kappauf et al., (1955), Kirk and Hecht (1963), and McCann (1969)), in two men performed better (Neal and Pearson (1966), Heimstra et al., (1967)) and in three others there was no significant main effect of sex but significant interactions with other variables (Bakan and Manley (1963), Krkovic and Sverko (1967), Whittenberg and Ross (1953)). There is clearly insufficient evidence for inspectors to be selected purely on the basis of sex.

3.1.5.1.5 Selection tests

A number of testing procedures have been applied to the selection of inspectors with generally disappointing results. Many early studies suffered from the drawback of correlating supervisors ratings, rather than objective measures of performance, against test scores. Wiener (op. cit.) points out that such ratings are probably based on the supervisor's perceptions of earnestness and co-operation and that the correlations between these variables and actual inspection performance are unknown.

Link (1920) obtained a correlation of 0.50 between output rate and tests of card sorting number cancelling and number group checking, with munition inspectors. Wyatt and Langdon (1932) were unable to obtain significant correlations using four standard industrial tests and eight inspection tasks. Sartain (1945) obtained a high multiple correlation, $R = 0.79$, between standard industrial tests and supervisor ratings.

Only low correlations were obtained using similar techniques, by Schuman (1945) and Tiffin and Rogers (1941).

One test specifically designed for inspection work exists, the Harris Inspection Test (HIT), which is a short pencil and paper instrument. Harris (1964) found significant correlations between the HIT and three out of four electronic inspection tasks. When the test was administered as part of a battery to 26 machined parts inspectors, however, no significant correlation was obtained between test scores and job sample measures of performance (Harris and Chaney 1966). By combining together the two most valid measures in the battery, the number comparison section of the Minnesota Clerical, and the Industrial Mathematics test a multiple correlation of R = 0.75 was obtained. No further validations of the HIT have appeared in the literature.

It is clear that there has been no systematic attempt to isolate the underlying individual factors which are important in inspection and to incorporate these variables in selection procedures. It is not surprising, therefore, that attempts to correlate performance with arbitrarily selected standard tests have been unsuccessful. The fact that the HIT, although successful in the application for which it was originally designed, did not predict performance in another type of inspection, suggests that it did not measure any general underlying ability.

We can conclude from the survey of the selection procedures currently in use in the inspection area, that a more fundamental approach is needed. Such an approach, based on the cognitive skills underlying inspection, will be set out in the theoretical section of this review (part 3.2.1).

3.1.6 Training for inspection

Training for inspection is a neglected area in industry. Many inspector training schemes described in the quality control literature (e.g. Browne 1965) have as their aim the imparting of the background information judged necessary to perform inspection. However there is virtually no emphasis at all on training for the perceptual skills necessary to detect defects and to recognize acceptable products. When training schemes which purport to accomplish this latter aim are described, the emphasis is often on the large savings that the scheme is alleged to have produced rather than the details of the technique and the method employed for evaluating it.

Learning effects in laboratory simulations of inspection tasks have been noted without specific training being given, e.g. Smith and Adams (1971) and Lion et al., (1968). In an industrial setting Chaney and Teel (1967, 1969) have employed a variety of techniques in training machined parts inspectors and photomask inspectors. Their most commonly employed technique, known as job sample instruction, essentially involves giving inspectors knowledge of results (KR) after inspecting test items containing typical defects.

KR (Knowledge of Results) has been employed in training for perceptual skills in a number of applied studies. Martinek (1965) reports a study in training photo-interpreters. Photo-interpretation, or image interpretation as it is alternatively known, consists of identifying military targets, which may be camouflaged, on aerial reconnaisance photographs. This task bears a close resemblance to an industrial inspection task. Martinek found that providing the photointerpreters with an error key

which analysed the commonly occurring errors of previous interpretations, produced significantly fewer errors of commission than a key in which the characteristic features of the various target types were set out. The error key can be regarded as a form of group KR, although it is incomplete. As there was no significant difference in the number of targets detected, this suggests that the result was due to a change in sensitivity rather than response bias. Another image interpretation study, Cockrell and Sadacca (1971), showed that KR was more effective in a team context when team members first scanned a photograph independently and then discussed the results together immediately afterwards. The greatest gains in proficiency were made by the least able members of the team and although detection performance was improved, significantly, there were greater gains in reducing the number of misidentifications and false alarms.

Another image identification study, Powers et al., (1973) attempted to improve performance by modifying the interpreter's scanning strategies rather than by KR techniques. Structured search practice increased the number of target detections at the expense of a greater number of false alarms. A 'speed reading' training technique, designed to reduce fixation times and expand the visual field, succeeded in halving the search time without changing the accuracy of the interpreter. Training using the 'error key' approach discussed earlier significantly reduced false alarms. Brock et al., (1974) provided a complete training programme for non-destructive testing radiograph examiners. The programme consisted of tape/slide presentations which gave examples of various types of defect, and then a KR phase where radiographs containing defects were presented and the student was given full KR after he had attempted to identify them. A self-administered test then determined if

the student should go on to the next module or re-take previous modules. There was a gradual increase in difficulty of the radiographs both within and between modules. The programme was highly successful in that it reduced the time to train to the required criterion from an average of 80 hours to 10.9 hours. Wallis (1963) describes the use of KR in the training of a complex perceptual task associated with a weapons system. Although no details were given of the classified system, substantial reductions in training time were found.

In view of the importance of training as a relatively low cost means of increasing inspection accuracy, it is surprising that no studies appear to have been performed to specifically investigate training techniques for inspection. It is felt that training is one of the areas which requires experimental investigation using tasks which are more representative of industrial inspection situations. In section (3.2.2) some of the theoretical considerations which will provide guidelines for research in this area will be considered in depth.

3.1.7 Conclusions regarding the general literature of inspection

It seems clear that inspection performance in a given situation will be the result of a complex interaction between some of the factors considered in the review up to this point. Although there is a paucity of research in a number of areas, two in particular will be considered in further detail in the theoretical review which will constitute the next part of this chapter. The first of these will be the implications for selection of some of the cognitive variables which may account for individual differences in inspection skill. The remainder of the

review will consider in detail some of the work that has been performed in the area of training for perceptual skills.

3.2 Theoretical literature survey

3.2.1 Cognitive variables in selection

An analysis of the cognitive skills necessary for the performance of inspection tasks provides useful guidelines as to possible new approaches to the problem of selection for these tasks.

In Chapter 2 the decision making aspects of inspection were emphasized in the context of SDT. It was implicitly suggested that decision making skills were learnt rather than innate. Although there may well be innate differences in the ability of individuals to utilize evidence concerning the existence of a defect to the full, it is not clear how one would devise selection procedures to identify such individuals, apart from simulation exercises.

Other aspects of the cognitive skills utilized in inspection seem to hold more promise. In an earlier section, when considering the visual search aspects of inspection, it was pointed out that in very many inspection tasks, the critical defect in an item is difficult to detect because it does not emerge perceptually from the background in which it is embedded. In fact this is by far the commonest situation. The defect virtually always occurs in noise. The noise may be an external random perturbation which degrades the detectability of a defect by obscuring or modifying the critical cues such as contours, angles or shade gradients which define the defect. Alternatively we can consider

a higher level 'cognitive noise' which results because the background within which the defect is embedded shares so many attributes of the defect that it is difficult to separate them perceptually. On this basis, if there are inherent individual differences in ability to separate a wanted configuration from the background in which it is embedded, this should provide a basis on which to select inspectors.

In fact a very considerable body of research exists on the ability of subjects to 'disembed' stimuli from their backgrounds. This dimension of individual differences is concerned with the field dependence - independence continuum. Field dependent individuals show great difficulty in breaking up an organized visual field in order to keep a part of it separate from that field. Field independent subjects on the other hand are readily able to extract a wanted configuration from the confusing field in which it may be embedded. These contrasting styles of functioning are found to be extremely stable with time and to affect performance in a wide variety of tasks. The work in this area is associated with Witkin and his collaborators (Witkin (1950), Witkin et al., (1962), Witkin et al., (1954)). Witkin developed an effective pencil and paper test for discriminating between field dependant and independant individuals, the Embedded Figure Test (EFT). This test has been validated in a wide variety of situations (see Witkin et al., (1971) for a review) and has demonstrated, for example, consistent sex related differences on the field independence dimension.

In view of the relevance of this aspect of individual functioning to many applied perceptual tasks it is surprising that so few studues have considered this variable. There are in fact very few studies applicable to the inspection area. Thornton et al., (1968) used an

identification task in which subjects had to locate and identify a series of small buildings in an aerial photograph. There was a highly significant correlation (0.72) between scores on the EFT and target detection performance in terms of number of targets detected during the time limits for the test.

Seale (1972) however, found no significant relationship between EFT scores in an aerial target acquisition simulation using aircrew as subjects. However, examination of the data suggested that the subjects were already a highly selected field independent group (airline pilots were used as subjects) and hence the test was likely to be too insensitive to differentiate between individuals. This may not detract from its general value as a selection instrument however.

There seems to be a good case for further work to be carried out using the field independence variable.

Other cognitive factors also deserve attention as potential dimensions upon which to base selection procedures. Field dependence is related to, but distinct from the dimension of distractability. It was originally hypothesized that field independence could be interpreted as the ability to resist distraction rather than to overcome the effects of the embedding context (Witkin et al., 1962). Subsequent work by Karp (1963) using factor analytic techniques suggested that distractability was a separate factor from field independence, but that the two factors were moderately correlated. Karp (op. cit.) differentiated between distraction and embeddedness as follows. In the distraction situations, the figural properties of critical items remain intact. In the embeddedness situation, the critical item or its parts are organized into new,

competing gestalts which serve to break up the original figure. A distracting context may be thought of as obscuring a critical item without changing the nature of the item, whereas an embedding context serves to obscure a critical item because it changes the nature of the item. Both types of field occur in inspection situations. Karp (1962) produced a number of tests designed to measure the characteristic of distractability. These will be described in detail in a later chapter.

Sack and Rice (1974) consider attention to have a directional aspect which can be analysed into at least three processes: degree of selectivity, resistance to distraction and shifting. Gardner and Moriarity (1968) discuss a factor termed 'field articulation', an index of a subjects ability to attend selectively to cues. Although the Witkin approach does not interpret field independence as an attentional phenomenon there is clearly a considerable overlap between the two concepts. Sack and Rice (op. cit.) consider distraction to be an involuntary change in an established attentional focus. Whichever interpretation of distraction is 'correct' there is no doubt as to the importance of this variable in inspection tasks. In many situations, particularly where the inspected item may only appear for a short time, as in paced tasks, if the inspector is too readily distracted by extraneous stimuli, he may miss defects. Smith and Barany (1970) present evidence that such non-observing does in fact take place in inspection tasks: the 'shifting' variable considered by Sack and Rice (op. cit.) is related to the ability to change one's attentional focus at will. It is clearly important to be able to readily change an attentional focus, as for example when one must lay aside one task and attend to another. The inspection of discrete items requires a readiness to shift attention from item to item as they are presented. This

variable also seems worth considering as for a possible means of selecting inspectors.

In summary it is apparent that the cognitive variables considered constituting an extremely promising area of research and one which will be pursued further in subsequent sections of this study.

3.2.2 Theoretical approaches to perceptual learning

Virtually all of the applied studies reviewed in section 3.1.6 have employed some form of KR as the basic training paradigm. This is partly because of the success of the KR approach in motor learning and the corresponding assumption that it is the most efficacious approach for perceptual skill training. There is still however some debate whether there may be other, equally efficient methods of training for these skills.

Most of the work that has been done has not used tasks which resemble the typical inspection situation. However, it is useful to consider in detail the approaches of two particular workers who have made substantial contributions to this field. These are Wiener, who has advocated the use of KR in these tasks, and Annett who has investigated cuing techniques in some depth.

It is useful at this point to define in more detail the two approaches. Knowledge of results (KR) has been defined as 'knowledge which an individual or group receives relating to the outcome of a response or group of responses' (Annett 1961). Forms of KR encountered in the detection context include immediate feedback as to whether a response was

a correct detection or false alarm or the provision of all or part of this information in the form of summaries.

KR is the classical learning paradigm and is said to exert its effects by reinforcing an S-R link, by reinforcing observing responses and by maintaining alertness via its motivational effect. The Law of Effect suggests that KR, contingent on a learners response, serves to reinforce the association between stimulus and response. The subject must make a response before the information can be obtained, and a 'corrected guessing' technique is used for training.

Annett, influenced by Gibson's (1953) suggestion that perceptual learning is a distinct type of learning requiring its own formulation, and not necessarily analogous to motor learning, has made a study in depth of the technique of cuing. Cuing has been defined (Annett 1959) as the provision of stimulus information before or during a response such that the response is made more effective or more likely to occur than without such information. Annett suggests that an overt response is only necessary as a means of acquiring information otherwise impossible to obtain, e.g. the 'feel' of a control. In identification tasks the stimulus and its name can be presented together and the learner does not necessarily have to make a response to the first in order to receive the second. Annett asserts that learning in perceptual tasks can take place via a simple association principle rather than a reinforcement paradigm. In his view the repeated pairing of the stimulus pattern to be learnt and its name, leads to the building up of a template against which a stimulus to be identified is compared. This theoretical position was part of Annett's general orientation that the facilitation learning by the provision of KR was not through its reinforcement or motivating

properties, but through the information conveyed to the learner (Annett 1969). Annett has presented an impressive array of evidence to support his position, although, as we shall see, his later work was more equivocal as to the superiority of cuing as compared with KR in promoting perceptual learning.

3.2.2.1 Cuing in perceptual skills training

Annett (1966) quotes an experiment in which subjects had to estimate the number of dots present in a tachistoscopically exposed field. Both cuing and KR training was superior to a simple practice control group. In a Landolt ring experiment in the same study, subjects were cued by placing the targets and non targets on different coloured backgrounds. The other conditions were a control and an 'easy' training condition in which larger gaps in the Landolt rings were used. The training effect was significantly greater for the cuing method. This suggests that simply making the task easier does not promote learning. Cuing makes the task easier without changing the critical dimensions of the stimulus. The final experiment in this series utilized a simulated sonar task in which 1000 Hz tone bursts were superimposed on a 50dB background of white noise. Four training methods were compared:

a) Cuing - a warning light was turned on half a second before each signal.

b) KR - as above, but the warning occurred after the signal, usually before the subject had responded.

c) Summary KR - a summary of hits, misses and false alarms was provided at five minute intervals.

d) Easy material - the background was reduced by 5dB so that most signals were readily detectable.

The cuing group was significantly superior to the others in enhancing performance. Annett concluded that these results were consistent with the view that perceptual learning takes place when a stimulus is unambiguously paired with its name or designation. He proposes that perceptual learning can be regarded as a simple example of paired associate learning. The failure of KR methods in some cases of training for auditory detection (Campbell (1964), Swets et al., 1964) is cited as further support for his case.

To further elucidate the relationship between cuing and KR methods of training, Annett and Clarkson (1964) conducted experiments using the same experimental set up as in the last study described. Five training groups were employed:

a) 100% cuing. A yellow light flashed half a second before each signal.

b) Retrospective cuing (non-contingent KR). A blue light flashed 2 seconds after each signal.

c) KR (contingent). Correct responses were followed immediately by a green light, incorrect responses by a red light.

d) Partial feedback cuing. The first signal was cued. Subsequently, if a signal was missed, the next signal was cued, if detected the next was uncued. False positives had no effect.

e) Partial feedback cuing and contingent KR - conditions for groups (c) and (d) combined.

According to the hypothesis that the most effective training procedure would provide the subject with as many authentic examples of the signal as possible and its distribution over the training period, it was expected that conditions (a) to (c) would produce the same rank order of effectiveness. Condition (d) was added in case cuing was effective,

but that the learner became dependent on it. It was intended as a form of conditional cue removal. Condition (e) was added when no training effect was found with condition (d). It was felt that this was because (d) gave no information on the signal distribution in time and hence partial KR was provided to enable the subject to obtain some knowledge of this through his responses.

Groups (a), (b) and (c) showed learning effects in the predicted order. Group (a) has maximum exposure to the signal plus full information about its frequency and distribution in time. Group (b) has fewer signal samples available but full information on signal frequency and distribution. In group (c) there are about the same number of signal samples available as in (b), but less distribution information. Group (e) with a proportion of authentic signals plus the opportunity to gain incomplete distributional information through contingent KR, is better than (b) or (c) but inferior to (a) which is again consistent with the hypothesis being proposed. The failure of group (d) to produce any training effect at all could be accounted for by the low proportion of signals cued and hence the reduced opportunity to accumulate the necessary experience. The cue withdrawal was initiated before any learning could be established. A notable feature of the results was that the improvements in detections found with KR were accompanied by an increase in false alarms, which was not the case with the cuing training. This is consistent with the SDT interpretation that cuing improves sensitivity but KR promotes a criterion change. In terms of the degree of vigilance decrement found, the cuing group showed a greater resistance to the decrement than the KR group. This was interpreted in terms of the expectancy theory of vigilance (Baker 1963) such that the subjects with the most accurate knowledge of the signal

distribution would show the smallest decrement.

Later experiments (Annett and Paterson 1966) using a similar experimental task provided further insights into the differing roles of cuing and KR in perceptual training. It was found that giving subjects information on the distribution of signals by flashing a warning light when a signal would have appeared, without actually presenting the signal, produced performance increases comparable to those obtained through cuing. This suggests that what is learned during cuing is not the nature of the signal itself, but some appreciation of its distribution. It should be noted however that in this task the signal, a 1800 Hz tone, had very few characteristics that one could learn. The results also suggested that the apparently lax criterion found in free response situations was at least partly due to an attempt by the subject to gain more information on the signal characteristics. With a fixed interval condition and complete KR there is much less difference in the style of performance between KR and cuing, although KR still induces a slightly lower criterion. In the third phase of Annett's study (Annett and Paterson 1967) subjects were trained in three attributes of sonar operation, i.e. pitch discrimination, intensity discrimination and duration discrimination. Cuing, KR, and a combination cuing/KR conditions were employed. All three methods were effective in training pitch and intensity discrimination but none for duration discrimination. There were no significant differences between the various training conditions. It was hypothesized that this was because a technique specifically designed to eliminate the effects of change in response bias was employed (2 alternative forced choice) and that the only difference between the various techniques is in terms of the change in response strategy produced. The most recent of Annett's work, Annett (1971), treats

cuing and KR as being essentially equivalent. This is in line with his general orientation that perceptual learning takes place purely by the pairing together in time of the stumulus and its name. If the experimental conditions are such that the subject does not have to make extra responses in order to obtain samples of the stimulus during training, then it does not make very much difference if the information is presented before or after the stimulus.

Annett's work has been discussed in some detail because of the comprehensiveness of his experimentation in this area. The issues arising from this work and possible further research that it might generate are discussed subsequently.

3.2.2.2 Knowledge of results in perceptual training

Wiener and his associates have produced a number of studies investigating the use of KR and other techniques mainly on visual vigilance type tasks. Wiener's early experiments (1963) agreed with Annett's findings that the provision of KR contingent on the subjects' responses (i.e. correct detections and false alarms) increased detections at the expense of a higher false alarm rate. Improvements due to full KR persisted after the KR was removed. A later study (Wiener 1968a) showed that detections continued to improve after 5 sessions of KR training, but that the training effect did not persist in a follow-up five weeks after the training. There was a significant increase in false alarms for the KR group over the first two sessions but not for the rest of the training, suggesting that a genuine increase in sensitivity was occurring. In these experiments complete KR was being given and hence there was no necessity for the subject to produce a high rate of responding to gain information.

An earlier paper (Wiener 1967) had shown that a group trained with KR on a visual meter monitoring task showed a significant improvement, compared with a control group, when transferred to a different type of visual monitoring task. There was no change in false alarms. This result tends to disconfirm the Annett hypothesis that perceptual learning consists of building up a 'template' of signal characteristics, since this would suggest that training would not readily transfer to tasks with different signal types. The results can, however, be accounted for quite readily by Annett's other proposal, that learning the signal distribution is almost as important as learning its characteristics. In Wiener's experiment the signal distribution (in terms of intersignal interval) was virtually the same for both tasks. Two other explanations were put forward by Wiener. One is that KR increases the subject's motivation to perform at a high level, and that this carries over to the transfer session, and the other that KR enables the subject to gain a general skill at maintaining vigilance in some unspecified way. In view of the fact that a vigilance decrement occurred in the transfer session, the latter two explanations seem less likely than Annett's proposal. In a direct test comparing cuing, KR, KR+cuing and a control, significant training effects were obtained with the KR, and KR+cuing groups but not with the cuing only group (Wiener and Attwood 1968). There was no significant change in false alarms on transfer with the groups receiving KR, but the cuing group made significantly fewer.

In SDT terms the results could be accounted for by the more cautious criterion induced by cuing, indicated by the lower false alarm rate in the transfer session. It is known (Broadbent and Gregory 1963) that during a vigilance task an increasing stringency of criterion is observed. This, combined with the already high criterion induced by

the cuing, could account for the observed absence of training effect in this experiment. An excessively high criterion would of course depress both false alarms and correct detections. The criterion effect would presumably exert a greater effect on detections than any learning associated with cuing, because of the simple nature of the signal in this experiment. A similar lack of efficacy of cuing and superiority of KR was reported for an auditory task in Annett and Paterson (1966). It was suggested in this case that the protracted nature of the training may have lead the subjects to become bored with the passive cuing training. However, the results can equally well be accounted for by the SDT hypothesis proposed earlier.

3.2.3 Some conclusions on KR versus cuing

We have seen from the analysis of Annett's and Wiener's work that their basic orientation to the question of training for perceptual skills is different, in spite of the fact that more recently Annett has been prepared to allow that cuing and KR may be largely equivalent. Annett's basic assumption is that perceptual learning is based on a simple association principle, and that learning will take place via a simple contiguity in time of a stimulus and its name. It follows therefore that both KR and cuing are important from the standpoint of the information they provide as to the characteristics of the stimulus and its distribution in time. Wiener's position is that the Law of Effect is the appropriate training paradigm and that it is primarily via the motivational effect of KR that the subject learns to maintain a higher level of arousal in the transfer session. Any knowledge of the signal characteristics that the subject learns through cuing or KR is by way of a bonus. Wiener also suggests that there is a danger that subjects may

become 'cue-dependent' and be unable to transfer any training gained to subsequent sessions.

The differing standpoints and perhaps the results of the workers in this field can be seen to be at least partly due to the differing types of task that they are interested in. Annett has not been concerned with vigilance phenomena as such, and although the signals he has employed have generally been fairly simple, the long term aim seems to have been towards the accurate identification of complex, near threshold stimuli, rather than reducing the vigilance decrement. The tasks employed by Wiener have, on the other hand, been largely visual, above threshold and extremely simple. The meter monitoring task was in fact chosen because it was known to produce a vigilance decrement. It is not surprising therefore that Wiener has been more interested in the motivational effects of KR. Although he has not stated as such, Wiener's orientation can be regarded as being concerned with enhancing signal detection through arousal mechanisms mediated via motivational variables. In this way, performance can be regarded as being improved both through an improved signal detection ability and a greater resistance to vigilance decrement. Annett would probably see the main gaol of training as improved signal detection ability, and any resistance to vigilance decrement as a bonus to be gained through a more accurate apprehension of the signal distribution. Lau (1966) and Wiener and Attwood (op. cit.) suggest that cuing may be more effective as the perceptual complexity of the task increases, whilst Annett and Paterson (1966) propose that KR may be judiciously mixed with cuing to provide a more interesting and therefore effective training regime for subjects.

In spite of the Wiener and Attwood experiment which showed no particular advantage for mixed KR and cuing training, several studies e.g. Weiz and McElroy (1964), Swets et al., (1962) and Swets et al., (1964) have shown this form of training to be advantageous.

It is clear that in spite of the considerable amount of research that has been expended on the question of KR versus cuing in perceptual training, the issue is still very much open. A number of omissions in the research to date can be identified. The most important of these is the absence of signals representative of real life tasks. In the context of this thesis, it would be of interest to apply some of the findings discussed in this section to the complex stimuli encountered in inspection tasks. Another important development would be to attempt to apply SDT in a more rigid way to the issues discussed up to now. Many of the changes in performance would be clarified by the application of SDT methods to isolate changes in sensitivity from those of response bias. This work would provide a useful link with the SDT approach to inspection developed in Chapters 1 and 2. A particular area of interest is the question of the development of a knowledge of the distribution of defects using KR or cuing techniques. The ability of an inspector to be sensitive to a change in defect distribution is important, since as discussed in Chapter 2, he can only employ an optimal criterion if his subjective estimate of the defect density is in accord with reality. It would be interesting to investigate if any of the techniques discussed up to now would develop any general ability to recognize a change in defect density. Presumably any technique which enhanced sensitivity would provide a larger sample of defects from which the inspector could more easily infer the defect distribution in time. The use of SDT would seem to be the most effective way to investigate this problem. These

possibilities would seem to be potentially fruitful areas for further research.

3.2.4 Other factors in perceptual training

Although the approaches discussed up to now represent the most consistent and prolonged studies of training for perceptual skills, a number of other models of perceptual learning exist and a variety of other variables can be seen as important in this area.

Wallis (1963) presented an interesting model of the perceptual learning process. Learning to identify complex patterns is seen as a blend of analytic and synthetic processes. Initially the trainee analyses the pattern to be detected into cues and features such as lines and angles in a visual stimulus. Eventually a process of synthesis takes place and perception occurs as an wholistic process, a Gestalt that is detected in its entirety. In his description of a training technique, Wallis stresses the demonstration of the relevant cues embedded in the complex whole by techniques such as drawing attention to one cue at a time and using training materials in which they are readily visible. As training proceeds, this analytic approach is gradually modified until an overall synthesis takes place. The main method of guidance utilized in this process is KR, and emphasis is placed on the importance of using realistic materials in training. It is pointed out that augmented feedback i.e. KR and cues, must be withdrawn at an appropriate time so that trainees do not come to depend on it for successful performance.

The problem of trainees becoming dependent on cues and being unable to successfully transfer from training has been pointed out earlier. A

similar difficulty could occur in KR situations under certain circumstances. Abrams and Cook (1971) propose that identification skills involve the development of internal references. If KR is provided continuously, the learner will utilize it to sustain performance at the expense of learning. Their experiments indicate that fading KR throughout the training session enhances the retention of identification skills. It is suggested that the removal of KR by fading creates the need for learning, and the continued, but reducing provision of KR the necessary information. Another finding which confirms the ideas of Wallis is that learning is enhanced by a gradual increase in the stimulus complexity during the training programme. Caution may be necessary in applying these results to inspection training however, since the stimuli were complex auditory signals.

Another issue arising from Wallis' analysis of perceptual learning is the relative efficiency of analytic compared with synthetic training techniques, otherwise known as part and whole methods. Annett (1971) reviewed a number of studies considering this variable. He conducted experiments comparing a large number of different analytic and synthetic methods. The overall result was that simple whole methods are as effective as any of the more complex part methods which attempt to draw attention to identifying features of complex stimuli.

3.3 Directions for research

The specific research proposals which emerge from the literature reviews of this and the preceding chapter can be seen to have three main themes. The first of these is the utilization of SDT in a sophisticated form in the analysis of real life and realistically simulated inspection tasks.

Secondly it is proposed to use SDT as a tool in the investigation of training techniques for the perceptual skills important in inspection. Finally the important but neglected area of selection will be considered from the standpoint of the cognitive skills necessary to perform inspection.

Prior to these more theoretically orientated areas, a description and analysis of the real life inspection systems which will form study vehicles for this thesis will be given and the results of on-line experimentation in an industrial context described.

CHAPTER 4 CASE STUDY I : THE INSPECTION OF BUBBLE CHAMBER PHOTOGRAPHS

4.0 INTRODUCTION

In this chapter a case study will be presented in which many of the theoretical topics discussed in the review chapter will be examined from the point of view of their applicability in a real inspection situation. The results of this case study and that considered in the next chapter will provide a useful orientation for the theoretical experimental work considered later.

Although the analysis of bubble chamber photographs may seem to be a somewhat specialized area, it will be demonstrated in this chapter that this task is directly comparable to inspection tasks found in industry.

The inspection system considered, which is in the Physics Department, University of Birmingham, has been described in a previous report (Embrey 1970). The work set out in this thesis is, however, previously unpublished.

4.1 General considerations

High energy nuclear physics research is carried out in many centres throughout the world. One of the most important research activities that these groups perform is the investigation of the structure of matter using large and very expensive particle accelerators such as the machine at Cern in Geneva. Beams of high energy particles produced by such machines are fired into devices known as bubble chambers, which consist of large containers filled with liquefied gas. The particles ionize the gas to give distinctive configurations of tracks, some of which may have been produced by new particles created as a result of

collisions between the incident particle beam and the gas atoms. It is the detection and analysis of these patterns which constitutes the bulk of the research effort in this field. The volume of data is such that it is not possible for experienced physicists to examine all data produced. Each time a beam of particles is fired into the bubble chamber, automatic cameras photograph the resulting tracks, giving rise to perhaps a million photographs from a particular experiment, each of which needs to be examined.

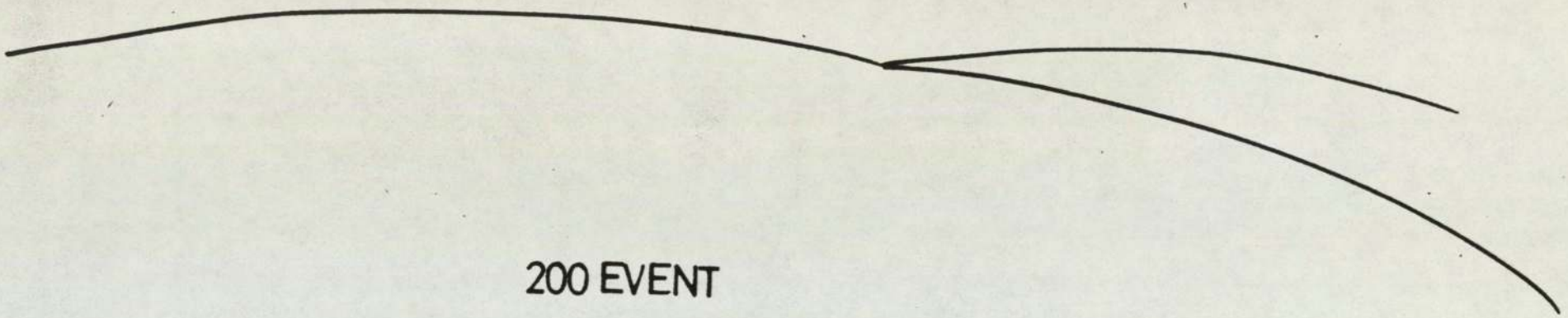
In order to cope with this inspection problem, data analysis groups have been set up in which the films are scanned and particular configurations of tracks, known as events, are detected and categorized.

The importance of the film examiner (or 'scanner') is not to be underrated. High energy nuclear physics research consumes a significant proportion of the funds available for scientific research in this country. In spite of the considerable expenditure on hardware it is salutory to note that the inspection efficiency of the unaided human operator is a vital link in the chain of analysis. For this reason the film scanning task constitutes a useful and valid area of study in its own right in addition to its implications for industrial inspection.

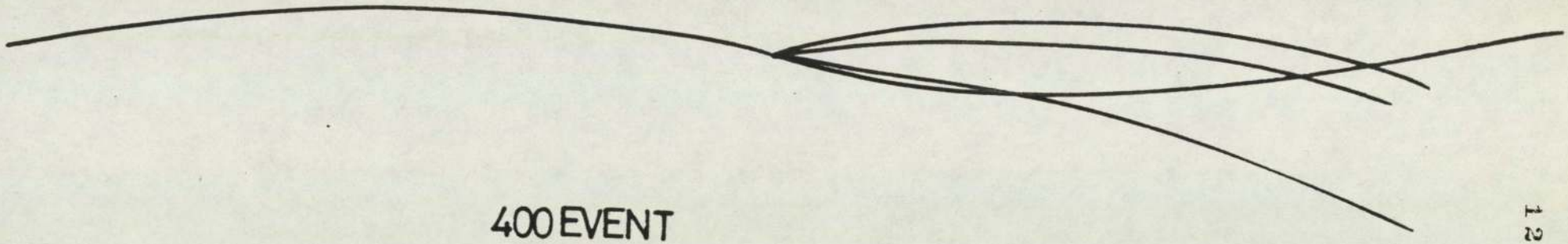
4.2 The scanning task

The items to be examined consist of photographs of bubble chamber events taken from three different angles, giving rise to three separate rolls of film, referred to as views 1, 2 and 3. Each roll consists of 750 frames. The appearance of a typical film frame can be seen from the enlarged photographs shown in Figure 4.2. The series of parallel curved

FIGURE 4.I DIAGRAMMATIC REPRESENTATION OF BUBBLE CHAMBER EVENTS.



200 EVENT



400 EVENT

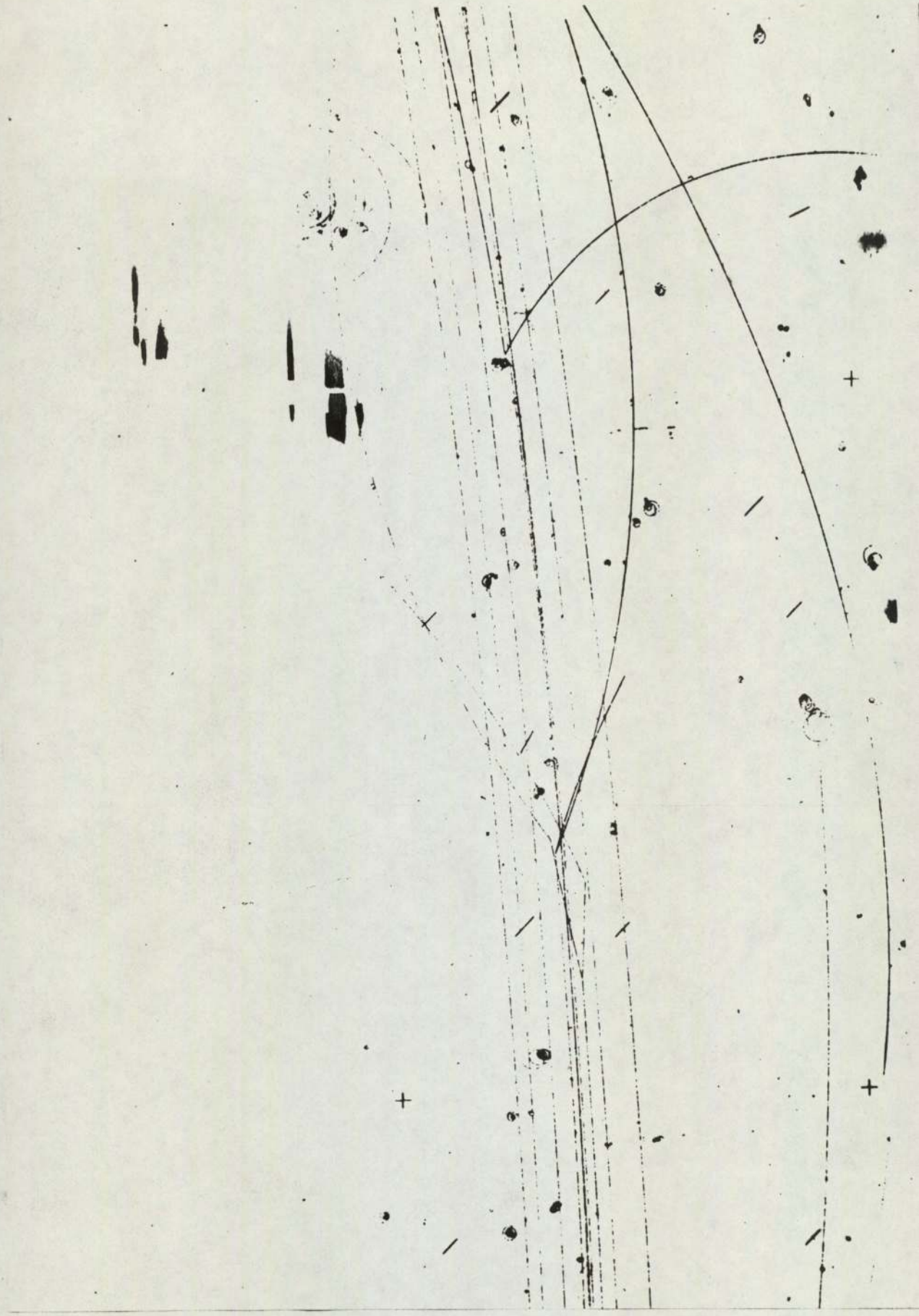


FIGURE 4.2 BUBBLE CHAMBER EVENT.

tracks entering at the bottom of the frame are known as beam tracks. As can be seen from the figures, most of the beam tracks pass through the bubble chamber without interacting. In the centre of the frame a typical 'event' can be seen where a number of prongs emanate from a 'production vertex'. The configuration of tracks which make up an event, known as its 'topology' can be described by means of a three number code. For the purpose of this study, it will be necessary to consider only two classes of event topology, those in which two or four prongs emanate from the production vertex, known as 200's and 400's respectively.

It is configurations of tracks similar to those illustrated which are of special interest to physicists. The task of the scanner consists of examining each film frame for meaningful patterns, in accordance with pre-assigned criteria. In many cases the scanners will be looking for a variety of different types of event on each frame although in some cases only one configuration is searched for. The film frame often contains a large number of unwanted patterns as can be seen in Figure 4.2. These tend to obscure the wanted configurations and can be regarded as visual 'noise'. The similarity with industrial inspection will be apparent.

The 'Shiva' scanning machines used to examine the film are illustrated in Figure 4.3. They consist of two scanning tables mounted side by side. Three rolls of film are loaded into a film transport mechanism at the side of the machine and the film images are then projected via an over-head mirror system on to the scanning table surface. The optical system contained in the scanning machine produces a magnification of 30 times to give a projected image of the same size as the scanning table i.e. 2.5 by 0.9 metres. The movement of the three views is controlled by the three switches to the left of the table surface. These enable the

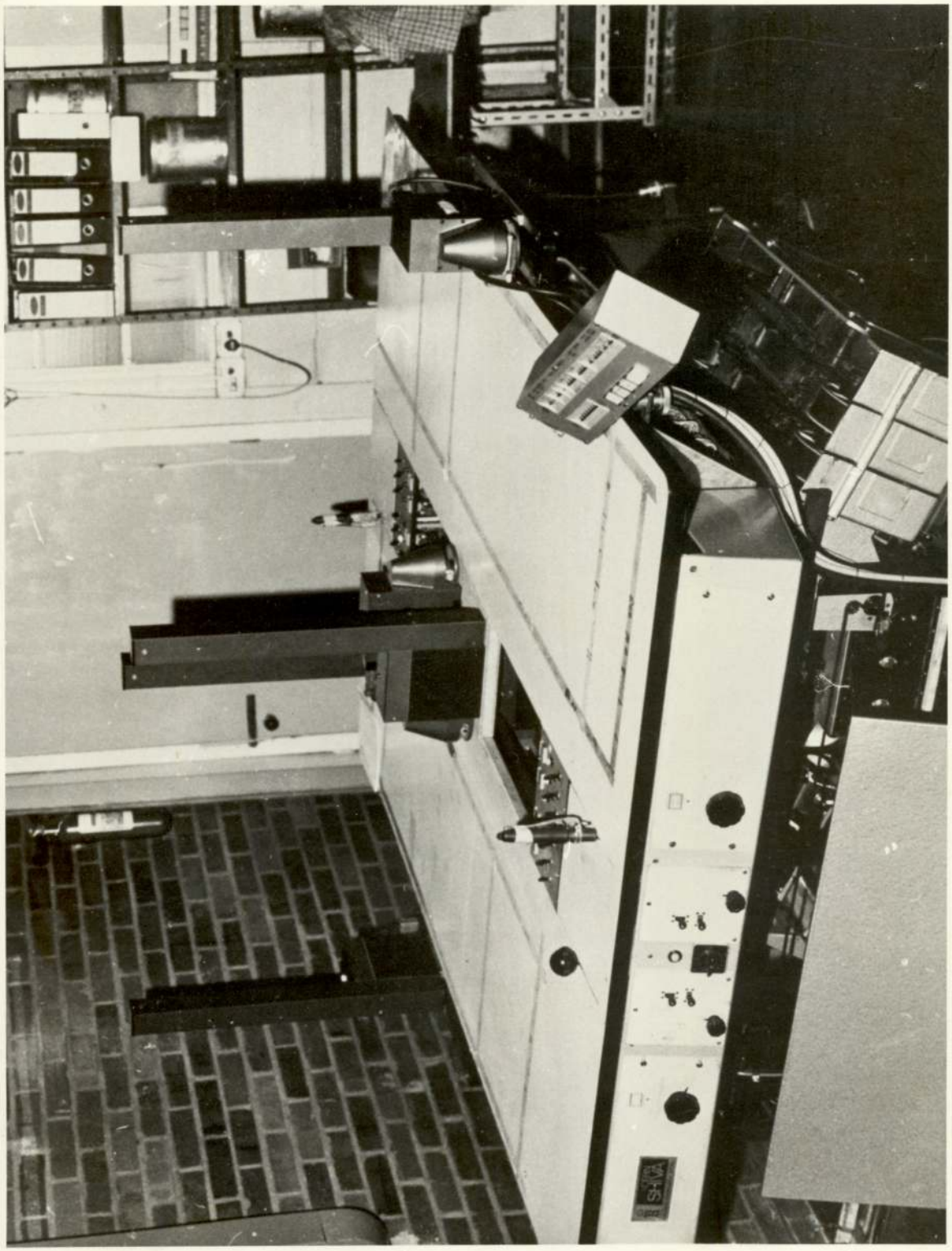


FIGURE 4.3 SCANNING TABLES

operator to advance any of the film views independently and to superimpose views if necessary. The whole optical system can be moved by the handle which can be seen to the left of the operator's position in Figure 4.2. This enables any part of the projected image to be moved close to the scanner for more detailed examination. Most scanners use this system extensively to scan the various parts of the display in which a suspected event lies.

4.3 Detailed task description

The scanner is required to examine every frame of the film specified, using at least two film views, and to find and record all the events on the film that satisfy the criteria detailed in the scanning instructions.

There is no detailed procedure laid down for carrying out the actual operations of scanning but the following is typical. The scanner advances view 1 of a particular film frame so that it is in a standard position on the scanning table. This is done by operating the view 1 film advance switch until a fiducial mark on the film coincides with a mark drawn on the scanning table. The operator then scans the projected image for wanted events, using the film transport handle to move parts of the field closer to the end of the table as necessary. Having scanned view 1, the scanner then changes to view 2, advances the film to the same frame as view 1 and then checks his findings from the first scan. This procedure is necessary because the appearance of an event can differ substantially from view to view, since each view is a two dimensional representation of a three dimensional space. Complex events are usually resolved unambiguously by use of the third view.

Having decided upon an interpretation, the scanner then feeds the information into a keyboard attached to the scanning table, from which it is transferred to on-line computer storage. Finally the scanner switches back to view 1, advances to the next frame and then begins the cycle of operations again. Each film is scanned twice, and lists compiled of any differences between the two scans. Finally the film is scanned a third time by a 'fine scanner', a more highly trained scanner who utilizes the comparison list and his own judgement to resolve ambiguities between the two previous scans. The 'fine scanner' is the final arbiter who decides which events are recorded.

4.4 Analysis of scanning as an inspection task

In the analysis of the data analysis group referred to earlier (Embrey op.cit.), a number of the variables important in inspection systems, e.g. selection, training and environmental aspects were discussed. The detailed consideration of these and other variables in the earlier review chapters enables a more complete analysis of the system to be produced.

4.5 Theoretical areas relevant to film scanning

4.5.1 Signal detection theory

SDT is clearly applicable to the scanning task in that it involves distinguishing the signal, in the form of the wanted event, from the 'noise' of the non-signal patterns in which it is embedded. It is of interest to consider if the equal or unequal variance model is likely to be appropriate in this task. If a population of experienced

scanners is being considered, they might be expected to know the characteristics of both signal and noise equally well and therefore the equal variance model would be more likely to apply than in laboratory based studies (Taylor op.cit.).

4.5.1.1 Factors affecting the criterion

Another consequence of a well practised subject population is that the scanners would be likely to be relatively homogeneous with respect to the criterion they adopt. This would be partly on the basis of experience per se, in that over a long period of time, there is considerable opportunity for feedback, informal or otherwise, to stabilize the criterion. Additionally the implied payoffs for correctly detecting an event or making a false alarm do not change substantially over time. The fact that all the events discovered by a scanner are subsequently checked by a fine scanner encourages the use of a lax, low criterion. On the other hand the low average probability of an event occurring (about 0.2 for the events considered in this study) would tend to raise the criterion. The way in which the criterion is affected by the combination of a priori signal probabilities and payoffs of the various types of division possible, was set out in chapter 2, page **27** . The probabilities of the various types of event that are scanned for is approximately constant over time, which also tends to give rise to a relatively constant overall degree of bias. Predicting the numerical value of beta is difficult because we do not have any quantitative estimates of the payoffs involved. However, by using the value of beta obtained experimentally, it should be possible to determine the subjective utilities employed by the scanners, assuming that the equal variance assumption model is appropriate. Alternatively, by assuming

some values for the costs and values of false alarms and correct detections, it would be possible to establish if the scanners were employing the theoretically optimum criterion. These questions will be considered during the experimental analysis.

4.5.1.2 Sensitivity considerations

In view of the relatively homogeneous criterion predicted in the previous section, performance variability would be primarily due to differences in sensitivity, either because of differences in sensory skills such as visual acuity, or intrinsic differences in the detectability of different event topologies. A fairly high overall value of d' would be expected, since most events are readily detectable by experienced scanners, given sufficient time to examine the film frame in detail.

There is a distribution of event discriminability across films which is a function of the contrast level, the number of beam tracks and the presence of extraneous tracks which are confusable with events. Within a particular film, different events of the same topology will vary in discriminability depending on, for example, whether the production vertex is obscured by an overlapping track, or the acuteness of the angles between the event 'prongs' and the beam track. Event prongs which have a very narrow angle are easily confused with the beam track. The number of 'awkward' events on a film is relatively small, but they give rise to protracted response latencies, whilst the scanner makes a very careful examination of the frame in order to resolve the ambiguity.

4.5.2 Visual search

It would be expected that much of the visual search data would apply to this task, as there is clearly a considerable search involved to cover the large area of the scanning table. However, the search is certainly not free search over the whole area, since the nature of the task constrains the scanner's attention to specific areas of the projected image. Events can only be produced from a beam track and these are concentrated in the centre of the frame. Typically the scanner will look along the curved, parallel beam tracks to see if their symmetry is broken by the oblique prongs of an event. A more general search would be carried out after the main event had been discovered, to ascertain if there were any other configurations associated with the main event.

In view of the inhomogeneous nature of the scan required, it seems unlikely that search time could be readily predicted from the techniques described by Bloomfield (1970). Because the task is self-paced, search time considerations are of less importance than in a conveyor belt situation. However, the overall length of time to completely scan a film will obviously be influenced by search variables, and so the overall throughput of the system is partly a function of this variable.

4.5.3 Vigilance aspects

It is difficult to predict a priori whether or not vigilance effects will significantly affect performance of the task under consideration. In terms of Kibler's comments cited in the last chapter (Kibler 1965), the signals encountered in scanning are more complex and (usually) of greater frequency than those found in laboratory vigilance experiments.

Also the fact that usually many categories of event are scanned for simultaneously, indicates that a more complex decision process is necessary than in vigilance experiments. Against these factors must be set the fact that many scanners subjectively find the task very boring. Certainly the author's own experience of scanning has tended to confirm this view. In theory scanning is carried out for periods of up to two hours continuously, although in practice the fact that all the scanning is carried out in one room means that there are many informal breaks when work stops for conversations with other scanners. In general it is felt by the scanners that the complexity of the task is such that it would be difficult, if not impossible, to perform whilst carrying out a conversation.

In view of these conflicting factors, it was felt to be of interest to investigate this variable experimentally.

4.6 Task characteristics

4.6.1 Pacing

The task is self-paced, which, as suggested by the review in the last chapter, should produce detection performance superior to a paced situation.

4.6.2 Enhancement of signal discriminability

Although the optical system produces a magnified image of the events, the actual inspection procedure is carried out on these magnified images, and hence no additional magnification is employed. The luminance of the scanning table surface (4 lumens) was within acceptable

limits. The only technique employed to enhance the discriminability of the defects was the informal one described in section 4.5.2 in which scanners look along the beam tracks in order to detect the oblique tracks of events.

4.6.3 Complexity

In considering how the complexity of the films being inspected affects event detection probability it is important, as pointed out in the last chapter, to consider the dimensions along which complexity is to be measured. In scanning, the complexity of the film can be regarded as the number of non-signal configurations likely to be present on a given film. Although there are occasional films in which there are a large number of beam tracks, which produce a high incidence of unwanted configurations, steps are usually taken to ensure that the beam tracks are widely spaced and are relatively few in number.

Another aspect of complexity that needs to be considered is that of the events themselves. Whether a four pronged event is more discriminable than say, a two pronged event, remains an empirical question. On commonsense grounds, one would expect differences, since the four-pronged event (known as a 400 in scanning terminology), has more of the event characteristics (oblique tracks) than a two pronged 200. In fact many scanning errors consist of misidentifications of events rather than missing them completely. In most cases it is necessary to ensure that both views are used to identify an event.

4.6.4 Signal rate

Because scanning is a self-paced inspection task, the signal rate in time will be determined by the average time that a scanner takes to scan a frame, and the overall probability that an event is present on a particular frame. Work by Colquhoun (1961), suggests that it is the probability of an event occurring given that a frame is presented that would determine detection efficiency, rather than the total number of frames scanned. Therefore one would not expect 'fast' scanners to detect a greater proportion of the events than 'slow' scanners, if one is considering only the event incidence in time. In general the probability of an event remains approximately constant for a given event type. We would not therefore, expect to encounter problems associated with the inspector being unable to modify his response strategy to take into account a sudden change in event probability, as was discussed in the last chapter. However, where a scanner was required to inspect for a different event type, with a different probability of occurrence than he had been used to, performance could be sub-optimal because of an inappropriate criterion.

The commonly occurring types of event have a probability of about 0.2. Usually, the scanner will be scanning for a range of events with probabilities varying from 0.2 to approximately 10^{-6}. These latter very rare events have an extremely high 'payoff' associated with them in terms of their interest to physicists, so that the criterion associated with them would not be as strict as might be expected from consideration of their probability alone. In the actual scanning task then, the inspector would be utilizing a range of criteria for the different event types.

4.6.5 Number of inspectors

Although there is only one inspector to each scanning table, the presence of the other inspector at the opposite end of the scanning machine, using the other scanning table, should enhance performance through the social facilitation noted in the studies reviewed in the last chapter.

4.6.6 Repeated inspection

All films that pass through the data analysis group receive two independant inspections from different scanners. Subsequently, comparison lists are compiled of differences between the two scans, and the film is then scanned for a third time by an experienced 'fine scanner', who utilizes the lists to resolve the differences between scans, as described earlier. Therefore the repeated inspection studies described in the last chapter are relevant here.

4.6.7 Environmental conditions

The environmental conditions are described fully in Embrey (op.cit.). The main problems are the poor ventilation and heating in the scanning room, and the bursts of high intensity noise from measuring machines in the same room.

4.6.8 Organizational and social factors

The organizational details of the system have been described in Embrey (op.cit). Considering social factors within the group, there tends to be friction between the old established workers, and the students who are

employed during vacations. The latter often quickly learn to perform the task satisfactorily but are easily bored.

4.6.9 Selection and training

Selection is on the basis of an interview with the supervisor of the data analysis group and the departmental personnel officer. Training is informal, and consists of an introductory lecture on the organization of the group and some of the basic elements of scanning. Subsequent training is largely 'on the job', the trainee being assigned to sit with an experienced inspector during the scanning procedure. In view of the complex nature of the stimuli, the perceptual skills required for scanning usually take a long time to acquire.

4.7 Conclusions regarding scanning from an ergonomics standpoint

In view of the fact that the task is self-paced and that repeated inspection is employed one would expect the system to be highly efficient in its operational objective of detecting particle interactions on film. Selection and training appear to have received little attention however, and the badly ventilated and noisy working conditions are likely to affect performance. There seems to be a possibility that there will be performance decrement with time, but this needs to be verified experimentally. The discrimination required to detect events requires both adequate visual acuity to acquire the information and the possession of high level perceptual skills to distinguish the wanted from irrelevant patterns. The scanner can employ learnt strategies, e.g. looking along the beam tracks, to enhance the detectability of the events. Although search is employed to locate an event, it is not clear how conventional search theory might be employed to predict the

time taken to achieve this. Signal detection theory would seem to be applicable to determine the degree to which the scanners employ an optimal decision criterion and the factors which affect their sensitivity for defects.

4.8 Experimental objectives

4.8.1 Specific practical goals

The analysis of the scanning task in the last section suggests several possible research objectives of interest. However, the experimental work eventually carried out had to reflect the specific needs of the organization concerned.

In the current investigation, two factors were of particular interest to management. These were whether different types of event differed in detectability and whether the noisy conditions in the scanning room degraded performance in the inspection task. When the possibility of performance deterioration as a result of prolonged periods of scanning was discussed, it was felt that this variable should also be investigated.

The interest in establishing whether there were intrinsic differences in detectability for the different event types, stemmed from certain statistical considerations connected with high energy physics. The calculation of various physical parameters was based on the probability of occurrence of different event types as inferred from the number of events detected by the scanners. These calculations assumed that all types of event were equally easy to detect. There is little data available on the effect of defect complexity as such on detectability.

Most of the studies using complexity as a variable have considered the overall complexity of the items being inspected, rather than defect complexity itself. As suggested in section 4.6.3, we would expect the number of prongs emanating from the production vertex of an event to determine its detectability, since it is primarily these characteristics which differentiate an event from the most commonly occurring form of background noise, the beam tracks. This hypothesis needs to be tested experimentally however.

The proposal to investigate the effect of noise levels on the detectability of defects was prompted by the fact that the management were trying to decide, on cost effectiveness considerations, whether or not to scrap the older machines that were responsible for most of the noise problems. They had not previously considered the possibility that the noise levels being generated might be affecting scanning performance, and hence were interested in obtaining objective evidence for this effect.

If definite evidence of performance deterioration with prolonged periods of scanning was obtained, management were prepared to consider the possibility of rescheduling the rest pauses.

4.8.2 Theoretical goals

This study provided a useful vehicle to investigate the practical utility of the theoretical orientation discussed extensively in Chapter 2: Signal Detection Theory. It was of considerable interest to evaluate the extent to which theories of this type could be translated from laboratory situations to real world tasks. As the earlier review suggested, many of the applications of SDT to date seemed to have not

adequately checked whether the underlying assumptions of the equal variance model apply.

If SDT in some form could be applied to the experimental data, then this would facilitate the separation of bias and sensitivity effects in the detection results and allow the comparison of the criterion used by the scanners with the theoretical optimum.

The effects of noise on detection situations employing patterns as complex as those found in this study had not previously been investigated, and it was of interest both theoretically and practically to see if those effects were different from those found in simpler detection situations. It was also of interest to investigate whether the nature of the noise had any differential effect on detection performance. Most studies had employed continuous white noise as the stressor, whereas the noise in the scanning room consisted of intermittent bursts as the machines were operated.

The investigation of time related decrements on the scanning task would provide additional evidence for the generalizability or otherwise of vigilance research to real world tasks.

Finally the possibility of interactions between the variables of noise pattern complexity and time on task was thought likely to provide further insights of theoretical interest.

4.8.3 Summary of research objectives

A. <u>Practical</u>

1. To investigate whether there were intrinsic differences in

detectability between events of differing complexity.

2. To evaluate the effects of various types of auditory noise on detection performance.
3. To determine whether detection of events was affected by prolonged scanning, i.e. time on task.

B. Theoretical

1. To investigate the applicability of SDT as a usable model in a real inspection situation.
2. To consider the effects of auditory noise on various parameters of detection performance. Different types of noise were to be considered.
3. To verify or otherwise the applicability of vigilance data to the scanning task.
4. To consider the interactions of the major variables present in the study.

4.9 Experimental philosophy

As the objectives of this study were to collect data in as realistic a situation as possible, it was necessary to conduct the experiment using real films and employing the scanning machines customarily used by the inspectors. The most authentic results would have been obtained by introducing a test film into the everyday work of the scanners and subsequently scoring it for accuracy. There were a number of disadvantages to this procedure. The unofficial breaks taken and other occurrences difficult to allow for, lead to the decision to conduct the experiment using a greater degree of control.

4.10 Experimental work

4.10.1 Hardware considerations

As far as the scanning task itself was concerned, two aspects needed modification. The first of these was the tendency of some of the scanners to advance the film so that only part of it was visible on the scanning table. They would then examine the tracks by moving a separate handle which moved the optical system independently of the film. In order to obtain consistent estimates of the total time to scan a frame, it was necessary to ensure that each frame was presented at a standard position at the commencement of each scan. Ideally the frame needed to be positioned on the scanning table such that the whole of it was visible. A second problem concerned the use of the various views available. Some scanners tended to scan using one view only, and when they encountered an event which was difficult to resolve on a single view, they would wind on the second and sometimes even the third view to obtain the additional information present on the corresponding frames. Sometimes these views would be very far behind the current frame and a considerable time might elapse until they were wound to the appropriate position. This procedure would clearly adversely affect any estimates of scanning time.

Both of these problems were resolved by modifying the film advance mechanism of the scanning machines. As described in detail in Embrey (op.cit.) an electronic film advance mechanism was designed such that by depressing a button, all three views of the film advanced simultaneously. When the frame image was at the correct position on the scanning table, a photocell sensor stopped the film advance mechanism.

During the actual experiment, subjects were asked to depress one of two buttons after they had scanned a film frame, depending on whether they felt that there was an event of a specified type on the frame. Depressing one of the buttons caused an oscillator tone to be recorded on tape, and simultaneously initiated the film advance sequence. The tapes were subsequently analysed using SETAR (Welford 1952) to give an output indicating the nature of the response (i.e. event present or absent) and the elapsed time since the previous response, i.e. the response latency in a self-paced task.

4.10.2 Experimental design - general

The basic conditions to be investigated were auditory noise levels and types, differing complexities of events, and time on task. Two of the noise conditions utilized continuous white noise, one level being a masking noise condition of 65dB and the other corresponding to the noisiest conditions in the scanning room of 85dB. The third noise condition consisted of an actual recording of the highly variable noise environment in the scanning room, the average intensity of which was 85dB. The object of this condition was to investigate whether the nature of the noise, apart from its intensity, had any effect on performance. The two most commonly occurring types of event, the four pronged and two pronged topologies, were chosen as the two levels of pattern complexity, which were known to be approximately equiprobable on the film to be used. A time interval of thirty minutes was chosen for the task duration, which was divided into three periods of 10 minutes for the purpose of the performance decrement analysis. Subjects started scanning at random points on the film, subject to the proviso that there were sufficient frames available for the fastest scanner to work for 30 minutes. It was not possible to control

completely for time of day effects because different subjects belonged to shifts which started at different times. Attempts were made, however, to ensure that the various sessions took place at approximately the same time within each shift for each subject. The subjects employed were seven experienced scanners, all with at least one year's experience. They all had normal or corrected vision. There were six males and one female (subject 7).

4.10.3 Experimental design - statistical

A 3 x 2 x 3 x 7 complete factorial repeated measures design was employed, the factors being noise conditions, event complexities, time intervals and subjects respectively. This design is discussed in Kirk (1968) p.237, Myers (1966) and other standard texts. Poulton (1969) has criticized such designs on the grounds of possible carry-over effects between treatments. In the opinion of the present author however, these criticisms are really applicable primarily in laboratory studies where learning effects between trials are almost inevitable with the time available for practice in typical experiments. In industrial experimentation, where very highly practised subjects are available, as in the present study, it is felt that such effects will be minimal, and hence the repeated measures design is considered appropriate. The desire to test a number of variables and the relatively limited number of trained subjects available made the subjects X treatments design a natural choice.

All the factors in the experiment were assumed fixed. In the case of the experimental treatments, the particular levels of interest of the variables considered were all included in the experiment. The justification

for using subjects as a fixed rather than a random effect was that the conclusions drawn from this study were intended to be specific to the group under study. Additionally, scanners are a specialized group and cannot be regarded as being randomly sampled from the population. The analyses of variance were performed by a program from the IBM Scientific Subroutine package, modified extensively to produce the particular design used. Where appropriate, arcsine or log transformations were used to reduce the heterogeneity of variance of the raw data or where obvious skewness of the distribution existed.

4.10.4 Analysis of data

The basic data from the experiment consisted of yes or no (event present or absent) decisions for each frame together with the time interval from the presentation of the frame to the response by the scanner. This was obtained from SETAR as described in Embrey (op.cit.). A computer program (DATAL, Appendix D) converted this data to give the measures set out below:

1. Performance measures based on SDT theories

(a) Parametric

d'

beta

(b) Nonparametric

(I) Sensitivity

Pollack Norman index

Latency sensitivity index (Navon 1975)

(II) Bias

Hodos-Grier index

ZFA

2. Inspection performance indices

   Indices A1 and A4 (McCornack 1961)

   Correct detection probability

   False alarm probability

3. Response latency measures

   Correct rejection (i.e. correct decision that frame did not contain an event) latency

   Correct detection latency

   False alarm latency

   Missed defects latency

The inspection measures A1 and A4 are defined as below:

$$A1 = \frac{NCR + NCD}{NFA + NCR + NCD + NOM} \times 100$$

$$A4 = \frac{NCD - NFA}{NCD + NOM - NFA} \times 100$$

where NCD = no. of correct detections of signals

NFA = no. of false alarms

NOM = no. of missed signals

NCR = no. frames correctly rejected as not containing signals

The program also performed arcsine and log transforms on some of the data prior to the analysis of variance. Most of the measures set out above have been discussed in Chapters 1 and 2. Any new measures considered will be discussed in the text.

4.10.5 Treatment of zero cells in SDT analyses

I am indebted to Dr Raj Parasuraman for suggestions on this topic. In this study, as in most detection experiments, on several occasions subjects made no false alarm responses during the experimental period

being analysed. This creates difficulties for the calculation of SDT parameters because without a pair of correct detection and false alarm probabilities beta and d' cannot be calculated. When these probabilities are obtained from detection experiments, the quantity we are actually measuring is the relative frequency of a particular response category over a number of n trials, which tends to the actual probability as n becomes large. If a given time period does not contain any false alarms, for example, this does not automatically mean that commission error probability during this period is zero. It simply indicates that the probability is too small to be estimated by the use of relative frequency techniques.

Several methods are available for the estimation of false alarm probability P(S/n) during a period in which n non-signal trials occur and no false alarm responses are made.

The technique that has usually been adopted, Jerison, Pickett and Stenson (1965), Wallack and Adams (1969), is to assume that half a commission error has occurred. This is equivalent to assuming that:

$$P(S/n) = \frac{1}{2n} \qquad (1)$$

Another technique is to take a weighted average of the probabilities associated with each noise trial in the observation period, i.e.:

$$P(S/n) = \frac{1}{2} + \frac{1}{2}^2 + \frac{1}{2}^3 \ldots + \frac{1}{2}^{n-1} \; / \; 2n$$

$$= \frac{1 - 1/2^n}{2n} \qquad (2)$$

A final possibility considers the likelihood of no commission errors occurring within a period.

Probability of a commission error not occurring during a single trial = 1 - P(S/n).

Assuming a fixed probability, the probability of no commission errors occurring in n trials = $(1 - P(S/n))^n$.

The likelihood of this can be tested by setting $(1 - P(S/n))^n \leqslant \frac{1}{2}$, from which $P(S/n) \geqslant 1 - \frac{1}{2^{\frac{1}{n}}}$, where $p(S/n) = 1 - \frac{1}{2^{\frac{1}{n}}}$ (3)

For large n (>50) equation (3) gives the best estimate of P(S/n), and was therefore used by the programs where zero false alarms occurred.

4.11 Results and discussion

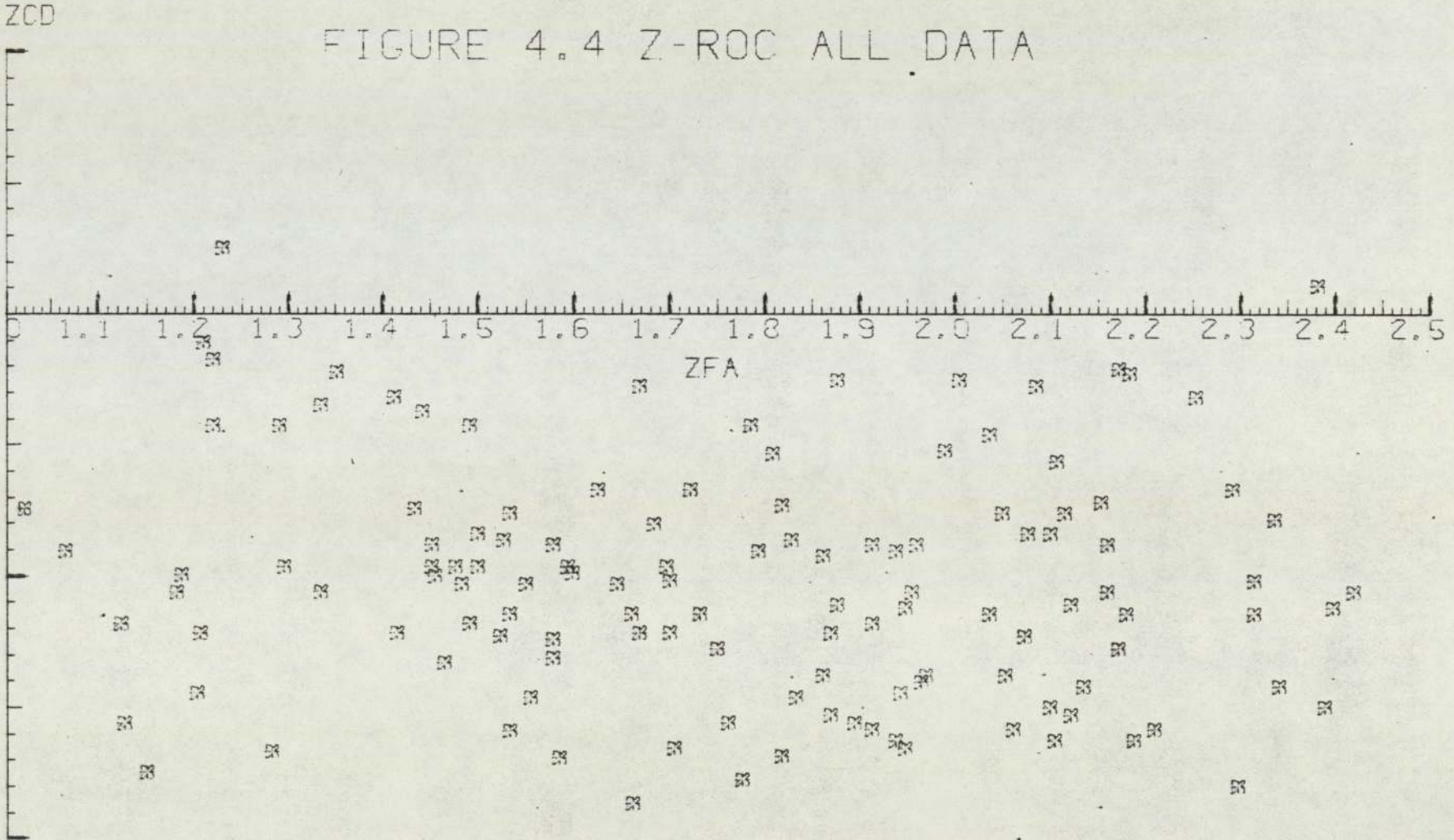
The summary data and means are given in Appendix B and the analyses of variance referred to in the text can be found in statistical Appendix A. The first group of results discussed will concentrate on the theoretical implications of the data prior to a consideration of their practical significance.

4.11.1 SDT considerations

4.11.1.1 The applicability of SDT to the data

One of the major goals of this study was to investigate the applicability of SDT in the task under consideration. The first test which can be applied is to plot the Z transforms of the false alarm and correct detection probabilities against one another to give the normalized or Z-ROC curve. Figure 4.4 shows this for all 126 data points. The data clearly do not fall on a straight line as predicted by SDT.

FIGURE 4.4 Z-ROC ALL DATA



A nonsignificant correlation confirmed this. On the other hand this result is not too surprising in that we are superimposing ZFA and ZCD values obtained across a wide variety of experimental conditions and subjects. Analysis of variance 1 for ZCD shows significant differences between subjects ($p < .001$) and time intervals ($p < 0.05$) whereas for ZFA there is a significant time interval x subject interaction. In order to obtain meaningful ROC curves it is therefore necessary to consider the ROC curves within time intervals within subjects. It was mentioned in Chapter 2 that because of the error inherent in the estimation of both false alarm and correct detection probabilities, the ROC curve should be obtained using maximum likelihood rather than least squares techniques, which assume an errorless independent variable. Considering the data within time intervals, within subjects, allows the utilization of the Grey-Morgan maximum likelihood (ML) fitting program, Grey and Morgan (1972). This program normally only accepts data from rating experiments in which a series of ascending confidence ratings have been made. Each pair of false alarm and correct detection frequencies in the data subset under consideration provides one point on the ROC curve, but if the raw frequencies were entered into the ML program without regard to order, the program would fail in attempting to fit a single straight line. A preliminary program (SIGROC, Appendix D) therefore first calculated a series of ZFA values from the false alarm frequencies. Since ZFA is monotonic with the magnitude of the sensory evidence regardless of the variances of the noise and signal + noise distribution, (Chapter 2), sorting the ZFA values into ascending order together with the associated FA and CD frequencies provided the program with the appropriate inputs. Part of the output from the program is given in Appendix A. The straight line Z-ROC fits the data in every case, as tested by chi-square. The mean ratio of the signal + noise to the noise variance is 1.362 which differs

significantly from 1 ($t = 14.47$, $p < 0.001$). Although the Grey-Morgan program, strictly speaking, was written with the rating experiment as its underlying model, its use in the present context is justified on the grounds that it is being employed purely as a device for fitting a straight line to the data using maximum likelihood techniques.

The use of the program also enables us to correct the beta values found by assuming the equal variance model, to the actual betas employed by the subjects in the unequal-variance situation. The Grey-Morgan program produced estimates of the variance ratio (the reciprocal of the slope of the fitted Z-ROC curve) for each block of time data within subjects that it was applied to. The unequal variance betas are obtained simply by multiplying the equal variance betas by the appropriate Z-ROC slope (McNicol (1972) p. 92). This was done for each of the 21 blocks of data for which the Z-ROC had been fitted.

Hence the data appears to be describable by the unequal variance SDT model. This is in accord with most laboratory studies using SDT with visual tasks, but does not agree with the earlier suggestion that the well practised subjects used in this study might be expected to know the signal characteristics as well as the non-signal attributes, and hence have equal variance internal distributions. It seems possible that because of the very low incidence of false alarms found in this study, the other cause of apparently unequal variance distributions, the greater sampling error involved in the calculation of the signal distribution variance, may be an important factor. The present finding differs from the only other published industrial study using SDT, Drury (1973), which found the equal variance model to fit the data. The main difference between the two studies was that Drury's subjects

were provided with rapid feedback during the latter part of his study, whereas no direct feedback for the scanners in this experiment was provided. However, even before feedback was provided in Drury's study, a slope of 1 for the ROC curve was obtained. Probably the main reason for the different findings in each study is in the nature of the signals in each case. The variability of glass faults is relatively small compared with the extremely wide range of configurations that are found on bubble chamber film, and this would tend to increase the variance of the signal distribution in the latter case.

4.11.1.2 Other tests of the SDT model

A number of methods are available for testing whether certain other assumptions of the SDT model hold in this situation. Ingleby (1974) points out that although the likelihood ratio criterion, beta, is the theoretically ideal measure of how much weight to attach to a sensory datum, it has never empirically been established that human subjects actually do set their criteria in terms of beta rather than, for example, the sensory evidence x itself.

It can be shown (McNicol (1972) p.64) that:

$$\log \text{beta} = d'x - d'^2/2 \qquad (1)$$

in the equal variance case, (x = ZFA, since this is monotonic with the sensory evidence used by the observer). In the unequal variance case, the expression becomes:

$$\log \text{beta} = x^2 [(\sigma_s^2 - 1)/2\sigma_s^2] + d'x - (d'^2/2) - \log \sigma_s \qquad (2)$$

If the observer is actually positioning his criterion on the basis of beta, the first implication of 1 and 2 is that there should be a linear relationship between x , i.e. ZFA, and log beta in the equal variance case, and a parabolic relationship in the unequal variance

situation. In fact, with the value of the variance ratio found in this experiment (1.3) it would be difficult to distinguish between a linear and parabolic regression with the range of probabilities which occurred. We can therefore take a good linear fit of the relationship between ZFA and log beta as reasonable evidence that a likelihood ratio criterion is being used.

The graph of log beta v. ZFA is plotted in Figure 4.10 and it can be seen that a good fit is obtained, ($r = 0.74$ $p < 0.001$).

Confirmation that the unequal variance model applies can be obtained by considering the change in ZFA for a given change in log beta at different values of d'. The parabolic relationship between log beta and ZFA of equation (2) suggests that ZFA should change less at high values of d' than at low values, with changes in log beta. We can test this by obtaining the regression equations for ZFA v. log beta for each subject. Since there are significant differences between d' for subjects (analysis of variance, Appendix A p.) a plot of the slope of the regression lines (a measure of the rate of change of ZFA v. log beta) against $^{1}/d'$ should be linear. As Figure 4.6 shows, there is a high degree of relationship with $r = 0.757$ ($p < 0.01$).

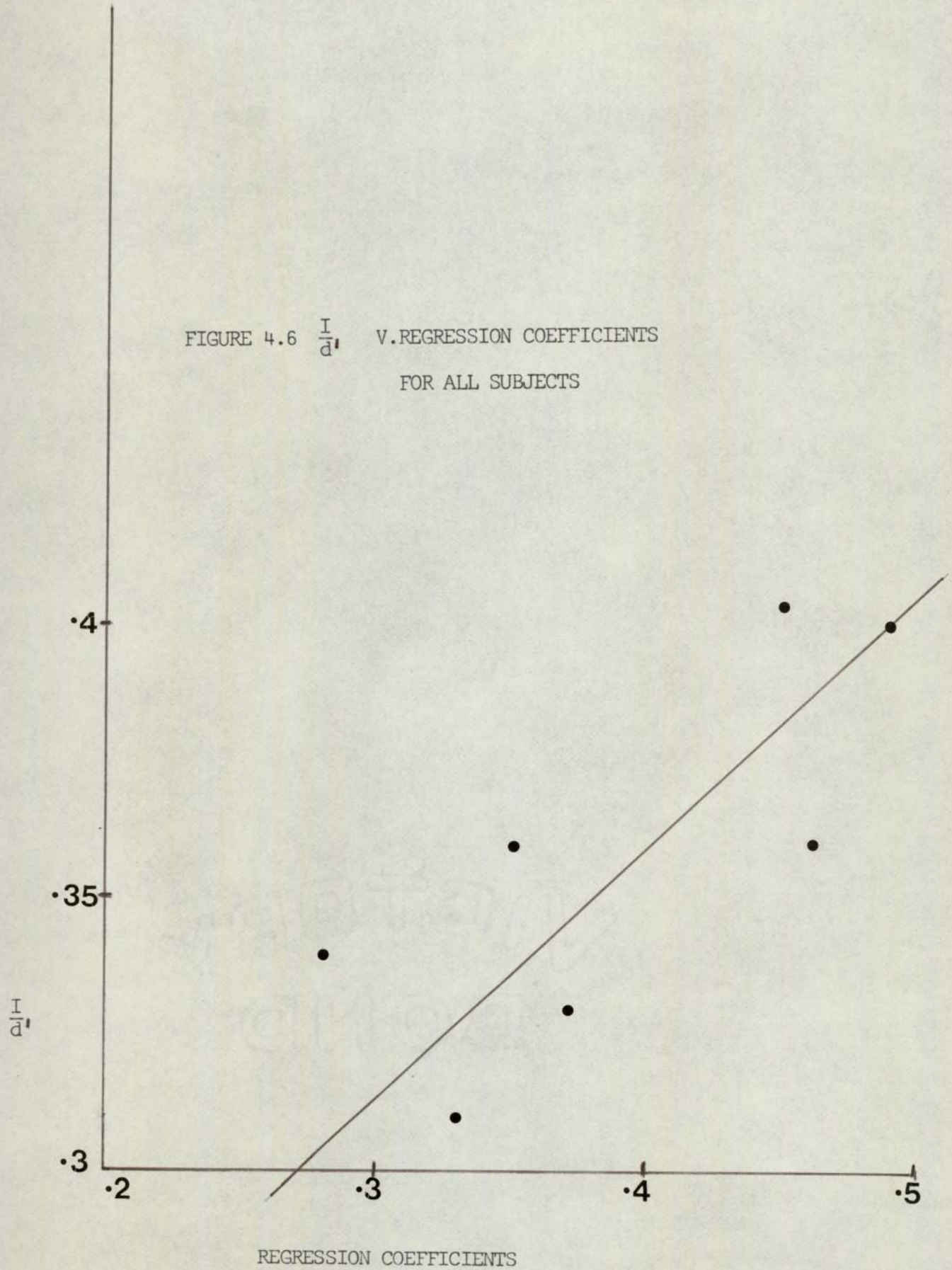
The preceding tests strongly suggest that the SDT model provides a good description of the data. The evidence is not entirely unequivocal however. The equal variance model suggests that beta should be related to the a priori probability P by the equation:

$$((1 - P) / P) \times (\text{relative cost factor}) = \text{beta} \quad (3)$$

(see Chapter 2 p. 27).

A significant correlation was only obtained between beta and $(1 - P)/P$

FIGURE 4.6 $\frac{I}{d'}$ V.REGRESSION COEFFICIENTS FOR ALL SUBJECTS



for one out of seven subjects. Although the fact that the unequal variance model does not produce a linear correlation, since the variance ratio is not very large, one might have expected a larger number of significant correlations. The result obtained is probably due to the fact that the a priori probability of the events does not vary sufficiently to produce a modification of the subjects' criteria.

One aspect of SDT predictions that has not yet been considered is the effect of the payoff matrix on the positioning of the criterion. It is clear that in the scanning task we do not have a symmetrical payoff matrix, in that missed signals are more expensive than false alarms, because the latter errors are likely to be picked up at the fine scanning stage. As mentioned in Chapter 2, the payoff matrix is given by the expression:

$$\frac{\text{value of correct rejections - cost of false alarms}}{\text{value of correct detections - cost of missed signal}} \qquad (3)$$

In order to provide an approximate check of the effects of these utilities, a survey was carried out amongst the scanners, and the mean utilities for the various decision alternative was found to be as follows:

value of correct rejections (of non-signal frames) = 4

cost of false alarms = 1

value of correct detections =10

cost of missed signal = 5

Substituting in equation (3), assuming on a priori signal probability of 0.2, the theoretical optimum value of beta obtained is 2.2. By comparison the overall mean beta obtained for the experiment is 2.71. The closeness of these figures supports the view that the inspectors are using a likelihood ratio criterion.

We can conclude, therefore, that the evidence suggests that the SDT model is an appropriate one in the inspection situation under investigation.

4.11.1.3 Relationships between performance measures

The fact that the unequal variance model is appropriate to the data makes it useful to consider the applicability of the non-parametric measures of performance in investigating the effects of the various experimental variables. It is also of interest to look at the relationship of some of the inspection performance indices such as A1 and A4 to the SDT measures.

As discussed in Chapters 1 and 2, the unique advantage of the SDT parameters is that they allow the separation of sensitivity and bias effects. They also have the advantage of resting on a solid theoretical foundation, and of providing a predictive capability. None of the non-parametric measures available for yes-no data are able to offer these advantages and we are therefore justified in judging their usefulness by the extent to which they correlate with beta and d'. In the present experiment it was possible to correct the betas obtained under the assumption of equal variances by multiplying each block of values by the slope obtained from the ROC curve for that particular block, as described earlier. This provided a baseline of 'actual' betas against which to compare the non-parametric measures. Unfortunately such a blanket procedure is not available to correct the d' values. The appropriate version of d' in the unequal variance case is Δ m or d'_e as described in Chapter 2. These can only be obtained from the ROC curve; d' values obtained under the equal variance

assumptions cannot readily be rescaled. In the comparisons that follow it was decided to use the equal variance d' as a baseline. Since the variance ratio is not large, the comparisons will retain some degree of validity.

Plots of the Pollack-Norman and latency sensitivity indices against d' are given in Figures 4.7 and 4.8. These indicate a high degree of correlation in the first case and a smaller but still significant r in the latter. These are confirmed by the product moment correlations in Table 4.1 below:

comparison	correlation	sig.
d' v. Pollack - Norman index	0.923	$p < 0.001$
d' v. Navon latency index	-0.184	$p < 0.05$

Table 4.1 Comparison of sensitivity indices

The relationship between d' and the Pollack-Norman index seems to be slightly curvilinear in nature, although assuming a linear relationship would lead to only slight errors.

The scatterplots of the corrected values of log beta against the Hodos-Grier bias index and ZFA are given in Figures 4.9 and 4.10 and Hodos-Grier v. ZFA in Figure 4.11. The corresponding correlations are given in Table 4.2

comparison	correlation	significance
log beta v. Hodos-Grier index	0.893	$p < 0.001$
log beta v. ZFA	0.743	$p < 0.001$
Hodos-Grier index v. ZFA	0.746	$p < 0.001$

Table 4.2 Comparison of bias indices

FIGURE 4.7 D' V. P-N INDEX

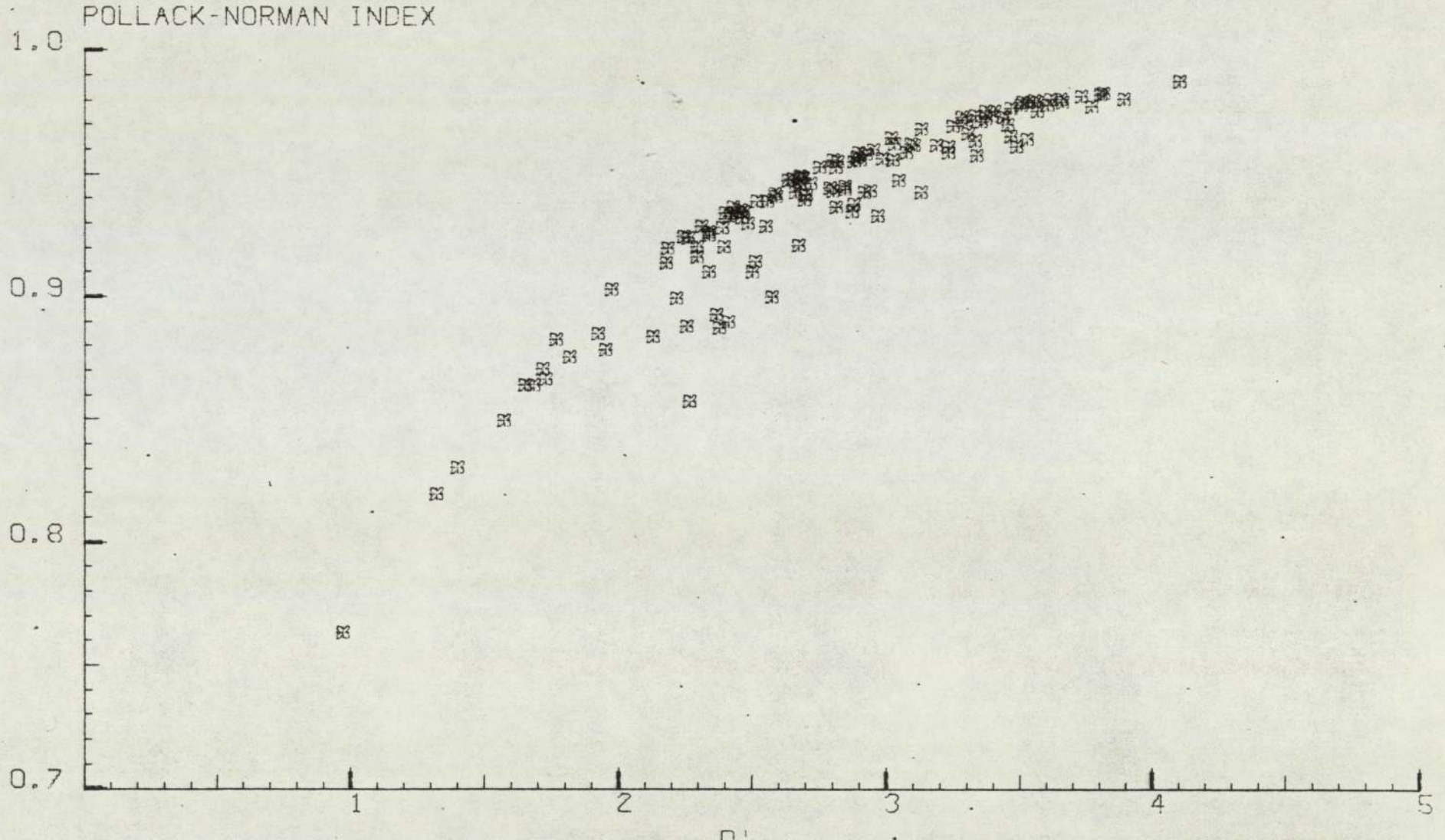


FIGURE 4.8 D' V. NAVON INDEX

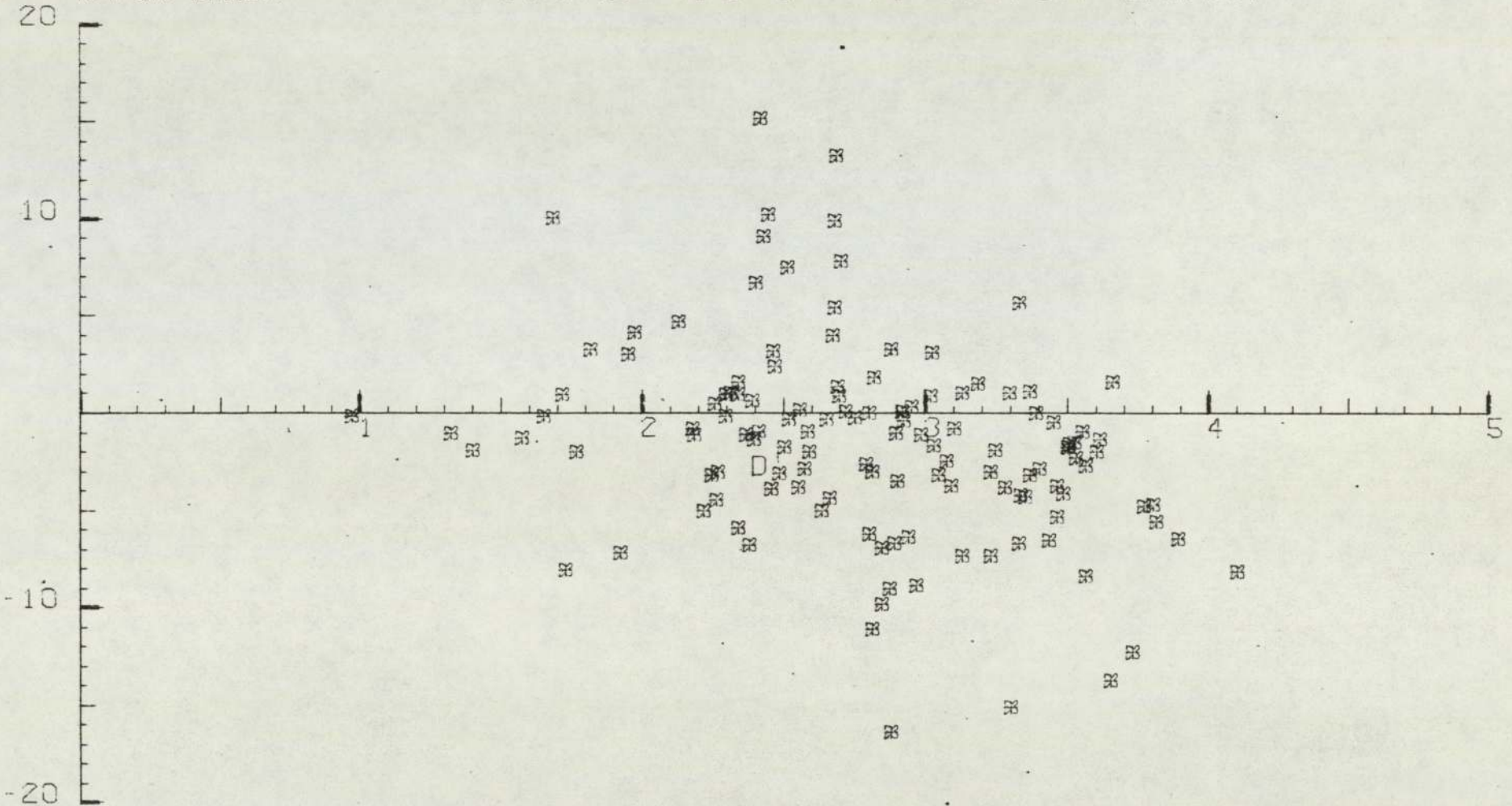


FIGURE 4.9 LOG BETA V. H-G INDEX

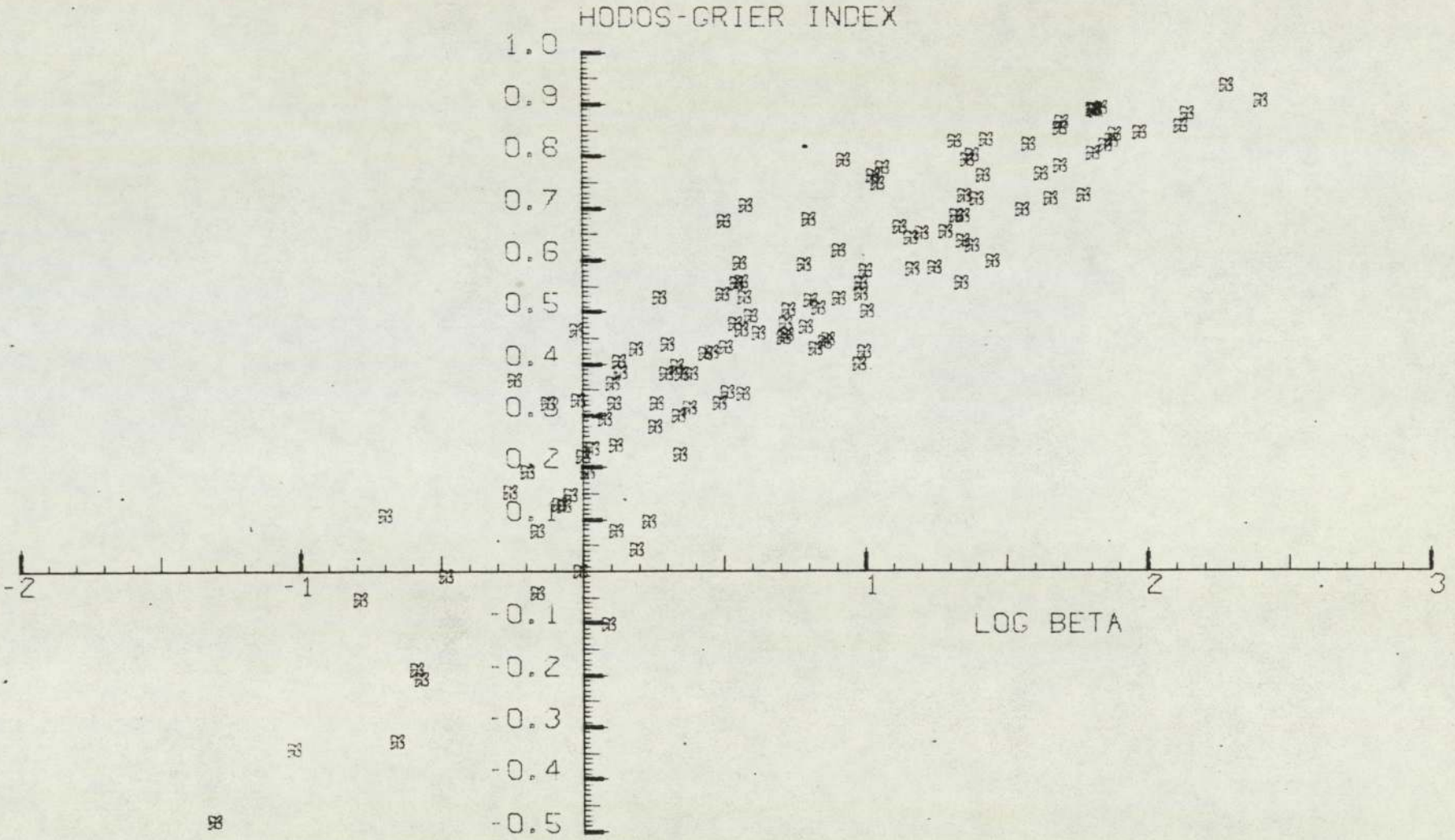


FIGURE 4.10 LOG BETA V. ZFA

FIGURE 4.11 H-G INDEX V. ZFA

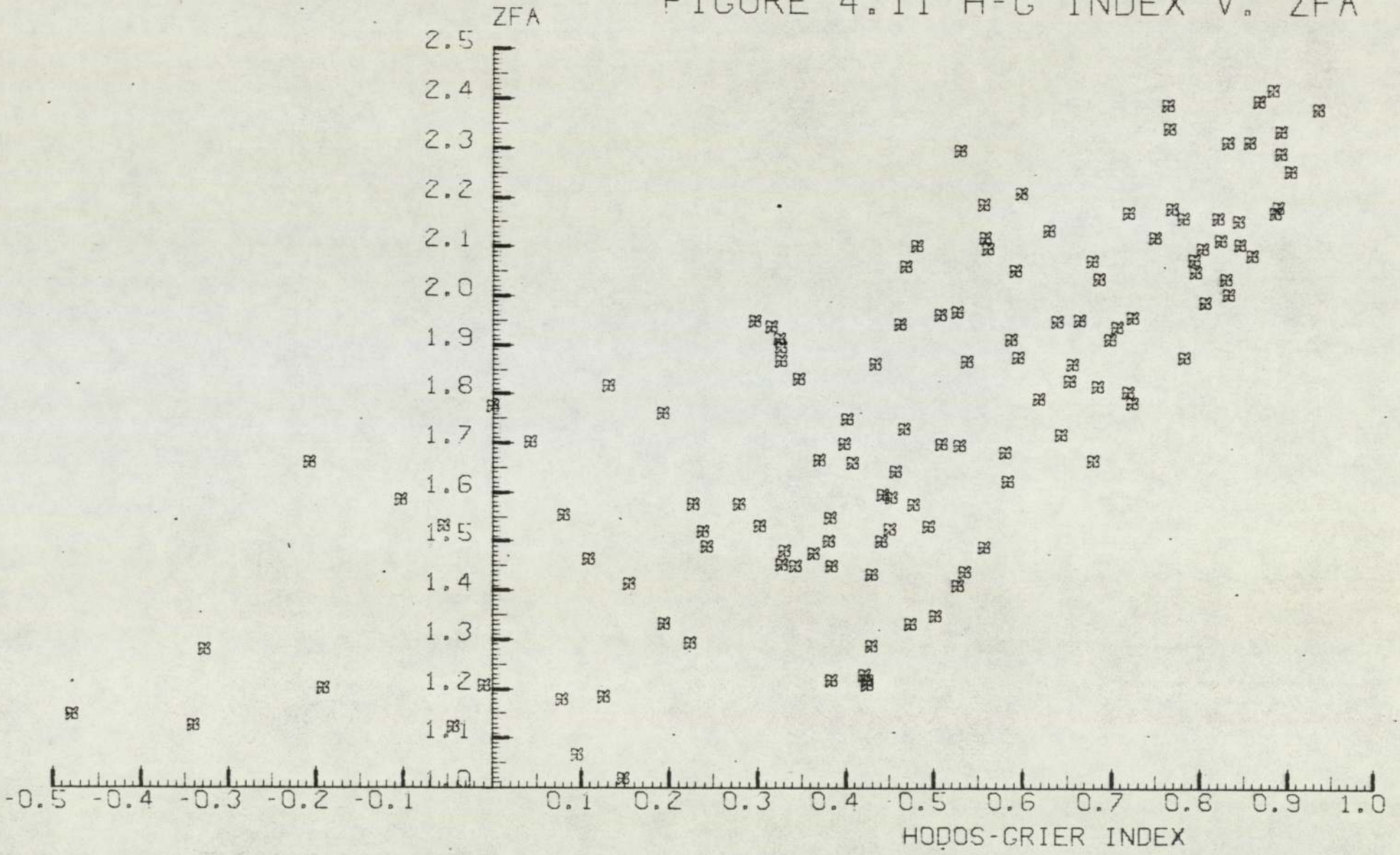
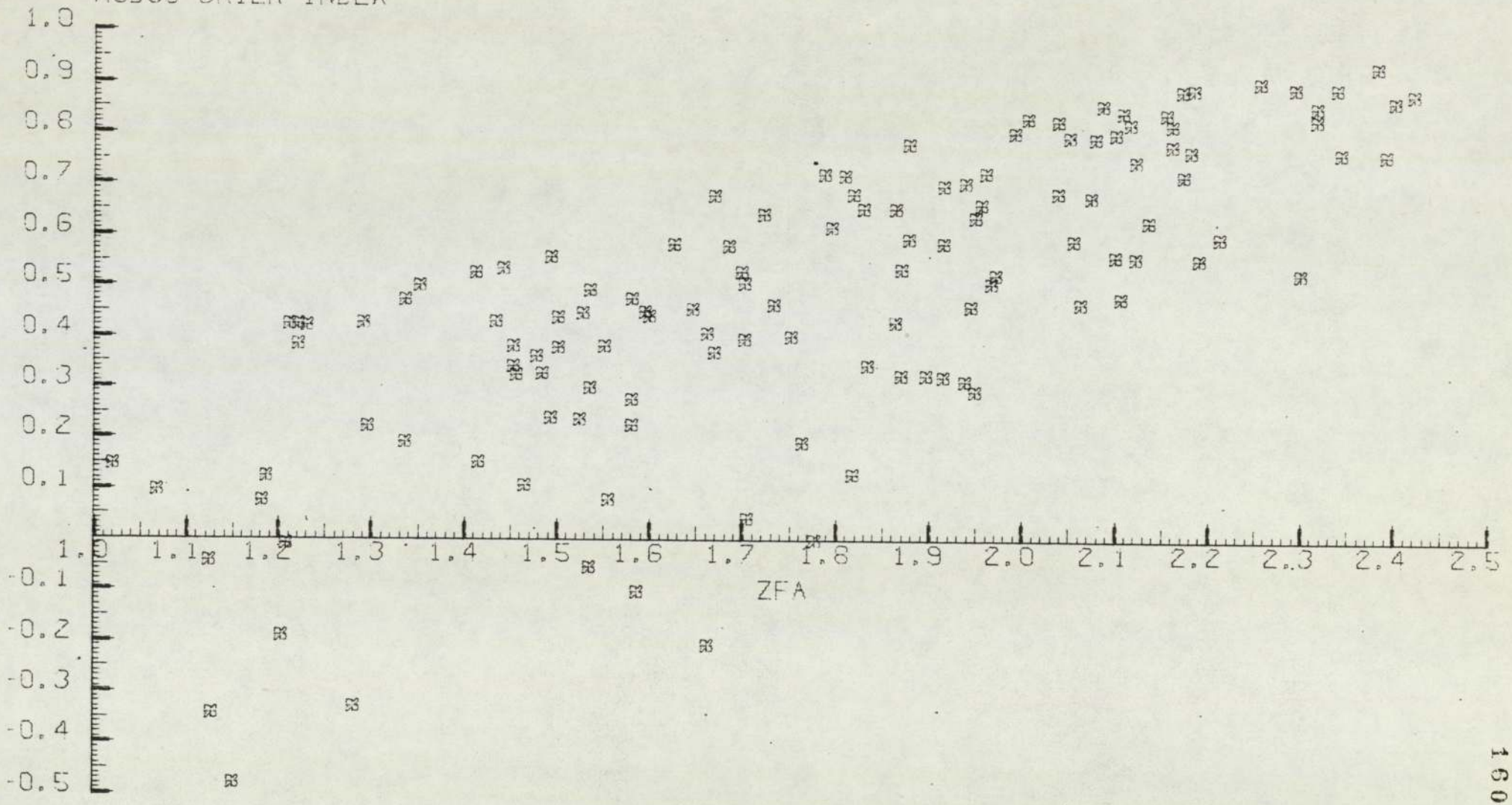


FIGURE 4.11 ZFA V. HODOS-GRIER INDEX



A1 INDEX V. A4 INDEX

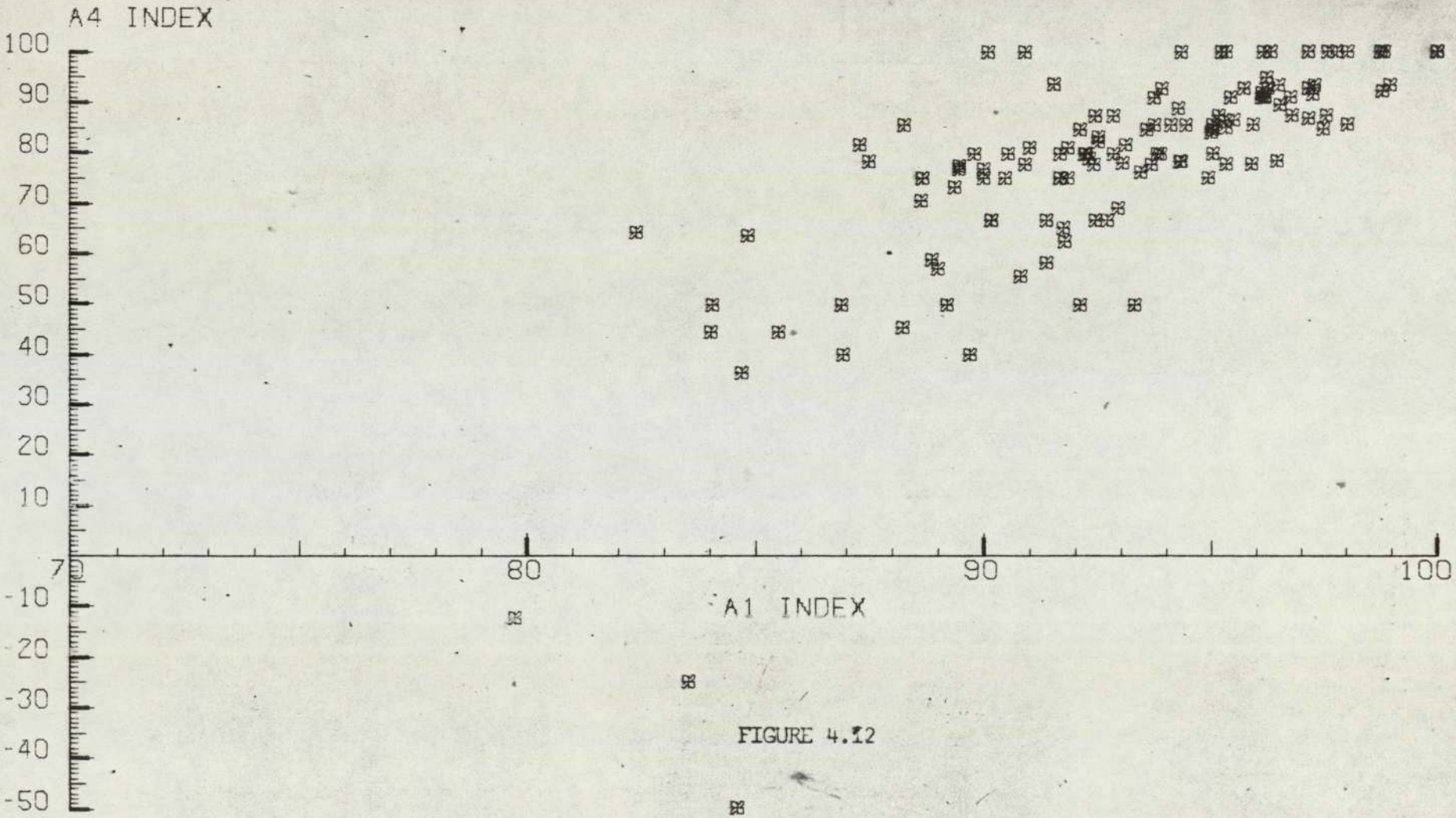


FIGURE 4.12

FIGURE 4.13 A1 INDEX V. D'

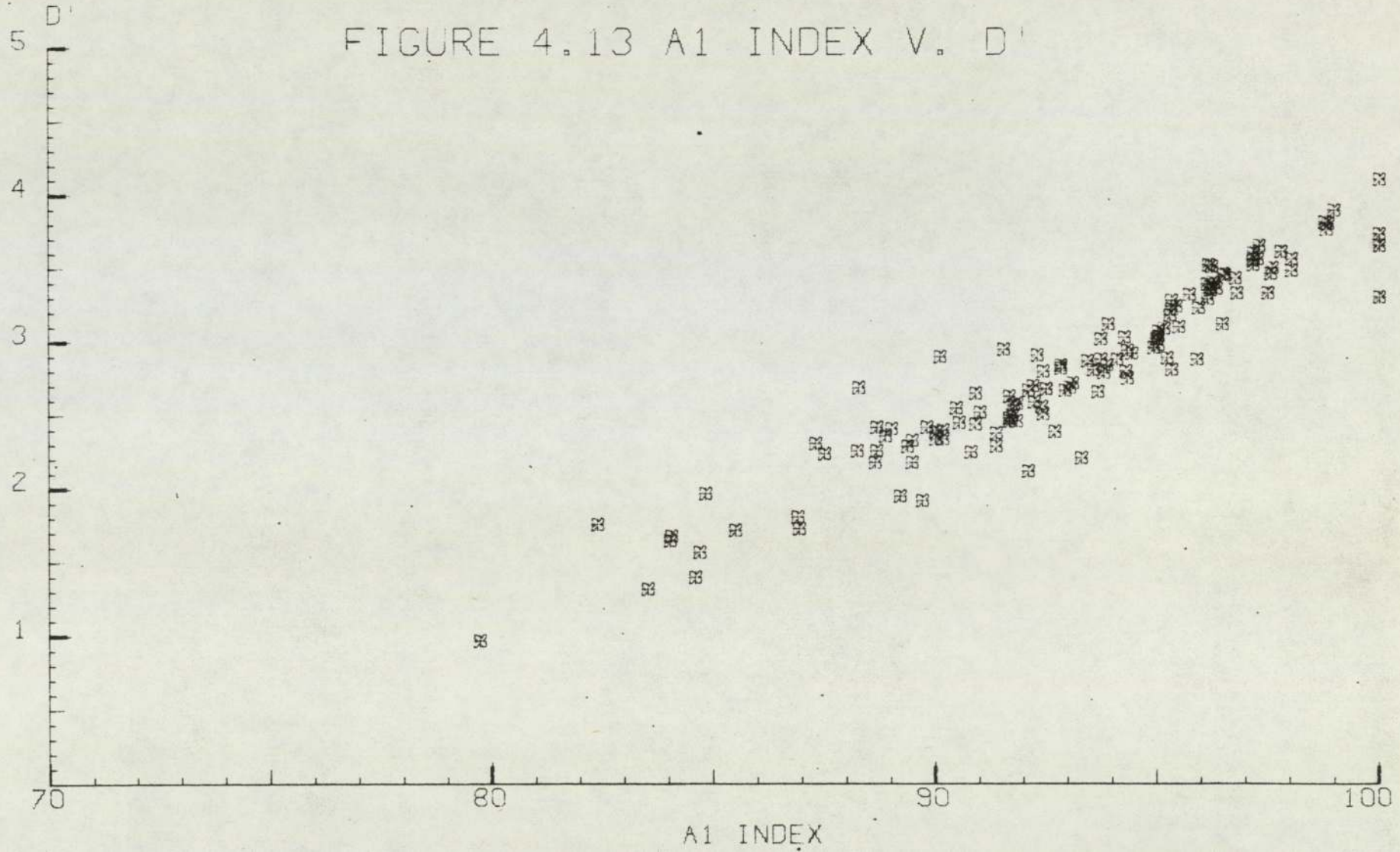
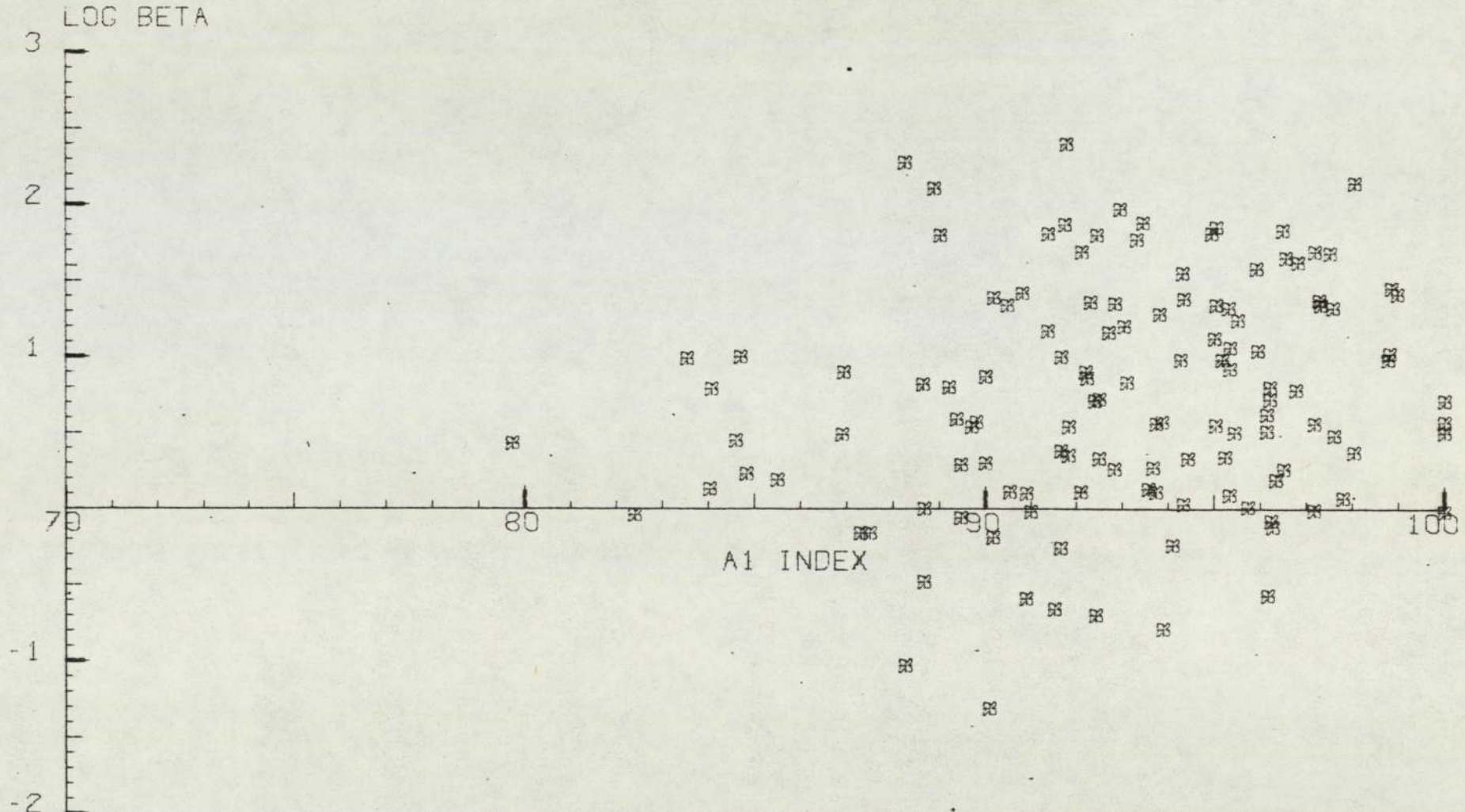


FIGURE 4.14 A1 V. LOG BETA



We can conclude that the non-parametric indices are closely related to the corresponding SDT parameters. This is not surprising in the case of ZFA, because of its monoticity with the evidence variable used by the observer as discussed earlier. The close empirical correspondence between d', log beta and the two corresponding geometrical indices is encouraging in view of the difficulty in establishing an analytical relationship. The fact that the latency sensitivity index also shows a significant correlation with d' suggests that the theoretical response latency model adopted by Navon in deriving the index was correct.

The next performance indices of interest are A1 and A4. Their scatterplot is given in Figure 4.12. The two measures are clearly highly correlated, which is confirmed by an r of 0.742 ($p < 0.001$). Plots of A1 against d', Figure 4.13 and log beta, Figure 4.14 also indicate a high and a more moderate degree of correlation, $r = 0.947$ ($p < 0.001$) and $r = 0.225$ ($p < 0.05$) respectively. These plots demonstrate the essential disadvantage of measures such as A1: they are dependent on both sensitivity and bias changes, and this clearly reduces the possibility of ascribing a cause to a given change in the performance index.

4.11.2 Effects of experimental variables on performance

4.11.2.1 Overall performance measures

Considering the analysis of variance for both the correct detection probability and its arcsine transformation, we see that there are significant differences between subjects ($p < 0.05$) and between time intervals ($p < 0.01$). In fact the time effect is an increase in performance with time, a comparison of means showing that the performance

on the last two ten minute periods of the experiment is significantly better than during the first ten minutes (Tukey test, $p < 0.05$).

The arcsine transformation of the false alarm probability shows a significant time x subjects interaction ($p < 0.01$). Examination of the interaction suggests that there are increases in false alarms with time intervals for five out of the seven subjects. The correct detection and false alarm probabilities taken together suggest that a change in criterion is responsible for the effects, rather than a sensitivity change.

The analysis of variance for the A1 inspection index indicates significant differences between subjects ($p < 0.01$) but does not indicate any time effects. No significant effects are found for A4, in spite of its high correlation with A1.

4.11.2.2 SDT performance measures

The analysis for d' gives significant differences between subjects ($p < 0.01$) but any conclusions drawn need to consider the fact that the unequal variance assumptions do not apply in this case.

In the analysis of variance for log beta the corrected values of beta were used as described earlier. Significant differences were found between time intervals, ($p < 0.05$), with a significant time intervals x subjects interaction ($p < 0.01$). In view of the importance of this variable, a simple main effects analysis was conducted to clarify the nature of the time effects for the different subjects. This is given in Table 4.3 below, and the corresponding graph showing the changes in log beta over time intervals for each subject for

Source	Sum of squares	df	M.S.	F	Significance
T at S1	1.56	2	0.78	2.07	
T at S2	2.78	2	1.39	3.69	$p < 0.05$
T at S3	1.63	2	0.82	2.18	
T at S4	6.75	2	3.38	8.97	$p < 0.01$
T at S5	1.5	2	0.75	1.99	
T at S6	3.4	2	1.7	4.51	$p < 0.025$
T at S7	1.9	2	0.95	2.52	
ERROR		24	0.377		

Table 4.3 Simple main effects analysis: time intervals x subjects interaction for log beta

which the differences were significant is given in Figure 4.15.

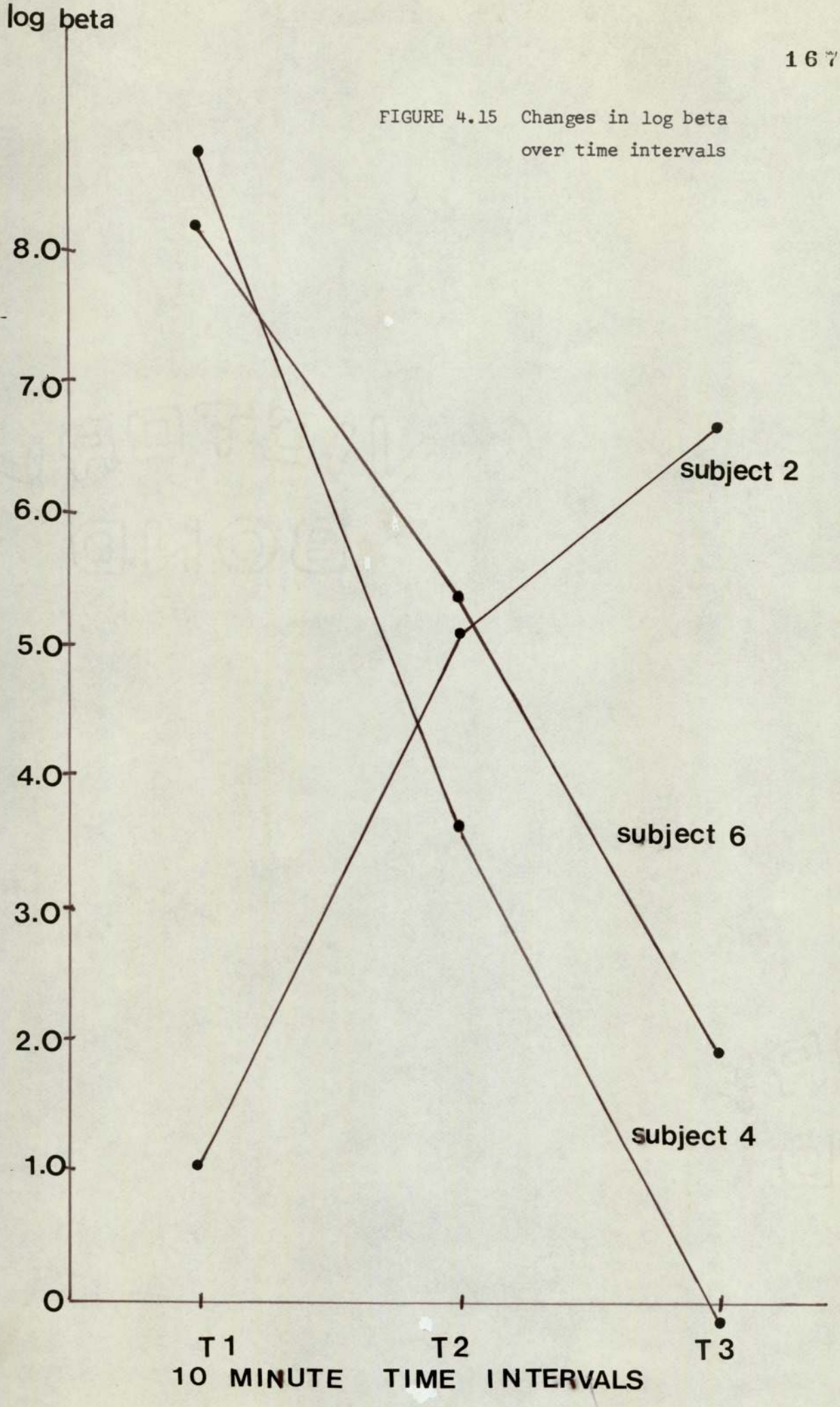
The simple main effects analysis indicates that log beta is significantly different between time intervals for subjects 2, 4 and 6. The significance of the difference between the means for these subjects is shown in Table 4.4 below (Tukey tests).

	Subject 2			Subject 4			Subject 6		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
T1		NS	$p < 0.05$		NS	$p < 0.01$		NS	$p < 0.05$
T2			NS			NS			NS
T3									

Table 4.4 Significance of the difference between log beta means (Tukey test)

With two of the subjects (4 and 6) there is a significant decrease in log beta between the first and last time periods, with subject 2 the trend is reversed. The overall trend, as indicated by the means, is for a decline in criterion with time on task.

FIGURE 4.15 Changes in log beta over time intervals



The Pollack-Norman sensitivity index shows significant differences between subjects on the analysis of variance ($p < 0.05$), paralleling the result for d'. The Navon latency sensitivity index analysis gives significant differences between subjects ($p < 0.01$) and two significant interactions, i.e. time x subjects ($p < 0.05$), and event types x time x subjects ($p < 0.05$). Most of these interactions occur because the quantities used to calculate the Navon index, the latencies for the response categories, are highly sensitive to the experimental variables. As a result, it is difficult to interpret them in this case.

The Hodos-Grier bias index shows a highly significant time effect ($p < 0.01$) and a significant noise condition x time x subject interaction. Consideration of the NxTxS summary table largely confirms the analysis for log beta, for times and subjects. The pattern of the results however, varies in an unsystematic way under the differing noise conditions.

ZFA shows a significant time intervals x subject effect ($p < 0.05$). This is analysed in Table 4.5 to give the simple main effects to compare with the corresponding analysis for log beta (Table 4.3).

It will be seen that the detailed analysis using ZFA does not exactly parallel that using log beta. Although subject 6 shows significant differences in bias across time intervals using both measures, subject 5 shows a significant difference using ZFA as the index and subjects 2 and 4 using log beta. It is apparent, therefore that different results are likely if the two indices are used in statistical analysis, even though the overall trend of the means is identical in each case.

Source	SS	df	S	F	sig.
T at S1	0.07	2	0.035	0.41	
T at S2	0.04	2	0.02	0.24	
T at S3	0.44	2	0.22	2.59	
T at S4	0.43	2	0.22	2.59	
T at S5	0.73	2	0.37	4.35	$p < 0.05$
T at S6	1.44	2	0.72	8.47	$p < 0.01$
T at S7	0.01	2	0.005	0.06	
ERROR		24	0.085		

Table 4.5 Simple main effects analysis for time x subjects interaction for ZFA

4.11.2.3 Latency measures

For the purposes of analysis, the latencies of the four categories of response, i.e. correct detections, correct rejections (of frames containing no events), false alarms and missed signals will be considered separately in addition to the total response time. In most cases, the latencies were subjected to a log transformation before the analysis of variance.

For reasons which will be discussed in the next section, a detailed analysis of some of the more complex interactions will not be given, particularly where these are associated with subject differences.

Considering the correct detection latencies, subjects show highly significant differences ($p < 0.001$) with a significant time by subject interaction ($p < 0.01$). Simple main effects analyses indicate that whether correct detection latency increases or decreases with time on the task depends on the subject, only subjects 2 and 4 showing significant changes.

The correct rejection latency shows a number of significant main effects and interactions. As usual there are significant subject differences, and the main effects of event type and time are also significant at $p < 0.05$ and $p < 0.001$ respectively. Conclusions drawn from these main effects need to be modified by the presence of significant noise x event type ($p < 0.05$) and time x subject ($p < 0.001$) interactions. Simple main effects analysis of the latter interactions shows that all subjects show significant differences between time intervals but that the change is not always in the same direction. As suggested by the time means, however, the predominant change is to a decrease in latency with time on task. Analysis of the noise x event type interaction shows that there are significantly longer correct rejection responses for the 200 as compared with 400 events, but only under the high white noise conditions.

Analysis of the log of the missed signals latency indicates, in addition to the usual subject differences, a significant noise type x subject interaction ($p < 0.05$) a significant noise x time x subject interaction ($p < 0.05$) and a significant event type x time x subject interaction. The means for noise types suggest a longer response latency under the high white noise condition compared with the other types. The event type means imply an overall longer response latency for the 200 events, but the significant interaction times and subjects suggests that this may not be constant over conditions.

Considering the false alarm latencies, the only significant effect apart from subjects is a noise x event type x subject interaction, ($p < 0.05$).

The final latency measure considered, the total latency, shows significant time and subjects main effects, ($p < 0.05$) and $p < 0.001$ respectively,

and significant event type x time ($p < 0.05$) and time x subjects interactions ($p < 0.01$). The means for event types indicate longer latencies for the 200 events and for the earlier time intervals compared with the later.

4.12 Conclusions from experimental study

4.12.1 The application of SDT

The evidence presented in the preceding sections suggests strongly that the unequal variance SDT model satisfactorily accounts for the data obtained. The original proposal that the equal variance model would apply is not supported by the evidence. In view of the fact that the subjects knew the characteristics of signal and noise equally well, as a result of extensive experience, alternative explanations for the good fit of the unequal variance model need to be sought. The greater variance of the signal distribution can be accounted for by the considerable intrinsic variability of the signal in this task and by the sampling error inherent in calculating the false alarm probabilities with the low incidence of false alarms observed in this experiment. These explanations are more tenable than the suggestion that the absence of feedback during the experiment prevented the scanners from learning the characteristics of the signal.

The fairly close correspondence between the optimum criterion for this task and the mean beta actually obtained, adds further weight to the use of the SDT model. It should be pointed out however, that there is a wide range of values of beta about the optimum criterion which will give most of the maximum value possible for the inspector's decisions (Swets and Green (1966) p.93.

The absence of overall significant differences in beta between subjects is in accord with the hypothesis proposed originally, that the scanners would be homogeneous with respect to their criteria, as a result of long experience on the task.

4.12.2 Comparisons of performance measures

Comparisons of commonly used indices of inspection performance A1 and A4, showed that they were correlated with both bias and sensitivity measures, thus limiting their usefulness in terms of suggesting causes for suboptimal performance.

Investigation of the relationships between the parametric and non-parametric indices of bias and sensitivity suggested that the non-parametric indices considered seemed to correlate highly with the co-responding beta or d'. At a detailed level of analysis however, they could not be expected to provide exactly the same results.

4.12.3 Effects of experimental variables

The first questions of interest are those relating to the effects of the experimental treatments on SDT parameters. The changes in correct detections and false alarms with time suggested a decreasing response criterion. This was confirmed by the analysis of variance for log beta, even though a significant decline was only obtained with two subjects, and an anomalous significant increase was obtained with a third. However, the overall pattern of the results was consistent with a reduction in beta with time. The change in beta to a more lax criterion is difficult to account for. In vigilance tasks, the opposite effect, the increasing stringency of beta with time on

task has been accounted for by the inhibition of neural responses, a greater stimulus intensity being required as time goes on, for a signal to achieve a given criterion (Mackworth (1969)). Another explanation advanced is in terms of an inappropriately low signal expectancy at the beginning of the session, leading to few signals being detected, which in turn leads to a lowering of the criterion (Mackworth, (1970)). In the present experiment, as has been emphasized previously, there was no reason to believe that the subjects began with an inappropriate expectancy. If this was not the case, we could account for the results by saying that the subject began the experiment expecting fewer signals than he actually encountered, and that he subsequently lowered his criterion as a result of detecting a high incidence of signals. The neural habituation theory also seems inappropriate, since it cannot be modified to predict a decline in the criterion. In fact the results seem to be more readily accounted for by a suggestion of Welford (1968), that an increase in arousal would move both signal + noise and noise distributions to the right, without modifying the position of the criterion. This would produce an apparent decrease in the criterion. On the other hand the task cannot be regarded as particularly arousing, and there seems no obvious reason why arousal should increase with time.

A more likely explanation would seem to be in terms of an initial modification of the subjects' utilities for the various decision classes, due to their perception of the experiment as a 'high risk' situation. The evidence suggests that initially the scanners approach the experiment with a criterion which was different than that employed in their day to day work. Because they were 'on test' it seems likely that they were initially utilizing particularly stringent criteria as to what constituted a wanted event. This would result in many 'borderline' events being rejected, even though in the real task they would

probably be accepted on the basis that the fine scanner would look at them again and make the final decision. In fact they could be regarded as putting themselves in the position of the fine scanner, who makes the final decisions about which events should be accepted or rejected, and who utilizes more stringent criteria. This effect can be interpreted in terms of payoffs, since during normal scanning, the fact that an event will be looked at again obviously encourages a lax criterion, because the cost of a false alarm is very low. In the experiment, one would expect the 'raised criterion effect' to decline as the scanners become more used to the experiment. Support for this hypothesis comes from an examination of the actual beta values for each of the 10 minute time intervals on the task, i.e. 3.356, 2.587 and 2.197. The final magnitude of beta obtained is very close to the calculated optimum i.e. 2.2. The anomalous subject who increased his criterion may have perceived the utilities of the situation differently than the other subjects.

The implications of this analysis are than the observed decline in beta was a characteristic of the experimental situation rather than an effect which occurs in the day to day performance of the task. Although this may be regarded as a 'negative' finding, it does seem to be of considerable importance when considering industrial experimentation in general. Any experimental study which utilizes an off-line investigation of the type described for signal detection experimentation, needs to control for variables of this type, or to perform a SDT analysis to isolate the effect. The results are also illuminating examples of the way in which subjects are able to modify their criteria without a concomitant change in d'.

The absence of any significant differences between event types was

expected, since there is no a priori reason to expect event complexity to affect beta. Although high intensity white noise is known to affect beta in vigilance tasks (Davies and Tune (1970)), this is generally at a higher level than the 85dB used in the experiment, and therefore the lack of any effect in this experiment is not surprising.
A surprising aspect of the results for d' and the non-parametric sensitivity indices, is that d' should be apparently unaffected by the characteristics of the event. The expected intersubject differences in sensitivity were found but no significant differences between event types, despite the strong a priori reasons discussed in section 4.6.3 for expecting these to differ in detectability. The lack of any effect of noise on d' is again probably due to the relatively low levels (85dB) employed.

The reasons for the lack of an effect on d' of the different complexity of the events are related to the self paced nature of the task, as will be discussed in the next section.

4.12.4 Latency measures

As has been discussed earlier, any response latency observed is the sum of the time to make a structured search to find a configuration of tracks which is a potential event, and the decision time to assign the configuration to the category signal or noise. Visual search theory and the various signal detection latency theories therefore do not in themselves provide good descriptions of the data. Some interesting insights emerge from a consideration of the latency results, although they will not be analysed in detail because of these theoretical difficulties and because response time is a relatively unimportant variable in practical terms for this particular task.

Let us first consider the degree to which the latency results provide support for the application of the SDT model. If we assume that the search time is some random time increment which is added to the decision making time, we can examine the total response latency in the light of SDT concepts. A further assumption that needs to be made is that the latency associated with each decision is a function of the distance of the observation point (in decision space) from the criterion. The further the distance from the criterion the shorter the response latency. This can be interpreted as the further the observation is from the criterion, the more discriminable it is, and hence the less time is required to sample sufficient evidence to make a decision. The notion of a difficult decision requiring a greater decision time is intuitively reasonable.

Considering Yes responses, incorrect Yes responses (false alarms F.A.) will, on average, be distributed nearer the criterion than correct detections (CD), implying that CD latencies will be shorter than FA latencies, i.e.

$$L(CD) < L(FA) \quad - (1)$$

Similarly for No responses, correct rejection (CR) latencies will be shorter than missed signal (Omissions, OM) latencies, i.e.

$$L(CR) < L(OM) \quad - (2)$$

Considering both Yes and No responses, we can say that in a situation such as the present experiment, where the probability of a signal is less than that of noise, No will be the dominant response. No responses will, therefore, on average, be distributed further from the criterion and will hence have a lower mean response latency. Hence both No responses, i.e. correct rejection and missed signal latencies

will be shorter than both Yes responses (correct detections and false alarms). This implies:

L(CR) and L(OM) $<$ L(CD) and L(FA) - (3)

Combining (1) (2) and (3), the latencies for the four categories of response should fall in the order:

CR $<$ OM $<$ CD $<$ FA

The actual mean latencies are, in seconds:

CR = 7.38, OM = 10.24 CD = 10.42, FA = 12.31

which is in the predicted rank order.

A possible explanation for the lack of an effect of event type on indices of discriminability can be found in the response latencies for the two types of event. The latency for the 200 event is consistently longer than for the 400 for all categories of response. In an unpaced situation it seems likely that the inspector is able to overcome the effects of a low signal to noise ratio by utilizing extra time to sample more attributes of the stimulus. Presumably the d' values which would be obtained from a short, fixed interval experiment, would be lower than those found in the self-paced situation.

The results for the effects of the experimental variables on the latencies of the various response categories cannot be simply accounted for, because of the reasons discussed at the beginning of this section. The large number of subject interactions found are likely to be a function of the differing scanning strategies adopted by the subjects, which tends to obscure the effects of the main variables. There is some evidence that the high white noise condition produces longer response latencies for some categories of response and this could be interpreted as evidence of a distraction effect. However, the effect

is small, difficult to isolate from the effects of other variables, and is unlikely to have any practical significance.

In summary, the most important information gained from the latency data is that it appears to provide further support for the application of the SDT model. The data are also useful in suggesting how in a self-paced situation, the subject may utilize extra sampling time to make the correct decisions for difficult discriminations.

4.13 Summary and general conclusions

The inspection task performed in the Data Analysis Group, University of Birmingham, has been considered from the standpoint of some of the variables considered in the review chapters of this thesis. An experimental study was performed to answer a number of practical and theoretical questions. The first of these concerned the applicability of the SDT model to the inspection task under consideration. The interrelationships between various performance measures was also investigated. Finally the effects on performance of several variables, of particular interest to the management of the inspection system, were analysed.

The evidence strongly suggested that the unequal variance SDT model provided a good fit to the experimental data obtained. It was found that in general the non-parametric measures of bias and sensitivity correlated well with their SDT counterparts, but that this correlation was not sufficiently close to produce the same results from detailed statistical analysis. A consideration of the relationship between SDT performance measures and other commonly used inspection measures showed

that the latter failed to adequately distinguish between changes of bias and sensitivity.

Having established that the SDT model was appropriate, the effects of the various experimental variables on the measures of bias and sensitivity was considered. The only significant effects obtained with d' and the non-parametric sensitivity measure of Pollack and Norman was between subjects. The criterion measure, log beta, was corrected for the effects of the assymetrical underlying variances, before being used in the analysis. A significant effect of time on log beta was obtained, the criterion decreasing with time on task. Various explanations were considered for this, and it was concluded that it was likely to be due to the experimental situation, the scanner initially utilizing a high criterion, because of the perceived payoffs of the various types of decision in that situation. As the experiment progressed the scanner utilized a criterion which was more in accord with the utilities found in the everyday task, as suggested by a prior survey of the inspection staff. No other significant effects were obtained with log beta.

Although the latency data was of secondary interest in this study, its analysis in general terms provided further support for the SDT model. A number of complex effects of the experimental variables on the latencies of the various categories of response were obtained. Interpretation of these was restricted because of the difficulty of separating search time and decision time. Considerable between-subject effects were found, probably as a result of differences in scanning strategies. Systematic investigation of the visual search aspects of this task would necessitate a two - phase experiment, in which the decision times would be evaluated independently of the search component.

Although it would have been possible to have conducted further analyses on the latency data in the form of distribution fitting, it was not felt that this would contribute to modelling the situation, because of the problems discussed earlier. The data provided general evidence that the simpler pattern produced an overall longer response latency, although this effect was complicated by subject differences. This result gave an explanation for the lack of any significant differences between the event types as measured by the SDT indices. It was hypothesized that in a self-paced situation the inspector was able to employ additional sampling of a near threshold potential defect,in order to make a decision about it. Self-pacing in inspection could be regarded as a means of enhancing sensitivity from this standpoint.

The conclusions for management were as follows. There was no concrete evidence of performance decrements as a result of the effects of auditory noise. The indirect evidence that the lack of differences found between the two types of event was due to the self-paced nature of the task suggested that no attempt be made to introduce pacing. The fact that there was no evidence for performance decrements with time during the experiment could not necessarily be extended to longer time periods.

In conclusion,the study demonstrated the practical feasibility of the SDT approach to inspection, and the unique insights that can only be obtained by this technique.

CHAPTER 5 CASE STUDY II : THE INSPECTION OF FILM IN THE QUALITY CONTROL DEPARTMENT OF ILFORD LTD.

5.0 INTRODUCTION

The industrial site for this study was Ilford Ltd., at Brentwood in Essex, a large manufacturing unit concerned with all aspects of photographic film manufacture, including Xray, ciné and roll film.

This study was particularly useful in that it served to emphasize the difficulties that can arise when attempting to apply some of the theoretical concepts that were discussed in the review chapters, to situations where the definition of acceptable quality is highly subjective. Although this study and the Data Analysis Group study were superficially related, in that they both involved film, in fact the nature of the inspection task itself was very different. Although the difficulties encountered in this study meant that an in-depth theoretical analysis of the results was not possible, many of the practical problems encountered provided an impetus for the experimental work described in later chapters.

The first part of the study involved a semi-structured interview with four of the senior film examiners and with the Quality Control Manager, Mr R F Salmon. The interviews were tape recorded and their content provided the basis for the general description and the task description which follows. Two experiments were then performed in an attempt to establish the basic performance parameters of the system and to investigate the utility of SDT in tasks of this type. The difficulties involved in the analysis of these tasks will be discussed in detail subsequently.

In spite of these difficulties, a report was presented to the firm which produced a positive response, and the contacts established at

Ilford proved to be very useful during the laboratory phase of the research discussed in subsequent chapters.

5.1 General description

The manufacture of photographic and X-ray film involves the coating of a plastic base material with a photosensitive emulsion. The coating machines which perform this operation operate in total darkness, because of the light sensitive nature of the coating, and produce large rolls of film approximately 44 inches in width. These are subsequently sliced up into a size appropriate for the use to which the film is to be put. The film is sold to a very wide market, with the emphasis on the commercial and industrial sector rather than the domestic field. Consequently the potential users have high quality standards, and quantities of film are sent back to the manufacturers if they are found unsatisfactory.

Quality control at Ilford is the responsibility of the Technological Services Division. The quality control section performs two functions. One is concerned with the maintenance of the physical parameters of film quality, such as coating thickness, grain size and base thickness, which are monitored by routine laboratory procedures. This aspect of quality control will not be considered in this study. The other function involves the visual inspection of film, which will be the main area of interest.

5.1.1 Visual inspection

Visual inspection takes place at two points in the manufacturing

process. The first of these is known as production testing. Samples of coated film in the form of large sheets are taken at an early stage of production, developed, and inspected by being placed over an illuminated table. The function of production testing is to provide rapid feedback if a malfunction in the coating machinery is producing a defective product. This section is run on a twenty-four hour basis in order to do this.

The other type of inspection is known as viewing. In the case of photographic film this involves taking samples of between 200 and 2000 feet in length, with widths of 16mm. or 35mm. from selected parts of the original roll. These are then exposed to give an even neutral density, developed, and sprocketed to produce cine films known as test references, or simply references. The references are projected on to a large screen in a cinema-like projection room. As the film is projected, a clockwork mechanism unwinds a tape in front of the inspector (or 'examiner') at a rate proportional to the rate of film transport through the projector. When a defect is observed, the inspector makes a mark on the tape indicating the nature of the fault. Measurement of the distance of the mark from the beginning of the tape enables the position of the defect to be subsequently located on the reference.

After a number of test references have been viewed in this way, any film which the examiner wishes to examine more closely is put on a device known as a viewing box, which enables the film to be wound at high speed to any position indicated by the tape as containing a defect. The film frames in this locality are then examined by transmitted light from an illuminated panel set in the front of the box, using a hand magnifier if necessary. This procedure, known as

'boxing' enables a detailed examination to be made of any part of the test reference of interest. Samples of film may be cut from the reference and sent to the laboratory for further analysis.

The report that is produced after examination, is usually qualitative in nature and generally consists of a description of the predominant defect characteristics, e.g. 'slight incidence of black spots but acceptable'. The overall philosophy of testing is to examine the film under the conditions that it will eventually be used. Thus ciné film is projected on to a large screen and 16mm. film on to a screen more representative of those used by amateurs. A numerical count of the defect is only made in detail if some problem exists which is associated with the occurrence of a particular defect, or if a count is required for statistical purposes.

The examiner has some prior information as to the potential defects which may occur on a reference. Each reference film can contains a 'coating card' which indicates whether any obvious problems occurred during the coating operation. Since many of the defects arise during the coating procedure, this gives the examiner some prior information as to the nature of the defects which may be expected.

Viewing is regarded as a very important operation, and normally no film is released on to the market until adequate screening has taken place.

5.1.2 Nature of the defects

A large number of configurations occur on film which can be categorized as defects. There are about 15 commonly occurring discrete

defects, which include items such as 'fibres', due to a foreign body being on the film when it was coated, and 'insensitive spots' due to the coating not adhering to the film base at some point. There are also 'continuous' defects, for example there may be a continuous slight variation in the film density. Apart from the more common defects there is a very large category of defects which occur occasionally, and even some which occur once and are never seen again.

The defects vary in size although they are generally very small compared with the area in which they appear. The angular size of an average defect (a 'fibre'), is about 1.44 minutes, and it has a contrast relative to the screen of approximately 10 to 1. If a defect occurs on a single film frame, as is often the case, the presentation time is of course extremely short, approximately 0.04 of a second for films being projected at the usual rate. The rate of occurrence of defects is extremely variable and may change suddenly at any time.

5.1.3 The definition of acceptable quality

The definition of what constitutes an acceptable film is highly subjective in nature. The ultimate arbiters of quality are the experienced film examiners, and any references which may be regarded as borderline will usually be referred to them. Inherent in the criterion of acceptability in a particular case, is the type of product and the customer for whom it is destined. In the case of cine film for example, the occurrence of a relatively large number of defects will be very readily apparent to the user, who will view a large sample of the film in a relatively short time. By contrast, with microfilm for example, the potential viewing population will be

small and the presence of a small defect even on each frame will not significantly degrade the quality of the film for the purposes for which it is intended.

5.1.4 Physical environment

The viewing area consisted of a large cinema-like projection room containing rows of seats. Around the periphery of the room were six viewing boxes with their associated winding machinery. There is usually a small 16mm. projector in use which projects the lower grade film on to a small screen on the sidewall. The main projectors were large, standard 35mm. cinema projectors.

The size of the main screen was 9'6" x 5'5" and the viewing distance 11'6", thus giving maximum visual angles of 79° and 58° respectively. The brightness of the screen as measured by an SEI photometer was 3.55 foot lamberts. Most of the time the ambient noise intensity was 60dBA. When two of the winding machines were being operated, the noise intensity went up to 84 dBA. Although the room was windowless, temperature and ventilation appeared to be adequate. The examiner viewing the main screen was generally seated in one of two large, deeply sprung armchairs provided for the purpose.

5.1.5 Selection and training

No formal procedures existed for selecting inspectors. There were difficulties in obtaining staff generally, partly because of the rather restricted long term career prospects. Most of the younger staff had been taken on to work specifically within the quality control area, whilst the more established examiners had generally

moved into quality control from the manufacturing side of the firm. None of the employees spent the whole of their time examining film, but most of them had at least one session a week.

No formal training programme existed. Most trainees sat with an experienced examiner whilst he was inspecting film until he was satisfied with their general level of competence. Many of the most experienced staff regretted the absence of a training scheme, but felt that there was at present insufficient time for a senior member of the staff to spend long periods with a trainee. It was felt that it took a considerable time to learn to examine film satisfactorily.

The educational background of the examiners was variable. Most of the older workers did not possess any formal qualifications, but some of the recent employees had G.C.E's.

5.2.1 Signal acquisition factors

It is clear that the acquisition of the signal is considerably more difficult than in the task considered in Chapter 4. The very short presentation time of the signal (0.04 seconds) means that the screen area cannot be searched and hence the probability of detection will depend on the size of the visual lobe relative to the size of the total screen area. The visual lobe size in turn depends on physiological factors such as peripheral visual acuity, and perceptual variables such as the ability of the inspector to distinguish between signal and noise. This latter skill is particularly important in film examining because of the very high level of noise stimuli that appears on the projected film. This is a result of dust particles

and other extraneous configurations put on to the film during its development. The perceptual skills required to distinguish between defects and non-defect stimuli take a considerable time to develop, and this is part of the reason for the length of time required to train an examiner. In SDT terms, increasing d' requires a thorough knowledge of the signal and non-signal characteristics.

From the standpoint of possible vigilance effects, the task characteristics would suggest that these are a distinct possibility. The signals are brief and are usually simple rather than complex. They often occur at a very low frequency in time and space, and the task is normally performed in an unstimulating environment. The decision process attending each response is usually simple. These conditions are in accord with those proposed by Kibler (1965) and Mackworth (1970) as giving rise to vigilance decrements. The length of time taken to examine the longest reference is 20 minutes, which is the minimum period at which vigilance effects would become evident. At the end of the 20 minutes there is usually a break of two or three minutes whilst a new film is loaded into the projector, before the next examining session begins. In this way, an examiner may inspect for several hours if there is a particularly urgent batch of film to examine. It seems possible that the short recovery periods between films might not allow performance to return to its original level, in which case a progressive decline over a number of sessions would be observed. This possibility suggested an experimental investigation, which will be described in a later section.

5.2.2 Decision making factors

Decision making factors are of considerable importance in this task. They operate at a micro and a macro level, affecting the decision making applied to an individual defect, and the global decision making employed to decide whether the test reference as a whole is acceptable or otherwise.

If we adopt the SDT paradigm, we would expect decision making at both levels to be influenced by prior probabilities and payoffs. At a defect detection level, the obvious effect would be that due to the expected incidence of defects. If the incidence of defects increased, it is likely that they would not necessarily be detected immediately because of the inappropriate criterion of the examiner. They could be interpreted, for example, as being due to extra dust particles on the film. In fact, examples of this effect were mentioned to the author during the interviews. In one case a particular type of defect had gradually increased in severity over a number of weeks. This was not detected until complaints were received from customers. When the films were re-examined with this new 'set' the presence of the defects was obvious. Considering the payoff matrix, at the defect detection level, false alarms could be regarded as relatively 'inexpensive' since they could always be checked at the boxing stage. This would promote a lax criterion.

The decision as to whether a test reference is acceptable or not is a global one which will depend only partly on the evidence gathered during screening. In addition to this evidence, which is in the form of the number of defects present and the overall appearance of the film, the final decision will also be influenced by prior probabilities

and payoffs, which may be different from those employed during the decisions concerning individual defects. The final acceptability decision will clearly be made partly in terms of the prior probability of the reference actually being unacceptable. This is normally a function of the particular film type, a new product being likely to have a higher a priori probability of being defective than an established one. The payoff matrix is a function of the likely effects of the various categories of decision. The probability of declaring a film acceptable will be influenced by the consequences of releasing a batch of film which is actually defective on to a particular market. Similarly, declaring a film unacceptable carries with it the costs of a possible false alarm, which include long delays whilst the batch of film is thoroughly investigated. There appears to have been no formal attempts to verify if the quality standards operated by the senior examiners are actually in accord with those required by the customer. The general feeling was that if complaints remained at a reasonably low level then quality was acceptable.

Another question which appears to have been neglected is the question of the drift of criteria of acceptability over time. We have seen from studies analysed in Chapter 3 (e.g. Thomas 1962) that in the absence of regular recalibration sessions, or absolute standards to refer to, then it is quite common for criteria to change over time to such a degree that they are no longer in accord with the professed quality standards of the inspection system.

A model of the film examination process in decision theory terms is given in Figure 5.1. It can be seen from the model that the actual film examination procedure produces an accumulation of evidence as to the number of defects and the general level of acceptability of the

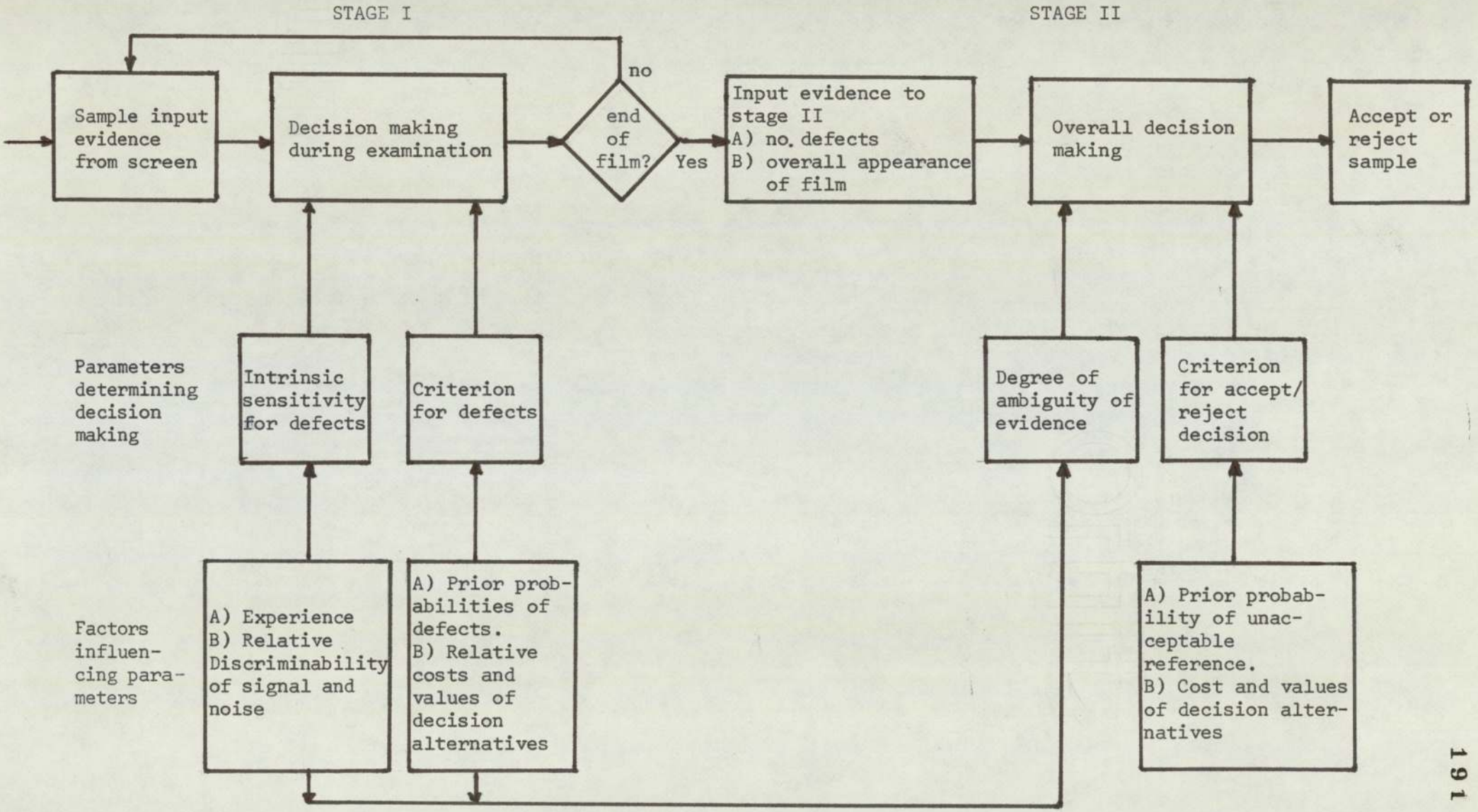


Figure 5.1 A decision making model of the film examination procedure

reference. The quality of this evidence will be influenced by the inspector's intrinsic sensitivity for defects and the criterial factors discussed earlier. If the inspector has no difficulty in distinguishing between signal and noise, (noise in this case could include other defect types if we are considering only one class of defects) then the evidence which is input to stage II will be highly reliable. During stage II, the examiner will have to decide, on the basis of the evidence from stage I, whether to accept or reject the sample. If we regard his sensitivity at this stage of the process as fixed, then his ability to decide between the two possible courses of action will be strongly influenced by the quality of evidence he receives. This is indicated in the model by the loop connecting the sensitivity and criterion factors of stage I to the degree of ambiguity of evidence box in stage II. In terms of SDT, the value of d' that the inspector is using at stage II is influenced primarily by the signal to noise ratio of the evidence. The payoff factors utilized at stage II discussed earlier, together with the prior probabilities of an acceptable or unacceptable sample, will determine the position of the criterion for the final decision.

If the examiner has a high d' at stage I, the evidence utilized at stage II will be truly representative of the actual state of the film, and hence the probability of an incorrect decision will be minimized.

5.2.3 Training

The SDT model proposed has various implications for training. The two stage nature of inspection suggests that any training programme will need to consider both phases. According to the SDT approach, training

should be aimed at two goals, enhancing the discriminatory powers of the inspector as much as possible, and modifying his response bias to the optimal position in a given situation.

Improving d' at stage I basically involves training the examiner to recognize the characteristics of defects as compared with spurious signals under the conditions that they appear on the screen, and also distinguishing between defect types. A consideration of the nature of the defects makes it obvious why training is so prolonged. The very short exposure time of the stimuli makes their initial detection difficult and hence any feedback from the trainer will only be effective on a relatively small proportion of the trials. Another difficulty concerns the lack of reference standards which would allow the trainee to gain insights into the appearance of the defects as they appear tachistoscopically on the screen. There is also an absence of a clearly defined nomenclature for the defect types. Some examiners gave examples of confusion that had arisen as a result of different inspectors using different names for the same defect. It is obvious that this situation is very confusing for a beginner. The training techniques which are appropriate for this situation have been reviewed in Chapter 3 and will be considered further in Chapter 7. The question of training an examiner to adopt an optimal criterion at stage I is a difficult one. An optimal criterion can be calculated from SDT from a knowledge of the a priori defect probability and the payoff matrix. The first difficulty involves the definition of the payoff matrix. Assigning numerical values to the various decision alternatives is particularly difficult in this case, where the results of the stage I decisions are used as input to the stage II decision making. Another difficulty concerns the a priori defect probability. As we have seen, it is pointless training an

inspector to rigidly adhere to a criterion appropriate for a particular defect incidence, if this is likely to change. The solution would seem to be to devise a training procedure which would allow the examiner to alter his criterion readily in the light of the prevailing defect incidence. Presumably this flexibility would also extend to changes of beta implied by changes in the payoff matrix. Although it may not be possible to assign precise values to the various payoffs, it should be possible to say in which direction the criterion should move in a particular case.

Considering the stage II decision making, it is clear that considerable experience is required to know what constitutes acceptable product for a particular market, particularly as the criterion is likely to be weighted by factors such as the urgency of a particular order and the possible cost penalties of not fulfilling it on time. Whether such complex utilities could be trained for explicitly seems doubtful. In any case, before a decision was made about a particularly important market, it is likely that confirmatory opinions would be sought from a number of experienced examiners, and further samples taken. In spite of this, there does not seem to be any reason why examples of acceptable quality film for various products and markets should not be utilized in training to give some indication of the required criteria.

It is not proposed to delineate a training programme at this stage. The review of perceptual training techniques in Chapter 3 would indicate that some form of KR or cuing technique would be an appropriate basis for training for sensitivity. The experimental studies in Chapters 6, 7 and 8 will provide further insights into this area.

5.2.4 Enhancement of the detectability of the defects

Another approach to improving the efficiency of the system, apart from training considerations, is to consider the ways in which defect detectability could be enhanced.

The most obvious way of doing this would be to slow down the rate of projection. This could only be done up to a point, since flicker would begin to be present at very low projection speeds. Nevertheless, it seemed reasonable to suppose that an effectively longer presentation time for each frame would effectively improve the detection probability. This proposal was put to management and met with an unenthusiastic response. The reason for this was that the proposal conflicted with the philosophy of viewing the films under the same conditions as they would be viewed by the customers. Even though more defects might be detected it was regarded as being more important to retain the fidelity of the inspection procedure to the conditions of eventual use. Although this attitude was certainly partly rooted in a desire not to disturb what was regarded as an effective system, there were other arguments that could be advanced to support it. The strongest of these was the fact that the senior examiners had established their criteria of what was acceptable for a particular market and product over a number of years. Any changes in the incidence of defects on the film, due to changed inspection techniques which were not related to true changes in quality, would require a complete recalibration of their standards in order to inspect to the same criteria as before. This emphasizes the fact that it is the integration of information over the entire examination of the film that is the important parameter of quality, not merely a count of defects.

Another suggestion made was to produce negative prints of the test references. This would mean that the defects would appear as light configurations on a dark background, rather than dark on light as at present. The proposed change would mean that afterimages of the defects would remain on the retina for a considerably longer period than the actual presentation time of the defect, thus considerably enhancing their detectability. Using fully dark adopted subjects, data from Alpern and Barr (1962), suggests that the after image would remain for an order of seconds. An additional bonus would be that spurious stimuli due to dust shadows would be absent.

Although this suggestion seems attractive, it was rejected on two grounds. The first of these was that, as discussed earlier, the test would be unrepresentative of its eventual use. Secondly there was the question of the time involved in the special processing of the film. There was constant pressure on the Quality Control Department from Sales to inspect the film references as quickly as possible so that the batch could be released for sale. It was felt that the extra processing time that would be required was not available.

5.2.5 Conclusions from ergonomics considerations

Film examining is a difficult discrimination task because of the short duration of the stimuli, the presence of a high level of visual noise and the highly subjective nature of the definition of acceptable quality. The present training technique would seem to be inadequate to produce trained examiners in a relatively short time, in the light of the difficulty of the task. A number of the techniques for training for perceptual skills discussed in Chapter 3

would seem to be relevant here. Consideration of the task from a decision making standpoint suggests that different types of training may be appropriate for different stages of the task.

It was felt that attention should be given to two aspects of the quality standards employed. The first of these was their degree of correspondence to those of the customers, and the second the possibility of drifts over time of these standards.

There seems to be a strong possibility of vigilance effects being important in this situation and it was felt that an investigation of these factors should be made experimentally. The visually demanding nature of the task also suggested that the visual skills of the inspectors should be examined.

These considerations formed the basis of the experimental studies described in the next section.

5.3 Experiment 2

The first experiment was intended to provide basic data on the overall efficiency of the inspectors and to investigate whether there were any performance decrements in time. Differences in performance between inspectors were also of interest.

5.3.1 Procedure

As the aim of the experiment was to gather data on the system under conditions as authentic as possible, six sample films were chosen by

the Quality Control Manager, which represented an average cross-section of the current work. Each of the six films was presented one after another, in randomized order, to each subject. The films were 2000 feet in length and each film lasted for twenty minutes. Reloading films meant that there was a gap of approximately three minutes between each film. The total experimental session lasted approximately $2\frac{1}{4}$hours. There were no breaks apart from the two or three minutes between film changes.

Each subject was given a set of printed instructions in which he was told to inspect the film in the usual way, noting any defects that occurred on the moving tape. To maintain authenticity, no attempt was made to keep the inspection room silent during the session, and hence there were several sessions in which extraneous noises and conversations occurred.

When all the experimental sessions had been completed, the tapes were collected from the subjects, and each film placed in turn on the viewing box. It was then examined very slowly by two senior inspectors, and a key chart prepared for each film, containing the locations and descriptions of each defect. The subjects' tapes were then examined and a similar chart prepared for each subject. Eight experienced examiners were used as subjects and they all had at least one year's experience and normal vision, according to a Snellen Chart.

5.3.2 Analysis of results

Considerable problems arose in scoring the experiment. Since the films contained a very wide variety of defects, it had been decided

to use one particular defect type, a 'fibre', to assess performance. After the experimental sessions had been completed and the key charts were being prepared, it was found that the overall density of defects differed considerably between some of the test films. The main reason for this was the very high density of 'coating bubble' defects on three of the test films. This led to a very high rate of response by the examiners and since it was impossible to resolve the responses marked on the tape to closer than 3 feet on the film, it was inevitable that the apparent detection efficiency on the high density films would be greater than that for the others.

A further difficulty was the variability in the numbers of fibres on the films, which ranged from two to twenty in number. It was therefore decided that the possibility of confounding together defects of differing discriminability was to be preferred to estimating probabilities with widely differing sample sizes. The performance index used was therefore the percentage of all non-coating bubble defects that were detected. These problems could have been avoided by preselecting the films to ensure approximately equal densities of a particular type of defect, but sufficient time was not available to do this. Additionally it would not have been representative of a typical selection of films. It was decided not to attempt to obtain measures of the incidence of false alarms, since the high incidence of coating bubbles made it impossible to tell if responses were in fact false alarms or correct detections of nearby coating bubbles. For this reason SDT measures could not be obtained.

5.3.3 Statistical analysis

The raw data consisted of the percentage detection efficiency for all non-coating bubble defects for each film examined by each subject.

These were analysed first as a two-way analysis of variance with repeated measures on subjects, the other variable being time intervals of twenty minutes, corresponding to individual films. The data were then de-randomized such that the factors were subjects and films and a similar analysis performed to detect any differences between films in the detectability of the defects they contained.

5.3.4 Results

Table 5.1 gives the percentage detection efficiencies for all defects excluding coating bubbles arranged in randomized order as presented in the experiment.

Subject/Time Interval	T1	T2	T3	T4	T5	T6	Mean detection efficiency for subject
B	72.0	40.0	55.0	32.25	26.30	82.50	51.34
E	47.4	55.0	30.0	69.5	53.0	80.0	55.81
A	92.0	65.0	96.0	74.0	70.0	47.4	74.06
G	52.0	69.50	26.3	64.00	23.5	35.0	45.04
C	70.5	60.0	78.0	31.5	88.0	35.0	60.50
H	52.0	52.0	15.8	48.0	30.0	29.5	37.88
F	35.0	16.0	65.0	36.0	26.3	53.0	38.54
D	44.0	26.0	26.3	20.0	17.6	20.0	25.64
Mean detection efficiency for time intervals.	58.11	47.93	49.05	46.9	41.83	47.8	Grand Mean 48.607

Table 5.1

Percentage detection efficiencies for all defects excluding coating bubbles arranged under the 6 test films used, i.e. de-randomized.

Subject/Test Film	F1	F2	F3	F4	F5	F6
B	55.0	26.3	40.0	32.25	72.0	82.5
E	55.0	47.4	30.0	53.0	80.0	69.5
A	70.0	47.4	92.0	65.0	96.0	74.0
G	35.0	26.3	52.0	23.5	64.0	69.5
C	35.0	31.5	60.0	70.5	88.0	78.0
H	30.0	15.8	52.0	29.5	52.0	48.0
F	35.0	26.3	16.0	53.0	36.0	65.0
D	20.0	26.3	44.0	17.6	20.0	26.0
Mean detection efficiency for each film	41.87	30.91	48.25	43.04	63.5	64.06

Table 5.2

Analysis of Variance for Table 5.1

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F	Sig.
T	1122.95068	5	224.59011	<1	N.S.
S	9629.47072	7	1375.63843	3.9	$p<0.05$
TS	12287.94338	35	351.08404		
TOTAL	23040.36333	47			

Table 5.3

Analysis of Variance for Table 5.2

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F	Sig.
F	6801.32423	5	1360.26489	3.9	$p<0.05$
S	9629.47072	7	1375.63843	3.9	$p<0.05$
FS	6609.57325	35	188.84494		
TOTAL	23040.36333	47			

Table 5.4

5.3.5 Discussion

The results indicate that there were significant differences between subjects in their detection efficiency, significant differences between the films in the detectability of the defects they contained but no significant effects of time on task on defect detection.

In view of the differing difficulty of the test films, the results cannot be taken as indicating unequivocally that there are no time effects on performance. These could have been obscured by the film differences.

The reasons behind the considerable inter-subject variability are worth considering in detail. Subsequent discussion with the examiners revealed that there were large differences in the type of experience they possessed. Although all of the inspectors had been examining for at least a year, the experience of some of them went back over 15 years. It was noticeable that the very experienced inspectors had a detailed knowledge of the mechanics of film manufacture and were able to utilize this to aid them in the inspection. For example, if a defect in the coating machinery caused a distinctive variation in the density of the film, the inspector would then be looking out for other defects associated with the malfunction. Other examiners were accustomed to inspecting mainly colour film, and hence were less familiar with the appearance of defects occurring on the monochrome films used in the experiment. A rigorous pre-selection of the subjects would have provided a more homogeneous sample, but would not have provided the realistic assessment of overall group efficiency that was required by management. Management were in fact surprised at the rather low overall detection probability for non-coating

bubble defects of 49%. Some possible reasons for this finding will be discussed in the conclusion.

5.4 Experiment 3

Experiment 2 had left several questions unanswered. The differences between films had made it difficult to detect any performance decrements with time, and the high incidence of coating bubbles had meant that false alarm rates could not be calculated. SDT measures could therefore not be applied.

For experiment 3, attempts were made to overcome these difficulties by obtaining four films which had an incidence of defects more typical of the norm. These films had also been selected on the basis that they were of equal difficulty and that they contained no distinguishing features that would lead to them being recognized as being different from the normal test references.

The object of the experiment was to test again for decrements in performance with time and also to obtain a wider range of performance measures than in experiment 2. It was also hoped to apply SDT measures to the data.

5.4.1 Procedure

An experimental session consisted of the alternation of the four test films with four normal references. This procedure was replicated a second time during another examining session to provide an increased estimate of experimental error. For ethical reasons, the examiners were told that test films had been introduced into the screening

session, but were not informed of how many films there were, and of their position. Each film lasted for twenty minutes, and hence they were viewed during the following time intervals from the beginning of the session:

test	1	20 - 40 minutes
"	2	60 - 80 "
"	3	100 -120 "
"	4	140 -160 "

Because of the interuptions in the normal flow of work necessitated by the experimental requirements, it was only possible to perform the study on three of the most experienced examiners. These were told to perform film examination in the normal way but to give the tapes produced, indicating the positions of the defects, to the investigator.

The results were scored in a similar manner to experiment 1. The positions of the defects as indicated by the tapes were compared with a key which had been prepared by a special examination of the films under optimal conditions. The detection of 'fibres' was again used to assess performance.

To investigate the visual acuity of the examiners in more detail, subjects were given the Bausch and Lomb Orthorater test for visual skills subsequent to the experiment. Only six of the original eight subjects agreed to take this test.

5.4.2 Statistical design

The experiment was analysed as a 3-way repeated measures analysis of variance, the four presentations of the film being regarded as sampling four time intervals, replications being the second factor and subjects the third. All effects were assumed fixed, by the same logic as was discussed in Chapter 4 section 4.10.3.

5.4.3 Results

Problems were again encountered when attempting to score the results. The main difficulty was the fact that it was impossible to be certain that a particular response was in fact a false alarm, when it occurred in close proximity to another non-fibre defect on the key chart. In fact virtually all false alarms appeared to be misidentifications of this type. This result would be expected if the subjects were able to discriminate readily between defects and spurious stimuli on the film, but were less sensitive to differences between categories of defects. On this assumption, the results can be analysed as a discrimination problem between the defect category under consideration ('fibres') and all other types of defect that occur on the film.

Where no apparent false alarms occurred, the approximation used in Chapter 4 was employed to obtain an estimate of false alarm probability.

Tables 5.5 and 5.6 give the correct detection and false alarm probabilities calculated as described earlier.

		TIMES				
Subjects	Replications	T1	T2	T3	T4	Subject means
S1	R1	0.455	0.273	0.636	0.909	0.568
	R2	0.364	0.455	0.818	0.636	
S2	R1	0.273	0.182	0.182	0.091	0.182
	R2	0.273	0.091	0.182	0.182	
S3	R1	0.182	0.455	0.273	0.455	0.398
	R2	0.455	0.636	0.455	0.273	
time means		0.333	0.348	0.424	0.424	grand mean 0.383

Table 5.5 Correct detection probabilities for Experiment 3

		TIMES				
Subjects	Replications	T1	T2	T3	T4	Subject means
S1	R1	0.083	0.056	0.083	0.083	0.108
	R2	0.056	0.250	0.167	0.083	
S2	R1	0.333	0.250	0.056	0.167	0.160
	R2	0.167	0.083	0.167	0.056	
S3	R1	0.056	0.056	0.083	0.056	0.070
	R2	0.083	0.056	0.083	0.083	
time means		0.130	0.125	0.107	0.088	grand mean 0.112

Table 5.6 False alarm probabilities for Experiment 3

The statistical Appendix A section 2, gives the analyses of variance for the arcsine transforms of the correct detection and false alarm probabilities in Tables 5.5 and 5.6. There are significant differences between subjects in correct detections ($p < 0.001$) but no other effects. All effects are non-significant for false alarms.

		Phoria			Acuity		Depth	Colour
		Vertical	Lateral	Both	Right	Left		
	Range of values from Orthorater tests	1 - 9	1 - 15	0 - 15	0 - 15	0 - 15	0 - 9	0 - 6
FAR	Subject A	5	8	11	9	9	7	1
	Subject B	9	15	11	8	12	8	6
	Subject D	5	11	11	9	8	4	6
	Subject E	7	7	7	7	10	5	6
	Subject F	5	11	10	9	8	5	6
	Subject H	5	5	8	8	8	3	5
NEAR	Subject A	5	7	7	5	7		
	Subject B	9	15	6	7	8		
	Subject D	4	2	12	10	11		
	Subject E	6	10	10	5	9		
	Subject F	6	7	10	8	9		
	Subject H	4	4	7	8	7		

Table 5.7 "Orthorater" results of visual screening of subjects

The efficiency indices A1 and A4 both confirm the significant subject effects for correct detections. The SDT parameters d' and log beta both showed significant differences between subjects ($p < 0.01$ and $P < 0.05$) which were reflected in the corresponding nonparametric Norman-Pollack and Hodos-Grier indices ($p < 0.01$ and $P < 0.05$ respectively).

The SDT indices need to be considered in the light of the fact that the plot of the z-scores for correct detection and false alarm probabilities shows no sign of being fitted by a straight line, thus bringing into doubt the validity of the SDT assumptions in this experiment.

The visual profiles from the Orthorater visual test are given in Table 5.7. An attempt was made to assess the adequacy of the visual skills using the templates provided with the Orthorater which defined minimum visual standards for various jobs. The 'visual inspection' profile provided, however, made no reference to the far visual skills required for film examination. Subjects F and D show the most adequate visual standards, their only deficiency being some degree of lateral phoria (an inability to fuse images from each eye). Subject H has both near and far visual deficiencies whilst subject A has below standard near vision and impaired colour vision. There was no correlation between these scores and performance on the task however.

5.4.4 Discussion

The salient feature of the results is the very low overall detection probability for defects obtained in this experiment. In view of the

fact that the subjects were three of the most experienced examiners, the overall detection probability of 0.38 needs to be accounted for.

The most likely explanation would seem to be in terms of the way in which the inspectors view the task of examining. As has been mentioned before, the prime function of examining is seen as globally establishing the acceptability or otherwise of a film rather than of obtaining a numerical count of defects, except for special purposes. For this reason the information on the tape will usually be more in the form of an aide memoire indicating the approximate disposition of defects rather than a precise record. Although in the experimental situation the subjects obviously made a more determined attempt than normally, to document the occurrence of defects on the tape, it must be recognized that the act of noting down every defect would be unfamiliar to them. This explanation is confirmed to some extent by considering the detection results for experiments 2 and 3. Experiment 2 was far more of a special situation where the examiners could concentrate on attempting to note and detect defects, knowing that none of the sample films was actually a 'real' reference which had to be assessed for acceptability. In the case of experiment 3, they would not be certain which of the references was a test film, and would therefore have to treat all of them in the same manner. This would involve concentrating on the more general aspects of film quality to the detriment of noting down specific defects. These contrasting strategies are reflected in the higher apparent detection probabilities of experiment 2 compared with experiment 3.

In view of these considerations, it is not surprising that SDT does not appear to fit the data according to the Z-ROC curve. From the

preceding discussion it is clear that from the data available it is not possible to decide unambiguously whether or not SDT applies in this situation. However, it is interesting to note that the significant differences in log beta between the subjects reflects the appropriate changes in correct detections and false alarms that would occur if the theory did apply, i.e. decreasing log beta is accompanied by increases in correct detection and false alarms. Log beta will always provide an index of such changes, whether or not SDT applies.

With regard to the changes in performance over time, no significant effects were found. It is notable, however, that the changes in log beta over time intervals, although not statistically significant, showed a clear decline similar to that observed in experiment 1 in the last chapter. The absence of significant vigilance effects can be ascribed partly to the short breaks between each film and partly to the wide variety of non-target stimuli that occurred during the session, which can be regarded as maintaining arousal.

5.5 Conclusions

This study exemplifies the difficulty of using the SDT approach in situations where obtaining accurate estimates of correct detection and false alarm probabilities is difficult, and in which the standards for acceptability are not rigidly defined. In spite of these difficulties, the two stage decision model described would seem to be applicable in this and similar situations where an accumulation of evidence over time is utilized to make an accept/reject decision.

The practical difficulties of scoring the data for the experiments described in this chapter made a valid test of the applicability of SDT at the defect detection stage impracticable. Although the results quoted do not provide direct support for SDT, this is not surprising in view of the small sample size and the rather gross assumptions made in scoring the data. The changes in correct detections and false alarms over time, although not statistically significant, are consistent with a criterion change in a more lax direction, as observed in experiment 1, Chapter 4.

Another experiment was planned to test the stage II decision making part of the model, by asking the inspectors to rate a number of films in terms of their degree of acceptability, from completely acceptable, to completely unacceptable. This would enable the generation of ROC curves and provide a more meaningful test of the stage II part of the model than a yes-no experiment. Unfortunately sales and production pressures meant that the experiment had to be deferred indefinitely.

No support was found for the hypothesis that there would be performance decrements with time on the task, despite subjective reports of occasional sleepiness by some of the examiners when normally performing the task. The suggestion was made that the variety of stimuli which occur on the film provide sufficient stimulation to maintain arousal.

The investigation provided the impetus for several lines of research which are described in subsequent chapters. The first of these was the possibility of training the inspector to be able to modify his criterion readily in the face of changing defect rates or payoffs.

The study suggested a need for this ability at both stage I and stage II of the decision making process. The need to provide a rapid means of training the examiner to discriminate between defects and noise stimuli, and between different categories of defects, was also identified as being an important research area. Finally the absence of any effective selection procedures suggested the possibility of work on this topic.

The analysis of the inspection system in terms of some of the dimensions proposed in Chapter 3 produced a number of specific recommendations for management.

It was suggested that a library comprising examples of the various types of defect be established, which could be referred to readily. Similarly a proposal was made to keep a range of film reference examples which provided clear examples of the quality standards required for particular markets and products. It was felt that a uniform system of nomenclature for both types of defects and levels of acceptability would further aid the establishment of clear quality standards.

In conjunction with these latter suggestions, the practice of regular calibration sessions was proposed as a means of ensuring that quality criteria, especially those defined largely subjectively, did not drift over time and that there was a high degree of concordance between inspectors. The relationship between customer standards and inspection criteria was felt to be an important one. Investigation of this relationship would, however, require an extensive market research exercise.

Returning to the training area, it was suggested that the results of further research be fed back to Ilford to aid in the specification of a training scheme. In addition to the perceptual skills discussed earlier, it was felt that information on film manufacture should be given to trainees, in view of its apparent importance in the identification of defect types.

Finally there was a recommendation that all potential film examiners should receive some form of industrial eye test such as the Orthorater. Although there was no apparent correlation between visual abilities and performance on the task, the examiners with the most impaired vision were in fact the most experienced. It was likely therefore that this would provide compensation for any visual deficiencies. Adequate visual skills were felt to be particularly important during the training stage.

CHAPTER 6 INVESTIGATIONS INTO THE EFFECTS OF DEFECT PROBABILITY CHANGES ON INSPECTION PERFORMANCE

6.0 INTRODUCTION

The effects of changes in the probability of defects on inspection performance has been referred to at several points as being an important consideration in inspection situations found in industry. An example has been quoted in Chapter 5, where quality levels of a product became unacceptable because inspectors were unable to modify their criteria to take into account the changes in defect probability that had occurred.

In this chapter, a brief survey will first be made of some of the relevant theoretical approaches. Two experiments will then be described in which some of the factors affecting performance in the changing defect density situation were investigated.

6.1 Theoretical considerations

The modification of performance in a changing defect probability situation can be seen as consisting of two aspects: perception of the fact that the probability has in fact changed, and the adaptation of the inspector's response strategy to the new conditions. We shall consider these two areas separately in the following discussion.

6.2.1 Probability learning

A large literature exists on the ability of subjects to learn probabilities. The classical probability learning situation differs, however, from that which occurs in inspection. Usually the subject is asked to predict which of two mutually exclusive events will occur, or

has occurred, on each trial of a series in which two events occur with fixed, but unequal, probabilities. For example, the subject may be asked to predict which of two lights will blink on at each trial of a series. He is allowed to witness the outcome of each event and is therefore completely reinforced. He may, in addition, be given monetary rewards for correct predictions and/or be penalized in some manner for incorrect performance.

A number of theories exist for analysing behaviour in these situations. One of the earliest, from game theory considerations, was that after a few trials the observer will distinguish the more frequent event, and predict it on all succeeding trials, thus maximizing his total number of correct predictions. This 'pure strategy' occurs very infrequently. Another theory, called the probability matching hypothesis, predicts that the subject will learn to match his response ratios to the probabilities of occurrence of the two events. This behaviour was first noted by Grant, Hake and Hornseth (1951). They observed that over a series of trials in which two alternative reinforced events occurred with fixed probabilities, the subject's probability of predicting a given event tended to approach or "match" the actual probability of the event. A large number of other studies, e.g. Bush and Mosteller (1955), subsequently showed that the subject begins by predicting the two events equally often, then after a slow initial rise he comes to predict the two events with the actual probabilities of their occurrence. Later work by Edwards (1961) produced evidence suggesting, however, that if a large number of trials is given, i.e. > 300, then the subjective probability estimate becomes more extreme than the actual probability.

Siegel and Goldstein (1959) produced an analysis of probability learning from decision theory considerations which has obvious affinities with SDT. They suggested that the maximization of subjective expected utility will account for both probability matching and pure strategy (i.e. always responding the most likely alternative) behaviour. Since the utility of an outcome is its subjective value to the subject, his behaviour will be a function of the particular conditions and reinforcements inherent in the task. The general hypothesis is that he will maximize his subjective utility in whatever terms he perceives it.

Siegel and Goldstein suggest that in a task with no external payoffs, (monetary rewards etc.), the subject receives satisfaction from predicting and having confirmed the rarer of the two events. This might be expected to relieve the monotony of the pure strategy of predicting the more frequent event on all trials. For whatever reason, it is hypothesized that the subject will adopt a mixed strategy which approximates to the matching hypothesis.

If however, the subject receives some monetary reward for a correct prediction, or a penalty for an incorrect decision, the theory hypothesizes that as these rewards are increased, the subjects prediction of the more frequent event will tend towards 100% i.e. will approach a pure strategy. Experimental evidence confirmed these predictions.

A number of other factors have been found to affect probability learning. These are reviewed in Lee (1971), but will not be discussed in detail here because of their limited relevance to inspection.

It is clear that there are affinities between the inspection situation and research on probability learning. The main difference is in the nature of the reinforcement received by the inspector. In the probability learning task, the subject is usually completely reinforced, whereas the inspector receives only partial reinforcement, depending on such factors as his ability to recognize defects. The occurrence of inspection error leads to the inspector having an incomplete knowledge of the true proportion of defects. A probability learning study by Estes and Johns (1958), involving ambiguity in the reinforcement, similar to that occurring in the inspection task, found reasonably close agreement between the frequency an event was predicted, and the frequency with which it was judged to have occurred.

It seems clear then, that the most fundamental difference between the probability learning situations usually studied and the inspection task is the amount of information the subject receives concerning the nature of the overall defect probability. In the probability learning task, the only information available is that obtained from the responses. On the other hand, this information is usually complete. In inspection tasks, the evidence from responses is partial, but there may be other sources of information available, as will be discussed in the next section.

6.2.2 Sources of information on defect probability

Two sources of information are available to the inspector, from which he is able to estimate the prevailing defect probability. These are external sources and evidence inferred from the task itself.

The first type of external evidence could be described as 'feed-forward' and consists of prior knowledge of the defect frequency, acquired either as a result of experience or from some other source. Examples of such information sources have already been quoted in Chapter 5, where film inspectors are provided with a card detailing the manufacturing history of a particular batch of film. In the glass industry, inspectors of float glass are given prior warning that the incidence of defects may change by the furnace operators working upstream from the inspection point (Gillies (1975)). The other form of external information is feedback, or knowledge of results concerning his accuracy, that the inspector receives subsequent to the actual inspection of the sample. In most real inspection situations such feedback is delayed and incomplete, and cannot readily be related to the separate decisions made during the task.

The most accessible source of evidence available to enable the inspector to modify his subjective probability estimate is from the task itself. The accuracy of the subjective probability estimate will depend primarily on the inspector's ability to distinguish signal from noise, and to a lesser extent on the appropriateness of his current criterion to the prevailing defect density. A dynamic situation exists, in that the more optimally the criterion is set, in relation to the actual defect density, the more accurate will be the information available to the inspector to modify its position still closer to the optimum. It should be noted that these considerations are strictly only of importance in a relatively low d' situation. Where the signal to noise ratio is high, changes in defect probability can be expected to have a relatively slight effect on performance. On the other hand most real inspection

situations involve discriminations between perfect and defective items which would normally be regarded as 'difficult'.

6.2.3 Modification of the subjective probability estimate and the criterion

In the preceding sections, we analysed the sources of evidence available to the inspector concerning the fault density. In considering how this information will be utilized by the inspector to modify his criterion we have to examine several further questions. First, it is necessary to know the extent to which the inspector will utilize the information to revise his subjective probability to correspond to the actual defect probability. Secondly there is the question of how large a sample of the new probability he needs (in a situation of change) in order to effect this change. Finally, we need to consider the ability of the subject to modify his criterion to the appropriate optimal position on the basis of his revised subjective probability.

Data on the first question is available from a number of studies in the area of the revision of subjective probabilities, e.g. Edwards (1962), Edwards and Phillips (1964), Stael von Holstein (1971). In general these studies suggest that observers do not revise their subjective probabilities to the optimal extent that Bayes' theorem would predict. There is very little evidence on the size of the sample needed by a subject to modify his subjective probability as a result of objective probability changes. The sample size might be expected to depend strongly on the discriminability of the signal employed. Nearly all SDT experiments have employed well practiced subjects and have used a fixed within-session probability. Within

these constraints, subjects have shown that they can use the appropriate beta value corresponding to the a priori probability (Swets and Green, (1966)).

One study does exist which investigated subjective probability considerations in an inspection context. Sims (1972) used a simulated inspection task in which subjects had to inspect printed facsimiles of printed circuit boards containing a range of defects. In the first experiment, prior to examining each item, the subject had to record whether or not he felt that it was going to be defective or not. Having made their predictions, the inspectors then examined the items and assigned them to accept or reject baskets. Three percentages of defects were used, i.e. 26, 14 and 2 percent and the subjects were either told the incidence in advance of defects or given no information at all. In the second experiment, the subjects were told that the defect level was 2%, but unknown to them there was a step increase in the defect rate to 14% midway through the session.

Unfortunately, the experiment was analysed purely in terms of the inspector's perception of the defect probability, and no detection data were given.

Considering the subjective probability estimates however, it was found that the subjects' estimates were significantly different for the differing probability levels within 30 trials of the start of each session. As expected, these estimates were in general more conservative than the actual probabilities. It was found that prior knowledge of the probability did not affect the accuracy of the subjective estimates. The results also suggested that the final

subjective estimates of probability were more closely related to the proportion of items classified as defective, than to the actual probabilities. The probability change experiment produced anomalous results. Only two subjects were employed, and in one case the subject accurately changed his subjective estimate of probability to the new probability level within sixty trials. The other subject however, did not show any significant changes. Unfortunately no investigation was made of possible differences in sensitivity which could have accounted for this finding. The result that subjects match their probability to the perceived incidence of defects is one which might have been predicted on intuitive grounds. The lack of any effect of prior knowledge is a surprising one and suggests that the subjects weighted intrinsic information from the task more than prior evidence. This could of course be a function of the degree of difficulty of the task. Presumably the harder the task, the more reliance would be placed on prior evidence.

6.3 Experimental objectives

In view of the foregoing discussion it is clear that there are several questions in the general area of detection performance under changing probability conditions that need to be investigated.

The first of these is the general question of the ability of an inspector to adjust his criterion to the on-going defect probability. The other considerations concern the efficacy of knowledge of results and prior warning of a defect change in aiding this criterion change. Although the Sims study suggests that the inspector can correctly estimate the new probability, this does not guarantee that

this will lead to the appropriate criterion being adopted. Another consideration was whether the inspector's performance strategy was different if the change occurred during a session, rather than from the commencement of the session. Two experiments were conducted to investigate these questions, and will be described in detail in subsequent sections.

6.4 Experimental design: general

It was decided to employ an experimental design which allowed the accurate evaluation of SDT parameters. The fact that a minimum of 500 trials are recommended to do this (Swets and Green (1966)) and that a number of experimental conditions were to be investigated, meant that the number of subjects that could be included in the study was limited by time constraints, particularly as it was felt that extensive practice at the task was necessary in order to minimize learning effects. In view of the exploratory nature of the study, it was felt that three subjects would produce meaningful results.

In order to facilitate obtaining SDT parameters, the experiment was performed using a rating scale approach (McNicol (1972), p.99) as will be described in detail subsequently.

The first experiment was designed to investigate the situation where the defect probability remained constant throughout a session, but varied between sessions. The presence or absence of feedback on performance was considered, and performance changes between blocks within the session were also included in the analysis.

The second experiment was concerned with within-session changes in defect probability. The probability of defect occurrence changed after the second block within each session. In the first session, no warning was given that a change was going to take place. In the remaining sessions, subjects were warned that a probability change would occur. In the one case they were given summary feedback every hundred trials, whereas in the other case no feedback was provided.

The conditions investigated in the two experiments are summarized as follows:

Experiment 4

1. Constant high probability of signal (p = 0.5)

a) Feedback

b) No feedback

2. Constant low probability of signal (p = 0.1)

a) Feedback

b) No feedback

Experiment 5

Change in probability during session (p = 0.5, 0.2)

a) No warning, feedback

b) Warning, no feedback

c) Warning, feedback

Five hundred trials were given under each condition and the order of presentation of the signals was randomized, subject to the constraint that the probabilities were constant within each block of 100 trials.

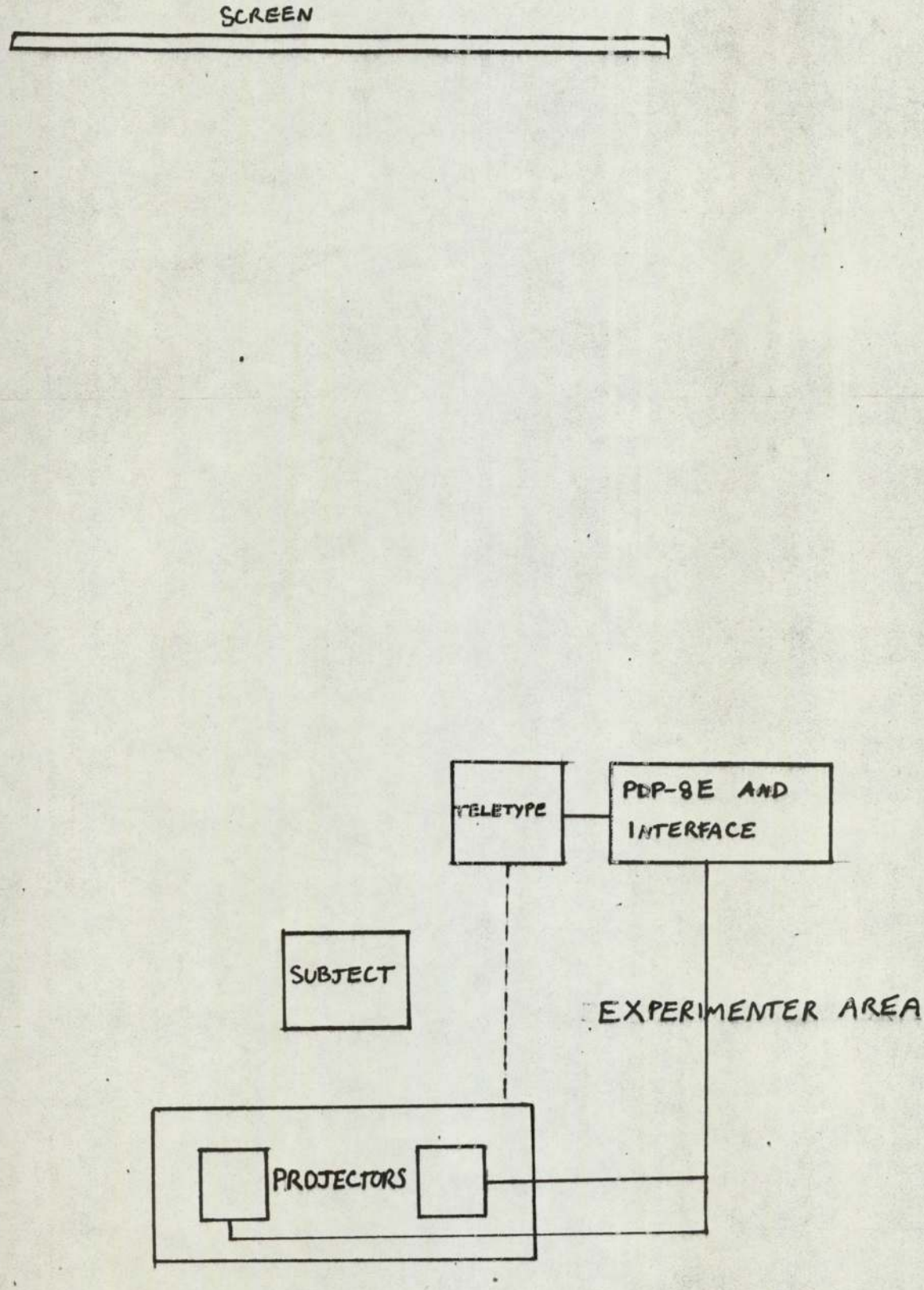
In view of the fact that vigilance effects would complicate the interpretation of the results, the experiment was designed specifically to minimize these effects. This was achieved by providing short rest periods every 100 responses, which meant that the subjects did not perform the task continuously for longer than about eight minutes at a time.

6.4.1 Apparatus

It was decided to build a general experimental facility that could be utilized for a wide variety of detection experiments in addition to those described in this thesis. The equipment was based on a PDP-8E computer and the stimulus material was projected by means of a pair of Kodak Carousel projectors, on to a screen in the darkened experimental room.

Essentially the requirement was for a device that could present signal and non-signal stimuli in a random sequence, would accept a series of responses corresponding to a rating scale and would provide various forms of knowledge of results for these and later experiments. It was also felt desirable to record the latency between the presentation of the stimulus and the response. The computer controlled the sequencing of stimuli, recorded the responses, and provided the feedback. A wide range of other facilities were also available which are described in detail in Appendix C . The equipment represents a highly general means of conducting a wide range of detection experiments. However, its design and construction, and the programming of the PDP-8E computer, which was the entire responsibility of the author, consumed a very considerable amount of

FIGURE 6.1 EXPERIMENTAL ARRANGEMENT



time, and in retrospect it is felt that a simpler if less elegant means of conducting the experiments would have been more appropriate.

The general experimental layout is shown in Figure 6.1. The first projector produced a constant display on the screen, consisting of a matrix of 9 equally spaced black discs, each 5 inches in size, giving an angular size of 2.6° from the table where the subject was seated, 10.5 feet from the screen. The second projector contained alternate opaque slides and slides with a small hole punched in them. When one of the latter slides was in the projector, the resulting pencil of light superimposed exactly on the central disc of the display. Filters were used with both projectors such that the increase in brightness produced was close to the threshold of detectability for each subject. A Compur magnetic shutter,which was controlled by the computer, operated to give a brief presentation of the stimulus. A teletype in the experimental area read in a steering tape which caused the second projector to either remain on the current slide, or to advance to the new slide, prior to the next operation of the shutter. This arrangement allowed a random sequence of any number of stimuli to be presented from a single Carousel magazine of 80 slides.

6.4.2 Procedure

At the beginning of the experiment the projector advanced to the first slide and the shutter opened for 0.25 seconds. This interval was chosen to approximate to the time of a single eye fixation, since it was originally intended to utilize the results in the analysis of subsequent search experiments. The subject then pressed

one of the response buttons and the response latency and the channel number of the response were output on paper tape by the computer. After a delay of three seconds from the response, the next stimulus was presented.

The six response buttons corresponded to six degrees of confidence that a signal had been presented on the preceding trial. From left to right, the ratings were: definitely non-signal, probably non-signal, possibly non-signal, possibly signal, probably signal, definitely signal. A symmetrical pay-off matrix was assigned to the responses in the following manner. Correct positive or negative responses were given 3, 2 or 1 points, depending on whether they were made at an extreme, intermediate, or low level of confidence. Incorrect responses gave minus these payoffs. Subjects were given 750 points at the beginning of the session and paid a bonus if they scored a further 750 points from their responses, i.e. if they scored half the total possible number of points from the 500 responses.

After each block of 100 responses the number of correct detections, correct rejections, false alarms and missed signals for that block was printed out on the teletype, together with the points scored, and the total cumulative number of points including the earlier blocks in that session. For the feedback conditions, the teletype was situated so that the subject could read this printout. A three minute break occurred between each block of 100 responses to allow the subject to do this. During non-feedback sessions the same length of break occurred. With both types of session, the experimenter made non-committal conversation with the subject during the break.

White noise was played at a low level through headphones for the whole of the session, apart from rest periods, to mask the sound of the shutter and projector operation. As the experiment was self-paced, its duration varied slightly, but most sessions were completed within about 45 minutes.

Prior to the main experimental sessions, all subjects had received at least 2000 practice trials using the same apparatus and stimuli. During the practice sessions, feedback similar to that administered during the test sessions was given, and the defect probability remained constant at 0.5. The last practice session had taken place one week before the first test session.

During the practice periods the nature of the scoring system was carefully explained to the subjects, with emphasis on the fact that both missed signals and false alarms would be equally penalized. The use of the rating scale response buttons was discussed, and any difficulties over what was meant by, for example, a 'probably signal' response were resolved. Immediately prior to each test session the subjects were told that the defect probability would be 'the same as during the practice sessions', or simply that it would be 'low', as appropriate. During the first session of the within-session probability change experiment, where no warning was given, the subjects were told that the defect probability would be the same as during the practice sessions. For the two subsequent probability change sessions, they were simply told that a change would occur at some point in the session. Those subjects who had not inferred that a change had taken place during the first session were explicitly informed of this immediately after the completion of the session.

All subjects received condition 1a first, which was identical to the practice sessions and was intended as a 'warmup'. The remaining conditions of the first experiment were presented in a random order, followed by the second experiment. Each subject performed one session per day. Attempts were made as far as possible to ensure that the subjects performed the task at the same time each day, but the exigences of lecture timetables meant that this was not always possible.

All subjects were final year undergraduates and were paid 50p per session, with a 25p bonus if they achieved the target score.

6.4.3 Statistical design

Experiment 4 was analysed as a 4-way factorial analysis of variance, the factors being blocks of 100 responses, presence or absence of feedback, signal probability of 0.5 or 0.1, and subjects. Experiment 5 was set out as a three way analysis of variance, with blocks of 100 responses, the three experimental conditions and subjects being the factors.

In both experiments all subjects received all combinations of conditions. As discussed in Chapter 4, the danger of carry-over effects exists in this type of design. In the present experiments, these effects were minimized by having a highly trained subject population and by including the blocks within the experiment as one of the specific factors. In view of the laboratory based nature of the experiments, a mixed model was utilized in the analysis of variance, with subjects being assumed a random effect and all other factors fixed.

6.4.4 Analysis of the results

The latency and response data which had been produced on paper tape by the PDP-8E computer, were fed into the PDP-15 computer for analysis. The responses were compared with the actual signal sequence and a number of performance measures produced.

The use of rating scale data allowed some performance indices to be calculated which had not been available for the yes-no experiments analysed previously. These indices have been discussed in detail in Chapter 2. They were the Altham-Hammerton sensitivity index, and P(A), the area under the ROC curve, which provides a sensitivity index independent of the underlying variances. The latter index was calculated using a simple numerical integration technique (McNicol (1972) p.115). Another quantity calculated was the score achieved by the subject obtained from the rating responses made and the payoff matrix. The remaining performance indices were those employed in previous experiments. These were the latencies for the various types of response, the false alarm and correct detection probabilities, d' and beta, and the corresponding nonparametric indices. The rating scale approach allowed the Grey-Morgan program to be utilized to fit ROC curves to each block of data. This in turn allowed correction of the beta values to take into account the ratio of the variances, and the calculation of d_e', the corrected value of d' discussed in Chapter 2. In most cases the unequal variance SDT model produced an acceptable fit to the rating data. In a small proportion of cases, an ROC curve could not be fitted because the subjects had used only the extreme response categories, effectively producing a single point on the ROC curve. In these instances, the slope for the ROC curve

fitted to the whole 500 points for the entire session was utilized. The mean ratio of signal to noise variance was 2.564, and there were no significant differences in the sigma ratio between experimental conditions.

6.5 Results - experiment 4

The analyses of variance are given in Appendix A and the raw data and condition means in Appendix B.

6.5.1 Signal detection results

The correct detection and false alarm probabilities for the five blocks of responses within the four experimental conditions of $P = 0.5$ and $P = 0.1$, feedback present or absent, are given in Figures 6.2 to 6.5.

The analysis of variance for the arcsine transforms of the two probabilities indicates significant differences between blocks and between probability conditions for the false alarm probability ($p < 0.05$ in each case) and significant differences between feedback conditions for the correct detection probability. Applying the Tukey multiple comparison test to the false alarm probability block means, indicates that this is significantly higher for the first block compared with the remaining four blocks ($p < 0.05$). With regard to the probability condition means the false alarm rate is significantly higher for the $P = 0.5$ condition than the $P = 0.1$ condition.

The analysis further indicates that correct detection probability is significantly higher under the feedback condition compared with the non-feedback situation ($p < 0.05$).

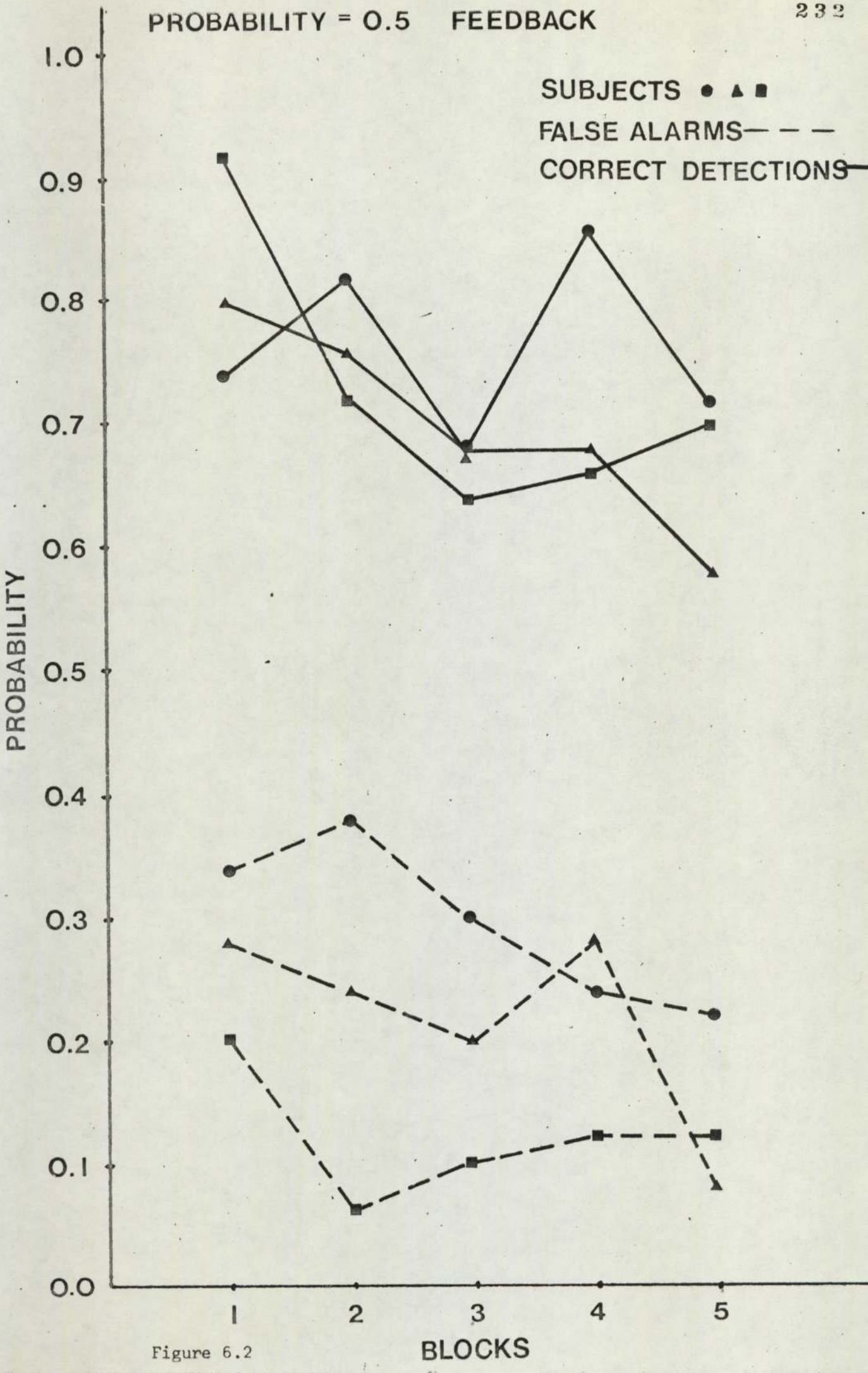

PROBABILITY = 0.5 FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS— — —
CORRECT DETECTIONS
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
1
2
3
4
5
BLOCKS

Figure 6.2

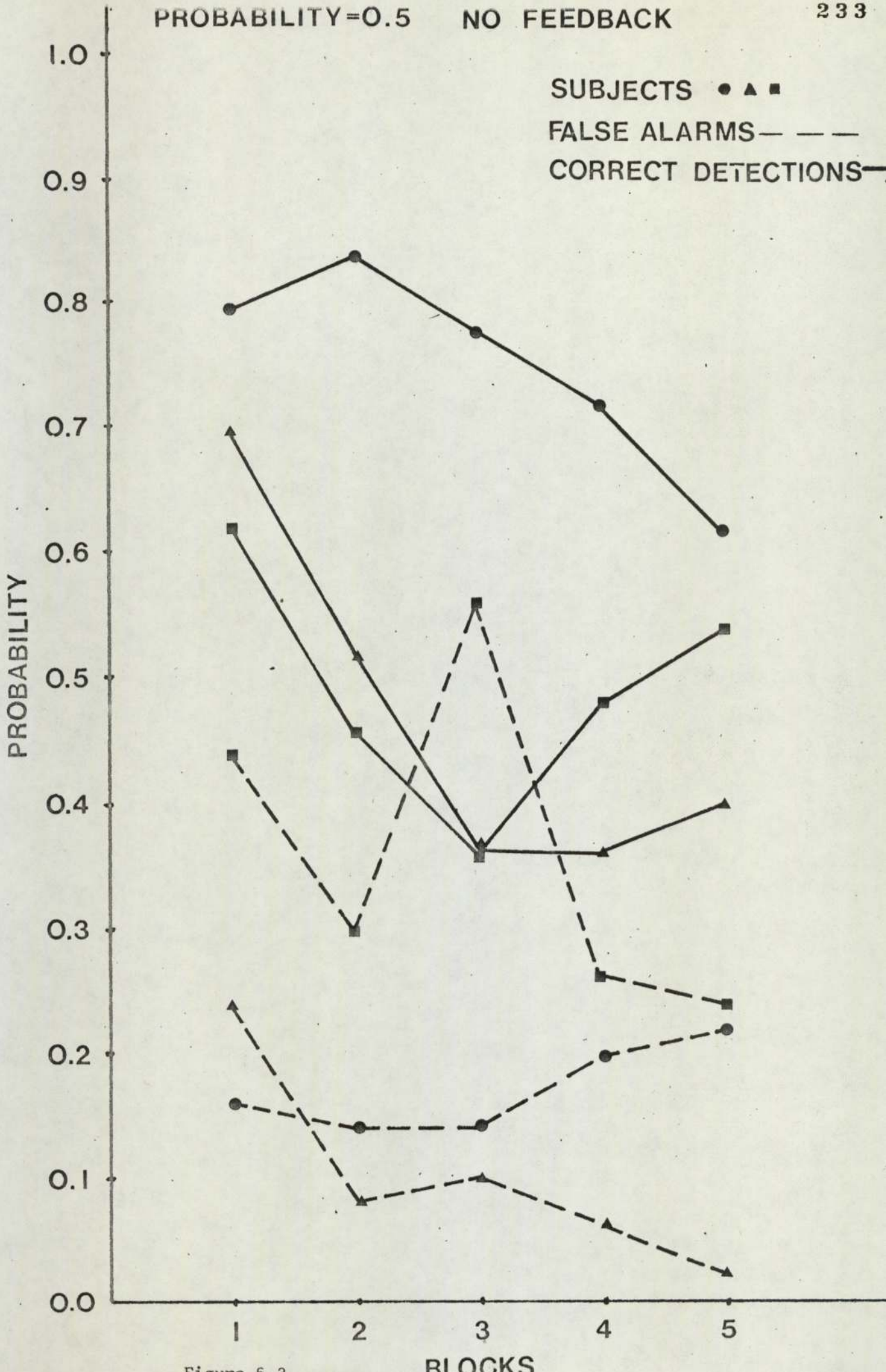

PROBABILITY=0.5 NO FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS— — —
CORRECT DETECTIONS——
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
1
2
3
4
5
BLOCKS

Figure 6.3

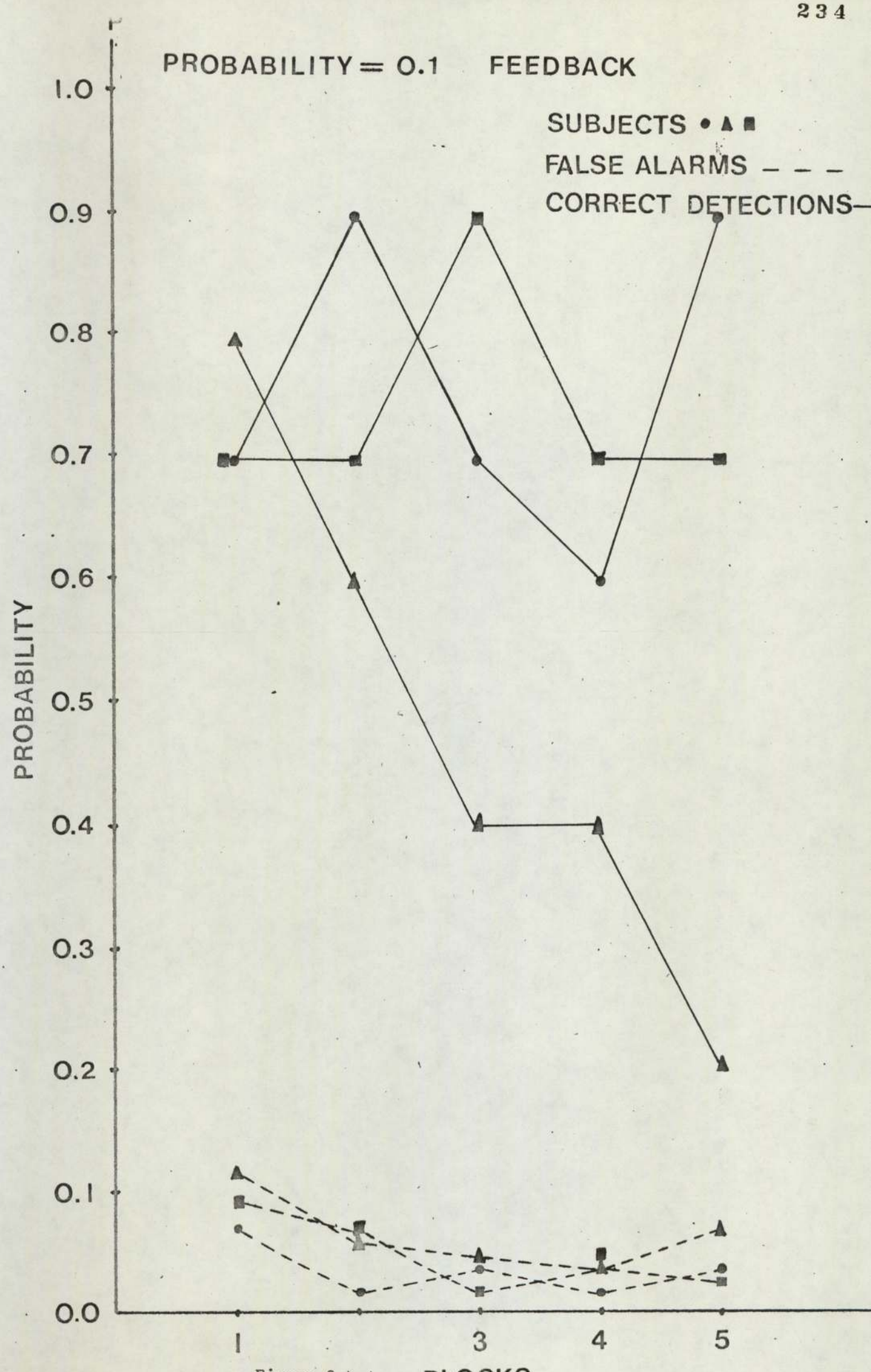

PROBABILITY = 0.1 FEEDBACK
SUBJECTS • ▲ ■
FALSE ALARMS – – –
CORRECT DETECTIONS—
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
1
3
4
5
BLOCKS

Figure 6.4

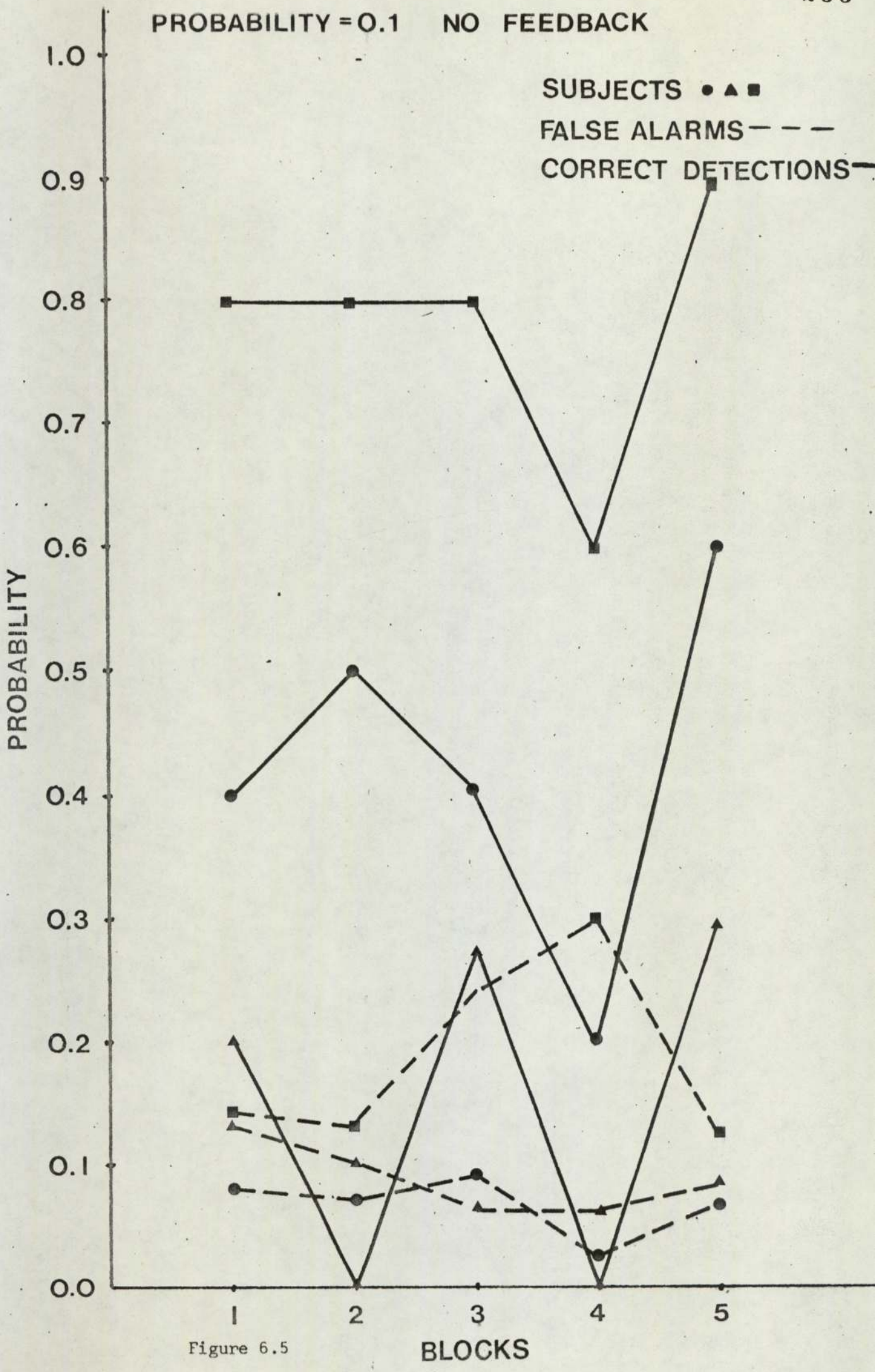

PROBABILITY = 0.1 NO FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS - - -
CORRECT DETECTIONS
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
1
2
3
4
5
BLOCKS

Figure 6.5

Considering the SDT measures, no significant differences are found for any of the sensitivity indices AH, P(A), d', de, or the Pollack-Norman index, under any of the experimental conditions. There are significant differences in bias as measured by log beta, between blocks ($p < 0.01$) and a significant feedback by probability condition interaction ($p < 0.05$). A similar pattern of results is observed for the Hodos-Grier bias index, the corresponding significance levels being $p < 0.05$ in each case.

Comparison of means for blocks shows that log beta increases significantly between the first and fourth ($p < 0.01$) and first and fifth ($p < 0.05$) blocks. This result accounts for the high false alarm rate observed during the first block. The correct detection probability was also highest during this block, as would be expected from the low log beta during this block.

The significant feedback condition x probability condition interaction is analysed in Table 6.1 below:

probability conditions	feedback conditions f/b	no f/b
P = 0.5	0.15	0.41
P = 0.1	1.35	0.87

Table 6.1 Comparison of probability and feedback condition means for log beta

A simple main effects analysis shows that log beta is significantly greater under the p = 0.1 condition, but only if knowledge of results is available.

The next variable analysed, the subjects' points score, is not strictly speaking a signal detection index, but is related to signal detection performance. It can be regarded as representing the value of the inspectors' performance to the inspection system, given the particular payoffs specified for the various degrees of response category which were described earlier.

There are highly significant differences between the probability conditions, the mean score for the low probability condition being very much greater than for the P = 0.5 condition ($p < 0.01$).

6.5.2 Latency data

The log transform of the total latency indicates significant differences between blocks. Multiple comparisons show that the first three blocks have a longer total latency than the last two ($p < 0.05$). The total latency is also significantly shorter ($p < 0.05$) under the P = 0.1 condition that the P = 0.5 condition. Analogous results are found with the correct rejection latencies, the significances being $p < 0.01$ and $p < 0.05$ in this case.

No significant effects are obtained for the correct detection and omission latencies. The false alarm latency however is very significantly longer for the P = 0.5 condition that the P = 0.1 condition ($p < 0.001$).

6.6 Discussion

The variables of particular interest in this experiment are the SDT measures of sensitivity and bias. The lack of any significant

effects for any of the sensitivity measures was not surprising, since none of the experimental factors was expected to affect sensitivity. The significant increase in beta found over blocks is unexpected, and closely resembles that found in vigilance tasks. In fact, some form of time related performance decrement is the only obvious explanation for this effect. If it occurred only with the P = 0.1 condition we could explain it in terms of the subjects attempting to adjust their criteria from the unbiased one appropriate to the P = 0.5 condition to a more stringent one for the lower probability condition. There is no obvious reason why vigilance effects should occur in view of the fact that frequent rests were provided. The effect also occurs in conditions where feedback is provided, which is contrary to almost all vigilance studies in which KR has been given. Additionally the subjects gave no verbal intimation that they found the experiment particularly boring. A possible explanation is in terms of the particular circumstances of the experiment.

The subjects were aware that the experiments were concerned with their ability to adjust to changes in defect probability, and may therefore have suspected that a change in defect density would occur within the session. There is some support for this notion in that the criterion, after reaching a maximum in session 4, declines again during session 5. Possibly the subjects had decided by that point that no change was going to take place.

The analysis of the feedback x probability interaction is of particular interest. It suggests that subjects are able to adjust their criteria in the optimum direction predicted by SDT, but only if they are provided with knowledge of results. The result suggests that in

this experiment, external sources of information were more important than evidence gained from the task itself, in shaping the criterion. It is difficult to compare this result directly with the Sims study, because no measurement of the criterion was obtained in that situation. In the current experiment we have no direct estimate of the subjective probability that the subject was employing. If we make the assumption that a change in subjective probability was translated directly into a change in criterion, in the present study, then the result is in direct contrast to the Sims study. The difference can be accounted for by the relative difficulty of discrimination of defects in the two studies. Although no measure of discrimination performance was given by Sims, it is clear from her description of the task, that it was considerably easier than the one under consideration. It is likely that Sims' task would provide far more intrinsic feedback for the subject than the present one.

These considerations lend support to the idea that the ability of a subject to adjust his criterion to the optimum for a particular defect probability is a function of the amount of information available on the defect density, whether it be from within the task or from external sources. Where the implicit task information is highly reliable, i.e. the signals are readily discriminable, the Sims study suggests that external evidence on the defect density is redundant.

The fact that the subjects obtained a higher net score under the $P = 0.1$ condition can be accounted for by the fact that in spite of extensive training, it was easier for the subjects to recognize a non-signal than a signal, as indicated by the greater variance of the

the signal distribution obtained from most of the ROC fits. This allowed them to make a high proportion of 'definite' (i.e. extreme rating category) responses when most of the trials were known to be non-signals. It is also possible that the subjects were adopting a nearly pure strategy with the low probability signal condition, as proposed by the Siegel-Goldstein probability learning hypothesis discussed earlier. Most of the subjects were aware that they could obtain a high payoff by responding 'definitely non-signal' on a high proportion of the trials, regardless of the evidence from the task. It seems possible that in a difficult discrimination task, with no a priori information as to defect probability, the criterion will be set according to the subjective probability generated by considerations discussed in the probability learning literature.

Most of the latency results can be accounted for by the SDT model. If the situation is as set out in Figure 6.6 below, with the signal variance greater than the noise variance, then with the assumptions set out in Chapter 4, the following predictions can be made.

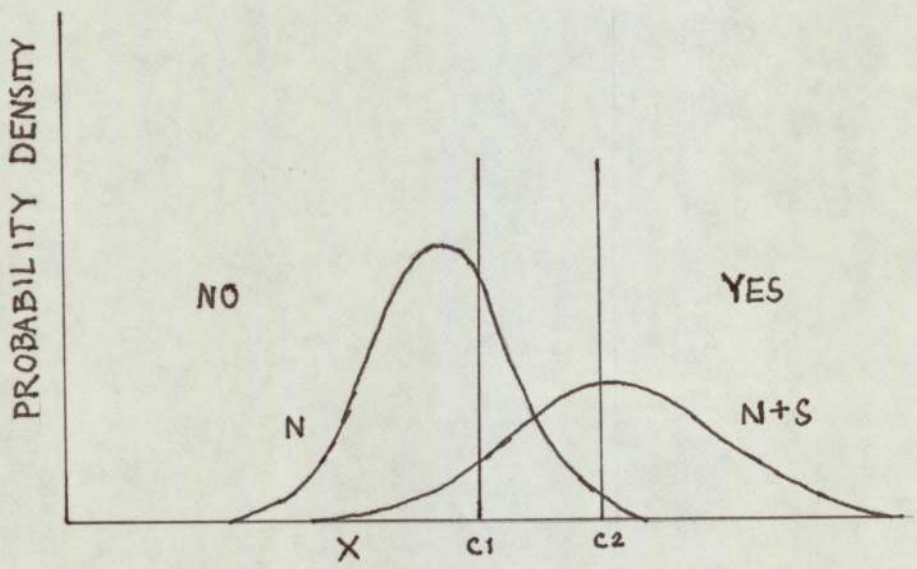


Figure 6.6 Response latencies under the unequal variance SDT model.

If the criterion moves from position C1 to a more stringent position C2, we would expect NO response latencies to decrease, since these would on average be distributed further from the criterion. In the equal variance situation, YES latencies would increase, because they would be distributed closer to the criterion and hence represent more difficult decisions. However, with the large signal variance situation illustrated above, the changes in both correct detection and omission latencies might be expected to be non-significant.

The results generally bear out these predictions. The criterion increases significantly with time on task and is overall greater for the P = 0.1 condition than the P = 0.5 condition. The correct rejection (i.e. correct 'no signal' decision) latency decreases significantly with both increases in criterion. The omission and correct detection latencies show no significant changes. The false alarm latencies are anomalous, since they should increase with the increase in log beta for the P = 0.1 condition, whereas they actually show a significant decrease compared with the P = 0.5 condition. This result can be seen as a consequence of the subjects habituating to a particular mode of response. In the P = 0.1 condition, since most of the responses were 'non signal' and the non-signal stimuli were more recognizable as such than the signal trials, the rapid high confidence responses made to non-signals probably tended to spread to other categories of response. This effect would be likely to dominate the effects of the greater bias. In the P = 0.5 condition, the equiprobable signal occurrence would prevent a style of response developing which was dominated by one type of trial.

The total response latency results are simply a reflection of their predominant component, the correct rejection response latencies.

6.7 Results - Experiment 5

The correct detection and false alarm probabilities for the blocks and experimental conditions are set out in Figures 6.7 to 6.9. The transformed correct detection probability showed significant differences between blocks ($p < 0.05$), the first two blocks differing significantly from the remainder. The transformed false alarm probability showed significant effects between blocks, with a highly significant blocks x conditions interaction, ($p < 0.001$). This interaction will not be analysed in detail at this point, since it is a direct consequence of the criterion effects to be discussed subsequently.

Considering the sensitivity indices, the results are somewhat ambiguous. The analysis for the Altham-Hammerton, Pollack-Norman indices and d' all indicate a significant blocks x conditions interaction, whereas d_e', and P(A) show no significant effects. This anomaly will be discussed in more detail subsequently. Log beta gives significant effects for blocks, conditions, and a highly significant condition by blocks interaction ($p < 0.001$). The simple main effects analysis is set out in Table 6.2 below:

Source	SS	df	MS	F	Significance
B at C1, no warning feedback	0.094	4	0.0235	0.28	ns
B at C2, warning, feedback	1.55	4	0.39	4.59	$p < 0.05$
B at C3, warning, no feedback	8.09	4	2.022	23.79	$p < 0.01$
error	1.355	16	0.0847		

Table 6.2 Simple main effects analysis of BxC interaction for log beta

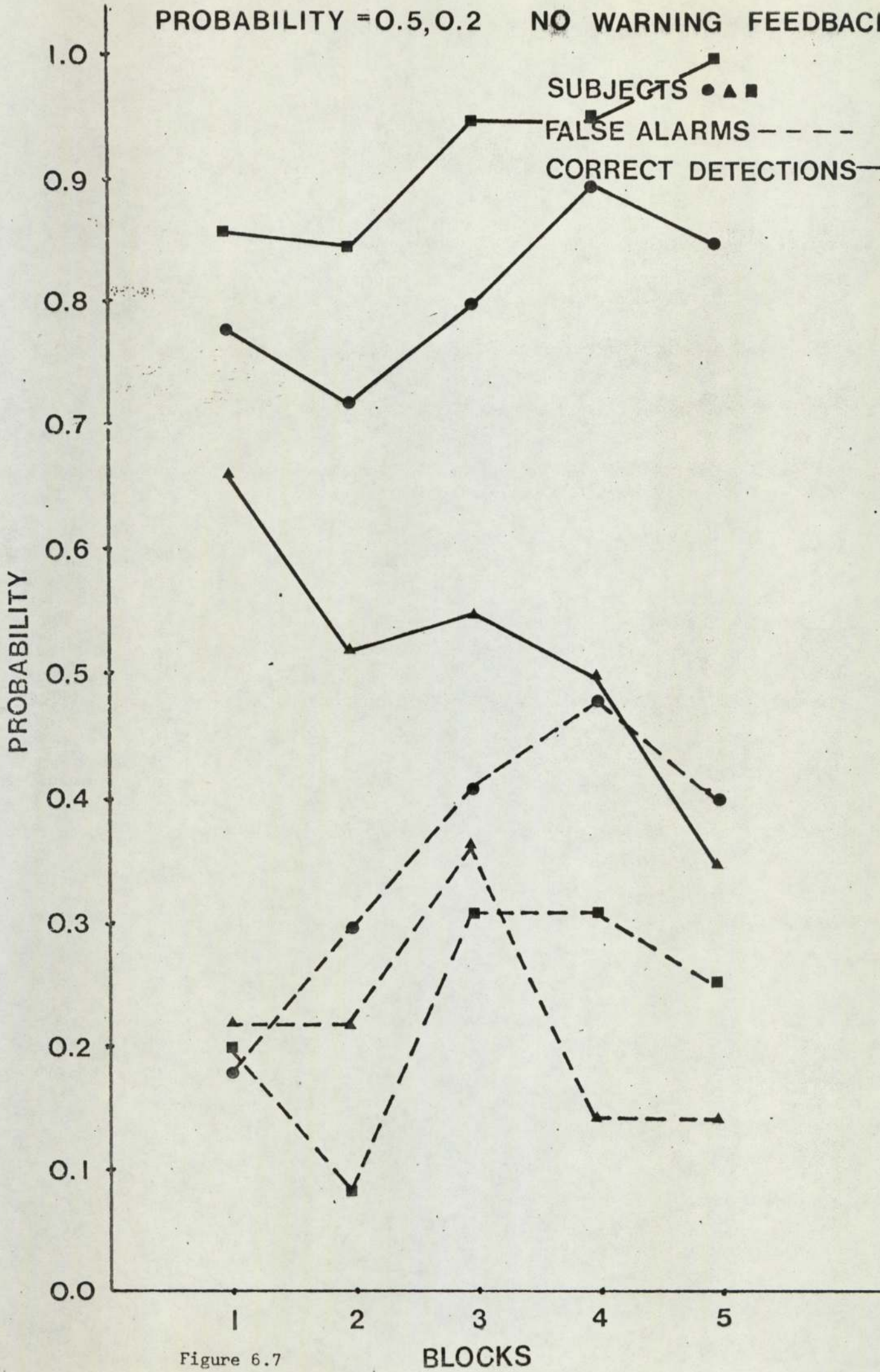

PROBABILITY =0.5, 0.2 NO WARNING FEEDBAC
SUBJECTS ● ▲ ■
FALSE ALARMS - - - -
CORRECT DETECTIONS
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
1
2
3
4
5
BLOCKS

Figure 6.7

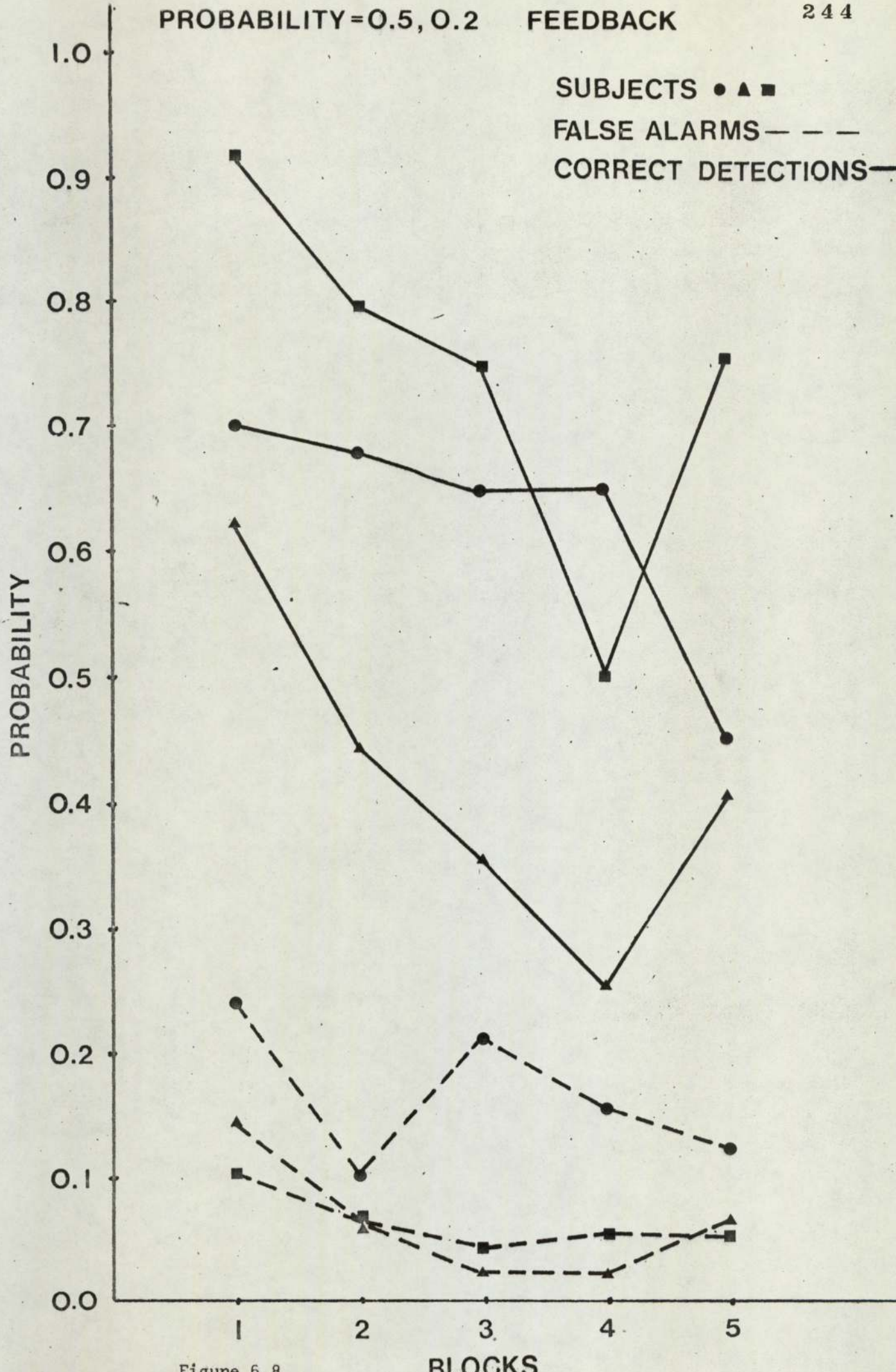

PROBABILITY=0.5, 0.2 FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS — — —
CORRECT DETECTIONS ——
PROBABILITY
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
BLOCKS

Figure 6.8

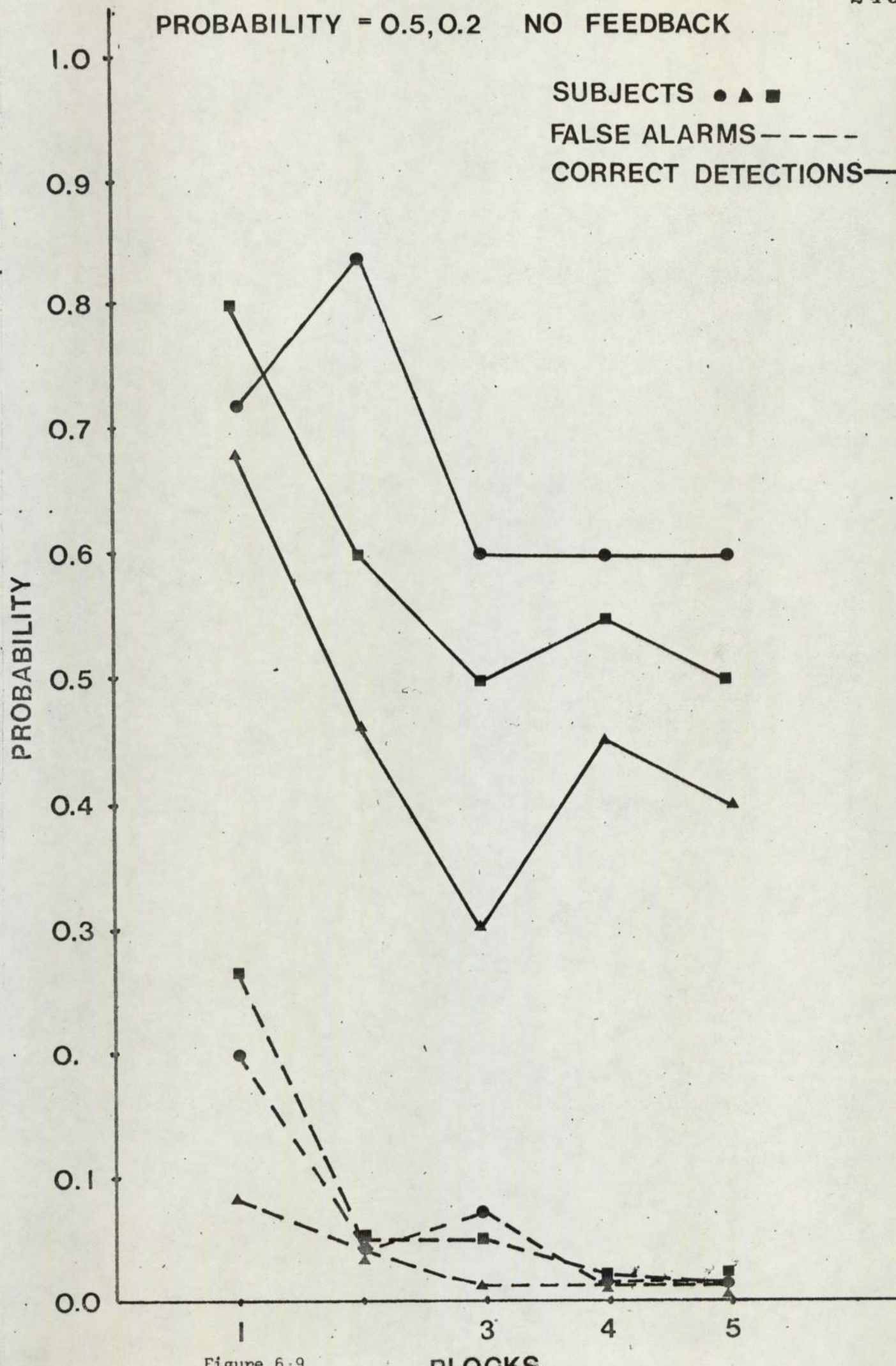

PROBABILITY = 0.5, 0.2 NO FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS - - - -
CORRECT DETECTIONS —
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.
0.1
0.0
PROBABILITY
1
3
4
5
BLOCKS

Figure 6.9

The simple main effects analysis indicates that there are significant differences between blocks for the two conditions in which warnings of change were given, but not for the no warning condition. Tukey multiple comparisons tests for C2 and C3 indicate that there are significant differences between blocks 1 and 2 combined together, and blocks 4, 5 and 6 combined, ($p < 0.05$, $p < 0.01$).

The latency results will not be considered in this experiment, since they do not provide any additional insights into the variables of interest.

6.8 Discussion

The first question to consider is whether any changes in sensitivity have occurred as a result of the within session change in probability. In view of the fact that both P(A) and de, which are known to be independent of the nature of the underlying variances in SDT, show no significant changes, the validity of the changes suggested by the other indices must be viewed with scepticism. This is particularly the case in view of the fact that sensitivity measures are not normally affected by signal probability changes. It seems likely, therefore, that the apparent change in sensitivity is an artefact in this instance.

The detailed results for the effects of the within session changes in probability on log beta provide very interesting insights. It appears that prior warning of a probability change is more effective in producing an appropriate criterion change than the provision of full feedback from the task. Also, when a warning is given, the subjects

are able to adjust their criterion in the appropriate direction, even in the absence of feedback. An interesting feature of the no feedback, warning condition is that the criterion continues to increase to the end of the session, whereas with feedback it remains fairly stable. The KR information can be regarded as encouraging a stable, moderately high criterion after the probability change, because even given a precise knowledge of the signal probability, from the KR, the subjects will always adjust their criteria in a conservative manner. Where KR is not provided, the subjects are likely to assume that the defect probability has changed to an even lower level than is actually the case. Adopting a more stringent criterion will lead to fewer defects being detected which will in turn produce an even higher criterion, as a consequence of the perceived signal probability.

6.9 Conclusions

These experiments have provided general support for the hypothesis that the degree of criterial change produced by a subject in a changing defect probability situation depends on the amount of information available on the probability from external sources, and from the task itself. The relative weighting of the internal and external sources of information would appear to be a function of the reliability of these sources. For between session changes in defect probability, even when the subject was aware that a reduced incidence of defects could be expected, significant criterion changes only occurred when knowledge of results was available. This could be explained by the fact that although the subject knew that the defect incidence was less, he did not know to what degree he needed to alter his criterion. The precise evidence obtained via feedback provided this information and encouraged the appropriate criterion change. In a situation of

high uncertainty, as with the within session probability change experiment, the importance of prior information is seen in enabling a criterion change to occur. Even when the subjects had full KR, they were unable to modify their criteria unless they had been alerted to the possibility of change. The fact that warning of change seems to be a more important factor in criterion adjustment, than feedback from the task, where within session probability changes occur, suggests further investigation is needed in this area. In both experiments the importance of external evidence of some form can be explained by the difficult nature of the discrimination task, and hence the unreliability of evidence from this source.

No attempt has been made to analyse the degree to which the subjects were able to achieve the actual optimum beta predicted by SDT. This is because, as shown by Green and Swets (1966) p. 92, there are a range of beta values about the optimum which will achieve a high proportion of the theoretical payoffs. This does not however, invalidate the practical importance of the inspector at least being able to modify his criterion in the direction appropriate to the current defect probability. As discussed earlier, an inability to do this can produce a very inefficient quality control system, if defect probabilities are liable to fluctuation.

Although these experiments were exploratory in nature and utilized a laboratory situation and a limited number of subjects, the results seem sufficiently interesting to suggest further work in an industrial context.

CHAPTER 7 TRAINING AND SELECTION FOR INSPECTION

7.0 INTRODUCTION

In this chapter experimental work arising from some of the issues discussed in the review of perceptual training techniques and selection methods in Chapter 3 will be presented. In the final phase of the study, two experiments were performed in which various types of training for perceptual skills were investigated. The performance measures from these experiments were used in conjunction with a number of tests of various cognitive skills, in order to establish the usefulness of such tests for the purposes of selecting individuals for inspection work.

7.1 Theoretical considerations in perceptual learning

It is useful at this point to briefly recapitulate the conclusions of the earlier review in the area of perceptual training.

The two basic methods of training that had been used to train for perceptual skills were cuing and knowledge of results (KR). Enhancement of signal detection ability by cuing was seen as a result of the simple paired contiguity in time of a signal and its name. This form of training promotes perceptual learning because it provides the subject with the maximum information concerning both the characteristics of the signal and its distribution in time.

It was suggested that the other form of perceptual training, KR, trains recognition skills primarily via the reinforcement of a simple S-R link. An important aspect of KR is its motivational effect in maintaining arousal during prolonged sessions. Improvements in performance using both techniques can be ascribed partly motivational effects,

partly to learning the characteristics of the signal, and partly to an increased knowledge of the statistical distribution of the signals. The applicability of the SDT approach to these latter two areas is clear. Increased knowledge of the signal characteristics implies an increase in d', whereas knowledge of the signal distribution produces an appropriate expectancy. This has been analysed in previous chapters as an accurate subjective estimate of the signal probability, leading to an optimization of the response bias, as measured by beta.

Early work by Annett and Wiener suggested that cuing enhanced d', whereas KR improved detection performance at the expense of increased false alarms mainly by producing a more lax criterion. Subsequently it was found that the latter result was to some extent an artefact of a free response situation, in that subjects appear to make a higher number of affirmative responses, and hence produce an apparently lowered criterion, in an attempt to gain more information about the signals and their distribution. Recent studies, e.g. Annett (1971) have suggested that in a situation where a subject has to respond to a series of successive trials, the differences between the two training techniques are slight, because they both provide essentially the same information.

Another aspect of perceptual skills training that was discussed was Wallis' (1963) suggestion that learning to recognize complex signals was best accomplished using an analytic-synthetic approach where the salient features of the stimulus are first learnt separately and then synthesized into a wholistic 'Gestalt'. The importance of the withdrawal of cues at an appropriate time was mentioned in this study, and the whole question of the dangers of a subject becoming dependent on cues or KR, to the detriment of learning, was further

emphasized by Abrams and Cook (1971). These workers employed a gradual reduction in the amount of KR to reduce dependence on feedback, and suggested that this technique facilitated the development of the internal referents necessary for identification skills.

From the point of view of its applicability to inspection, much of the existing work on perceptual skills training suffers from several disadvantages. The stimuli employed have usually been unrepresentative of those found in inspection tasks, and much of the work has been in areas such as sonar where auditory signals are employed. These factors were considered in the design of the experiments described in the following sections.

Experimental work - general

In view of the points discussed earlier it was decided to attempt to simulate the critical features of a real inspection task in both experimental studies. The task chosen was one that had already been analysed in some detail, the Ilford film inspection task described in Chapter 5. It will be recalled that considerable problems had been experienced in scoring experiments with the task in its original form, because of the difficulty in resolving the position of the defects on the film with a sufficient degree of accuracy to prevent ambiguity. It was therefore decided to alter the task in the following manner. A particular type of defect, known as 'insensitive spots' was chosen as being representative of those encountered by the inspectors in their everyday work. This defect type consists of a number of tiny spots on the film where the coating has not adhered adequately to the base. Because the film coating is

not darkened at these points, the appearance of the defects when the film is projected is that of a number of small points of light on the otherwise grey background of the correctly exposed film. Actual samples of defects were cut from films and turned into a series of slides. Similarly samples of 'perfect' film from other parts of the same roll were made into 'non defect' slides. This was done to ensure that the background density for both types of slide was of the same apparent brightness in both cases. These slides were then used in conjunction with the equipment described in Chapter 6, to simulate the appearance on the screen during film examination of a single frame containing insensitive spot defects. The speed of the shutter was adjusted to give a viewing time for each slide of 1/15th of a second. This was the closest speed available to the 1/18th second employed in the actual task.

A continuous overall background illumination of the screen was provided by a second Kodak carousel projector and was -0.4 log foot lamberts as measured by an SEI photometer. When a slide was projected on to the screen, the brightness increased to 0.2 foot lamberts, which approximated to that found in the actual inspection situation.

The basic operation of the equipment was similar to that described in Chapter 6. At the start of the session the computer operated the slide projector magazine to load the first slide, and then the Compur magnetic shutter to provide a presentation on the screen of 0.06 seconds. A software clock within the computer was started at the same time and this enabled the subjects response latency to be measured. The response arrangements were as in previous experiments. A series of six buttons was provided to enable the subject to make

the rating responses described previously. As before, a symmetrical payoff scheme was assigned to the responses such that 1, 2 or 3 points were gained for correct responses of increasing certainty and a similar number of points were subtracted for the corresponding incorrect responses. The computer outputted the response type and time on paper tape after each response.

The detailed arrangements for each experiment will be described separately in the following sections.

7.1 Experiment 6 - comparison of cuing and feedback techniques

The first experiment was designed to investigate the effectiveness of cuing versus knowledge of results in the simulated inspection task. To do this, three conditions were investigated. These were cuing, feedback and a control condition. All three groups were given a short practice session to familiarize them with the equipment and the method of response. The scoring scheme was carefully explained to them and twenty practice trials were given, at the end of which the individual slides were discussed to indicate what were the characteristics of the defective as compared with the non-defective slides. It was pointed out that even the 'perfect' slides contained many configurations which were confusable with defects and that they were only to categorize as defective those slides which contained 'insensitive spots'.

The subjects were assigned at random to one of the three experimental groups and all subjects first received five hundred trials as the pre-training phase. As in experiment 5, attempts were made to

minimize vigilance effects by providing the subjects with a three minute rest after each 100 trials. Low level white noise was played through headphones throughout the experimental sessions, but subjects were allowed to remove these and converse with the experimenter during the breaks.

For the training sessions, the control group received a further 100 trials similar to the pre-training session. Before each slide was presented to the cuing group, the computer typed a message on the teletype, which was visible to the subject, of the form: 'the next trial will be a defect' or 'the next trial will be a signal'. The subjects were told to read the message, and to use it to ready themselves to observe the characteristic features of the defect or non-defect. They were also told to make an appropriate response after the slide had been presented. For the KR group, a message was typed by the computer after each response and consisted of one of the following four types:

1. You have just missed a defect.
2. That was a false alarm.
3. That was a correct detection of a defect.
4. That was a correct detection of a non-defect.

The final post training session was a repeat of the first session, in which five hundred trials were given.

Twenty one subjects were employed in the experiment, seven being assigned randomly to each condition. They were all undergraduate students from various disciplines and were paid 50 pence per session. The pre-session and training session were administered on the same

day, and the post session, as far as practicable, at the same time on the following day. The probability of a defect remained constant within blocks of a hundred trials throughout the experiment and was 0.2.

Analysis of the results

As described for experiment 5, the paper tape output from the PDP-8 was transformed into magnetic tape files on the PDP-15 computer and the various performance measures employed in experiments up to this point calculated. The statistical design employed was a 'split plot' experiment (Keppel (1973) p.433). The between subjects factor was the training conditions, and there were repeated measurements on blocks of 100 responses within the 500 blocks for each subject. The quantity entered into the analysis of variance was the difference between the pre and post training sessions for the various performance measures. This method of analysis is generally regarded as being superior to including before and after training as a separate factor, because it compensates for the differing initial performances of the subjects.

Where the summary data for all 500 responses were used in the analysis, the design becomes a simple one-way completely randomized analysis of variance.

The rating data from the performance summaries for all 500 responses of the pre and post training data were run through the Grey-Morgan ROC curve fitting program to obtain the variance ratio, the sensitivity index d'_e and the five beta values employed corresponding to the six rating categories utilized by the subjects. A fit was obtained for 41 of the 42 sets of data, although in some cases signal to noise

variance ratios of 23 had to be assumed by the program to carry out the fitting procedure. This is not surprising in view of the fact that the first 500 responses the subjects were unfamiliar with the characteristics of the signal. The results in general suggest that the SDT model applies, however.

Results

All analyses of variance will be found in Appendix A, together with summary data.

The first analysis was conducted on the SDT information output from the Grey-Morgan program. Only the summary data for all 500 responses in the pre and post training sessions was considered. The sigma ratio was first analysed as an index of sensitivity change, because one might expect the variance of the signal distribution to decrease as the subject learns the characteristics of the signal. Such a decrease was only found for the cuing condition, but the differences between the conditions were non-significant. A similar result was obtained for d_e, the corrected sensitivity index. The next variables considered were the log beta values for each cutoff corresponding to the rating categories employed by the subjects. The only significant effect was a significant difference between subjects for cutoff 4 ($p < 0.05$).

For the remaining analysis, the separate blocks of 100 responses were considered as repeated levels of the same factor for each subject. No significant effects for any of the sensitivity measures were found apart from d', which indicated a significant change between blocks.

In view of the unreliability of d' in the unequal variance situation, and the lack of corroboration from any of the other sensitivity measures, this result must be regarded as artefact.

Considering the bias indices, a significant difference between training conditions was obtained for beta ($p < 0.05$). Tukey's test indicates a significant difference from the control, but not from the other training condition. The Hodos-Grier bias index indicates significant differences between blocks, ($p < 0.05$) the change in bias as a result of training being significantly greater for block 1 compared with blocks 3 and 5 (Tukey test, $P < 0.05$). The only other significant effect obtained was significant differences between blocks in the changes in false alarm probability due to training. The false alarm probability declined significantly for block 1 compared with blocks 3, 4 and 5.

Discussion

The overall absence of significant changes in sensitivity produced by either of the training conditions could be due to the relatively small number of training trials employed. The increases in bias observed as a result of training are in the opposite direction to those found with KR in free responding situations. The direction of change was in the correct direction in this experiment, since the final value of beta was closer to the optimum than the original. We can conclude therefore, that cuing seems to promote a greater change in bias in the correct direction than KR, although the difference is not statistically significant in this experiment.

The changes in bias indicated by the Hodos-Grier index, and reflected by the false alarm probability changes, probably occur because during the earlier blocks of the pre-training session, the subjects were utilizing a more inappropriately low criterion than during the later blocks, as a result of their initially limited knowledge of the signal distribution. The effects of the training in increasing bias would therefore be more marked for the earlier than for the later blocks.

7.2 Experiment 7 - further training techniques

Although the last experiment yielded some interesting results it was felt that a wider variety of training techniques should be investigated, particularly from the point of view of improving performance in situations where the probability of a defect occurring varied within sessions. It was also felt that a greater amount of training might produce changes in sensitivity which had not been revealed by the first experiment.

The first two conditions that were used in the final experiment were feedback alone and cuing alone. These conditions were included to replicate the first experiment, but to provide a greater number of training trials. The next condition was a combination of cuing and feedback, in which alternate training trials were either cued or KR was supplied after the response. The object of this technique was to investigate whether cuing and KR utilized together were more effective than either used alone. The fourth condition provided an even greater amount of information to the subject by giving him alternate cuing and KR as in condition 3, and also supplying him with a summary of his performance in terms of correct detections, false alarms

etc. at the end of each block of 100 responses. The final condition was similar to condition 1, the feedback only situation, but in addition to the feedback information, the subject was provided with the points score he had obtained as a result of his responses, at the end of each block of 100 trials. This condition was included to establish if the provision of information concerning the payoffs associated with the subject's response strategy would enable him to optimize this strategy more effectively than if feedback alone were provided.

In addition to the five training conditions, two other factors were included in the experiment. As described in the introductory section of this chapter, and in the full review in Chapter 3, one of the difficulties associated with training for perceptual skills is that subjects may become dependent on the presence of the particular training aid to maintain performance. This may prevent the performance improvement obtained during training being transferred to post training sessions. Abrams and Cook (op. cit.), had found that gradually reducing the amount of feedback during training improved the retention of any detection skills acquired. It seemed worth investigating if this effect was also obtained with cuing. All the training conditions described earlier were therefore performed under two levels of this factor. In the one case, feedback and/or cuing was provided on every trial. At the other level of this factor, the amount of information provided by cuing or feedback was gradually reduced throughout the session.

The other question of interest which was investigated in the experiment was whether the experience of a varying probability of defect occurrence during training would facilitate performance in post

training sessions where the within session defect probability also varied. It was hypothesized that the ability to adjust the criterion optimally in a changing defect probability situation would be enhanced if the subject had prior experience of a range of defect probabilities. The training conditions were therefore performed both with a constant level of defects and also with a defect incidence that varied from block to block during the training sessions.

Procedural details

The equipment utilized was essentially the same as for experiment 6. Certain changes were made, however, in the way that the feedback and cuing information was conveyed to the subject. It was found that during the first training experiment, the use of the teletype to type messages to the subjects had considerably slowed down the experiment. Also, subjects become irritated at having to read a long message before or after each response. An alternative means of conveying information was therefore designed and built. This consisted of four coloured lights on a block painted panel. The lights were operated by the computer to provide the appropriate cuing or feedback information. For the cuing conditions a red light was illuminated prior to the presentation of a defect, and a green light before each non-defect slide. For the KR condition four lights were employed. A green light was used to indicate a correct detection, and white, red and yellow lights were illuminated following correct rejections, false alarms, and missed signals respectively.

Each subject performed six sessions, each of 500 responses in blocks of a 100 with rest periods between each block, as in experiment 6.

Prior to session 1, a short familiarization session was given as before. Session one was the pre-training session, in which the defect probability remained constant at 0.2 within blocks of 100 responses. The second session, which was performed on the same day as the first session, was designed to investigate detection performance in a varying probability environment. The defect probabilities for blocks 1 to 5 respectively were : 0.15, 0.15, 0.40, 0.20, 0.10. The subjects received prior warning that the probabilities would vary from session to session. The next two sessions, which occurred the following day, were both training sessions. The training sessions were separated by a break of at least 45 minutes to minimize any effects of fatigue. In the variable probability training sessions, the subjects were told the probability prior to each block of 100 responses. The probabilities were chosen to be representative of those found in the test sessions.

On the day following the training the two final sessions were administered. The first of these was identical to session one, and the second similar to session 2, but with a different sequence of defect probabilities for each block, i.e. 0.25, 0.15, 0.20, 0.10 and 0.30. The differing sequence was chosen because the event probabilities employed in session one had also been utilized during the training period, and there was some possibility that the subjects would learn the sequence of probabilities. It will be noted that the probabilities used for each block of 100 responses, in both the pre and post training variable probability sessions, summed to give the same overall probability of 0.2 as in the fixed probability sessions. This was done to ensure that any performance differences between the fixed and varying probability sessions was a result of the variability itself rather than because of an overall difference in defect

probability. Prior to session 6, the subjects were told that the defect probability would vary, but that it would be different than during the training sessions.

The subjects were recruited from advertisements in schools and their ages ranged from 16 to 19. All had normal or corrected vision.

Statistical design

The basic structure of the experiment is set out in Figure 7.1 below.

training conditions	constant KR and cuing		fading KR and cuing	
	fixed defect probability	varying defect probability	fixed defect probability	varying defect probability
feedback alone cuing alone feedback + cuing feedback + cuing + summary feedback + score	B1 B2 B3 B4 B5			

Figure 7.1 Experimental design

The results were intended to be analysed as a four-way analysis of variance. The factors were training conditions, constant or fading cuing or KR, fixed or varying defect probability, and a repeated factor for each subject of 5 blocks of a hundred responses within sessions. The scores to be entered in the analysis of variance were the differences between pre and post training in the case of sessions 1 and 5. Sessions 2 and 6 were to be analysed separately to assess the degree to which the subject was able to modify his criterion to take into account the changing probabilities. It was originally

intended to use at least two subjects within each cell of the analysis. Unfortunately circumstances beyond the control of the author meant that the original design had to be modified. There was a delay of six months in obtaining essential interface circuitry from the computer manufacturers and this, together with other unforeseen delays, meant that the design had to be changed after the experiment had begun.

It was decided to use a strategy discussed in Kirk (1968) p.227 and Winer (1962) p.216 and p.267. These authors suggest that by assigning one subject only to each cell of a completely randomized analysis of variance, an estimate of the error mean squares necessary to conduct the analysis of variance can be obtained from the highest order interaction, by making certain assumptions which will be discussed in the next section.

The use of the additive model in the analysis of variance

If we consider the summary data for each session and do not analyse the within subject variable of blocks within sessions, the experiment can be regarded as a three-way analysis of variance, with training types, fixed or varying probability and fixed or fading feedback or cuing as the factors.

There are two alternative models for the analysis of variance that can be postulated. Using the usual notation for factorial experiments, these are:

1. $X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \epsilon_{ijk}$
2. $X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijk}$

Model 1 assumes that all sources of variation other than the main effects and first order interactions are part of the experimental error ϵ_{ijk}. In model 2 a second order interaction is assumed to occur. This interaction term may be considered as a measure of the non-additivity of the main effects and first order interactions, i.e. the extent to which the observation cannot be predicted from a knowledge of three main effects and interactions, and the experimental error. Model 1 is known as the additive model and model 2 the nonadditive model, using Winer's terminology.

On a priori grounds, in the present experiment, there did not seem to be any reason to suppose that the second order interaction would be significant. A test is available, however, due to Tukey (1949) which enables one to decide which of the models is appropriate. If the first model is applicable, the variation due to sources other than main effects and first order interactions can be regarded as consisting of two components. The first of these is that due to the linear X linear component of the ABC interaction, and measures the degree of nonadditivity. The remainder is known as the balance. If the additivity component is significantly larger than the balance, model 1 is rejected.

An F test that will test this hypothesis is given by:

$$F = \frac{MS \text{ nonadditive}}{MS \text{ balance}}$$

The computational procedures necessary to obtain this quantity for a three way analysis of variance are given in Winer (1962) p.267. The F test is usually made at a high critical value ($p < 0.25$) in

order to reduce the probability of a type 2 error. If the F value obtained does not exceed the critical value, then the additive model is appropriate for the analysis, and the error term used is the ABC interaction. A program was written to perform the calculations, and used to provide a nonadditive F test for each block of data prior to performing the analysis of variance. Twenty student subjects were assigned randomly to the experimental conditions.

Results - fixed probability sessions

The assumption of the additive model was upheld with most of the data analysed. In the cases where the additive model was rejected it was sometimes possible to obtain a transformation of the variable concerned which did fit the model.

The first data analysed were the differences between sessions 1 and 5, the fixed probability sessions, for a range of variables. The Grey Morgan programme was used to check the validity of the SDT assumptions and to produce the first variables to be analysed, the sigma ratios, the sensitivity index d'_e and the beta values corresponding to the rating categories employed.

The only significant effects obtained were for the beta value corresponding to cutoff 2. Unfortunately, this was one of the cases for which the nonadditive hypothesis was rejected, and hence no conclusions regarding these effects could be drawn.

The analysis was then carried out on the range of variables considered in previous experiments, including the latency measures, the

parametric and nonparametric indices of bias and the net changes in payoffs obtained by the subjects as a result of the training given. The actual inputs to the analysis of variance were the differences between these quantities for the summary data for the whole of sessions 1 and 5. No significant effects were obtained for any of the resulting 22 analyses of variance.

It was therefore decided to repeat these analyses for each separate block of 100 responses. The latency data were not included in these analyses. For the first block of data significant effects were obtained for the changes in payoffs and for d'. In both of these cases however, the additive model was rejected by the F test. A significant interaction was obtained for the arcsine transformation of the false alarm latency, between the fixed or variable probability during training factor and the training conditions ($p < 0.05$). Simple main effects analysis of this interaction indicated that the feedback alone training condition produced a significant reduction in false alarms, but only when a varying defect probability was employed during the training period. A significant effect was obtained in block 4 for beta, which indicated that the increase in beta due to training differed significantly between conditions in which fixed and varying probability was employed. Examination of the means indicated that the fixed probability conditions produced an increase in beta whereas the varying probability conditions showed a slight decrease, the changes being +349 and -0.88 respectively. This effect was not, however, reflected by significant changes in the correct detection or false alarm probabilities, or the nonparametric bias measure. In block 5 there was again a significant effect of the use of fixed or varying probability in the training session, this

time on the sensitivity as measured by the Altham-Hammerton index. The means suggested that the decline in sensitivity as measured by this index was greater for the fixed than for the varying probability conditions.

Considering the results in general, for the fixed probability sessions, the overall pattern that emerged was that the criterion increased slightly to a value closer to the optimum. This was reflected in a slight increase in the overall payoffs obtained. An anomalous result was that all the sensitivity indices indicated a slight decline in sensitivity between the pre and post training sessions.

Varying probability sessions

With the modified experimental design which was adopted, it was not possible to make direct statistical tests on the ability of the subjects to modify their criteria to match the changing within session probabilities. By considering the overall changes between sessions 2 and 6, however, indirect evidence could be obtained regarding this question.

Considering the results in general terms, there was an overall slight increase in all the sensitivity indices and in beta, after the training sessions. There was some evidence that the subjects were generally adopting a more optimal strategy, as evidenced by the increase in payoffs obtained.

There was a significant interaction between the fixed or varying probability during training factor and the fixed or fading factor for d'

($p < 0.05$). Although none of the other sensitivity indices reached significance, high F ratios of 4.58 and 3.25 were obtained for the same interaction for the Altham-Hammerton and Pollack-Norman indices, which suggested that effect was a genuine one. Examination of the interaction in detail indicated that a combination of fixed probability with fading feedback or cuing produced the greatest increase in sensitivity.

Both the false alarm probability and the correct detection probability showed significant decreases for the fixed or varying probability during training factor ($p < 0.05$). In the case of the false alarm probability, this factor interacted significantly with the training conditions, ($p < 0.05$). The simple effects analysis for this interaction is given below:

Source	SS	df	S	F	Significance
C at A1	0.083	4	0.02075	11.99	$p < 0.05$
C at A2	0.029	4	0.00715	4.085	N.S.
error	0.00693	4	0.00173		

Table 7.1 Simple effects analysis for AxC interaction for false alarm probability.

The analysis indicated that the false alarm probability decreased significantly after certain types of training, but only when a fixed probability was employed during the training session. The comparison of means for the training conditions indicated that there were significant differences between training condition 4 and conditions 1, 3 and 5. Training condition 4 is the maximum information session, where feedback, cuing and a summary were all provided. Similar

results were obtained when the data for the correct detection probability was considered in detail. The decrease in correct detections were more marked in the fixed probability condition. As with the false alarms, the greatest decrease in correct detection probability was found with a combination of training condition 5 and a fixed probability during training.

Discussion

Considering the fixed probability sessions, the overall picture is of an increase in the criterion, but a slight decline in sensitivity. This latter finding is very difficult to account for. With the considerable amount of experience that the subjects had received of examples of both defective and non-defective slides, it is very difficult to understand why at least some improvement in sensitivity did not occur.

The fact that the significant effects which did emerge from the analysis appeared in separate blocks, and were usually unsupported by changes in other related variables, leads one to suspect their validity. By performing separate analyses of variance on each block of data, a total number of 360 F tests were performed. It seems likely that at least some of these would achieve significance by chance, and this may account for the absence of any coherent pattern.

The varying probability session results produced more definite conclusions, which were more in accord with previous work. An overall increase in sensitivity occurred between the pre and post training sessions. Certain combinations of conditions seemed to promote a

greater increase in criterion than others. In particular the fixed probability version of the feedback + cuing + summary training condition promoted a greater degree of change towards the optimum than the other training methods. This was corroborated by the significant changes in false alarm and correct detection probabilities.

The finding that the training condition which gave the maximum amount of information about the signal distribution, also produced the greatest criterion change towards the optimum is in accord with the previous work on perceptual training reviewed earlier. Why this change should be greater when a fixed probability was utilized during the training session is not clear. Support for the suggestion that this combination of training variables enables the subject to optimize his criterion most effectively comes from a consideration of the analysis of variance for the total score obtained. (Appendix A). Although it does not achieve significance, an F of 3.53 for the interaction of training conditions x probability type used during training was obtained. Consideration of the summary table for this interaction shows that the magnitude of the increase in payoff for the fixed probability, feedback + cuing + score training condition was considerably greater than for any other combination. Since the payoff obtained is strongly influenced by the response bias of the subject, this provides reasonable indirect evidence that this form of training enables the criterion to be adjusted effectively to the changing probabilities.

The result that d' appears to be increased by fading the cuing or feedback employed during the training sessions confirms the finding of Abrams and Cook (op.cit.) and Wallis' suggestion that supplementary

cues and information must be withdrawn during training in order to promote effective learning of the signal characteristics.

If we consider the results in general, the most difficult aspect to account for is the relative absence of significant effects for the fixed probability sessions. This can be partly ascribed to the insensitivity of the design employed. Only 4 degrees of freedom are available for the error estimate, and an F value of at least 6.91 is required to achieve significance at the 5% level. On the other hand significant effects were obtained for the variable probability conditions. There seems to be no obvious reason for this effect.

In summary, the results have the following implications for training inspection and other perceptual skills. Where a variable signal probability situation occurs, the form of training which appears to produce the most effective optimization of the criterion is one in which KR, cuing, and knowledge of the payoffs resulting from the responses is given, in conjunction with a fixed signal probability. Sensitivity appears to be enhanced when the feedback or cuing employed in the training technique is gradually reduced.

7.3 The use of tests of cognitive skills in the selection of inspectors

In Chapter 3 it was proposed that various tests of cognitive skills might be suitable as selection tests for industrial inspectors. Three types of cognitive skills were proposed as having relevance to the inspection situation. The first of these was known as field independence, the ability of a subject to selectively disembed wanted items from a confusing background. In view of the fact that

most inspection situations involve the recognition of the characteristics of a defect in a perceptually ambiguous background, it was proposed that some of the embedded figure tests which measure field independence might prove suitable for selecting inspectors.

The next type of perceptual skill which seemed to be relevant to the inspector situation was the ability of the individual to resist distraction, where a distracting context may be regarded as one which tends to obscure a wanted signal without changing its nature. This cognitive skill was regarded as being related to, but distinct from, the dimension of field independence measured by embedded figure tests. The final cognitive skill of interest was known as shifting, and could be described as the ability to change one's attentional focus at will. In an inspection context, this ability would be important in allowing the inspector to readily shift attention as each item was presented.

In the study to be described in this section, an attempt was made to investigate the extent to which performance on the detection task described in the earlier sections of this chapter correlated with various pencil and paper tests designed to measure individual differences in terms of these three dimensions of perceptual skill. If sufficiently high correlations were obtained between the tests and the various performance measures, this would suggest that they could, after suitable validation, be employed for selection purposes.

Procedure

The first groups of subjects consisted of the 20 participants in experiment 7, the last training experiment described. For various

reasons, three of these subjects were unable to take the tests, and hence the final number of subjects employed was 17. It was felt that the subjects used in experiment 7 were, however, unrepresentative of a real inspector population. For this reason, a number of inspectors from the Quality Control Department of Ilford Ltd., which was analysed in detail in Chapter 5, were invited to take part in the study. Fifteen inspectors agreed to participate, which constituted nearly the whole of the examiner population. These subjects travelled to Birmingham on successive days, and took part in sessions 1 and 2, the fixed and variable probability conditions of experiment 7.

Subsequently, each group of subjects were administered the tests described in the next section.

Description of tests

The tests fell into three broad categories, corresponding to the three dimensions of cognitive skills discussed earlier.

The first two tests were designed to measure selectivity of attention, or field independence. The Group Embedded Figure Test (Witkin et al., (1971)) is an adaptation of Witkin's original Embedded Figures Test for group administration. The test consists of 18 complex figures, within which are embedded simpler figures. The complex figures are shaded, to emphasize large, organized Gestalts which serve to increase the difficulty of the disembedding task. The subject is prevented from simultaneously viewing the simple form and the complex figure containing it, by printing the simple forms on the back of the test booklet, and the complex figures on the booklet pages. The test

is divided into three sections, the first containing 7 very easy items for practice and each of the remaining two sections containing 9 more difficult items. Three minutes were allowed for the practice session and 5 for each of the remaining sections. The other test employed was the Closure Flexibility test, Thurstone and Jeffrey (1965). This is an adaptation of the Gottschaldt Figures from which most of the embedded figure test work originated. The test measures the ability of a subject to form a closure in the face of some distraction and was developed as a result of factor analytic studies by Thurstone (1944) and Pemberton (1951). Each item in the test consists of a figure, presented to the left of the page, followed by a row of four more complex drawings to the right. Some of the more complex drawings contain the given figure in its original size and orientation, and subjects are required to check the appropriate drawings. Ten minutes were allowed for this test. For both tests the subject's score was the number of correct answers given in the allotted time.

To assess distraction, five tests from Karp's Kit of Selected Distraction Tests (1962) were used. The Distracting Contexts Test I (DCT1) involves the subject locating a simple geometric figure within a matrix of extraneous lines and figures. In the distracting contexts test 2A (DCT 2A) the subject is required to locate a series of simple geometric figures within a large matrix of such figures. In the second version of this test (DCT 2B) as an added source of distraction, coloured overlays were superimposed over the simple figures. Two minutes were allowed for each of these tests. The arithmetic operations test consisted of 24 simple arithmetic problems spaced evenly on horizontal rows, interspersed with a series of irrelevant jokes, instructions and pictures. The subjects were given one

minute to complete as many of the problems as possible. The final distraction test was a cancellation task in which subjects were required to cross out the letters a, t and c each time they appeared on a page of randomly arranged single spaced letters. Three minutes were allowed for this test.

Two tests were employed to assess the shifting variable. The first of these had been employed in a study by Sack and Rice (op.cit.) and consisted of the first and last pages of an anagram test developed by Gardner, Lohrenz and Schoen (1968). After a practice list of anagrams, the subjects were given two minutes to solve as many of the 20 anagrams as possible. Their score was the number correct in this time. Mendelsohn et al., (1966) amongst others have postulated that anagrams require voluntary shifting in attention. The second test alleged to measure the shifting variable is the Reversed Triangles Test, Sanguiliano (1951). In this test the subject is given one minute to draw as many triangles as he can, each separate and with the apex upwards. This is then repeated with the apex downwards, and in the final one minute session he is required to alternate the triangles with points upwards and downwards. The score is the total number of triangles drawn during the last session.

In the test session the tests were presented in the order described, i.e.:

1. Group embedded figures test
2. Closure flexibility
3. DCT 1
4. DCT 2A
5. DCT 2B

6. Arithmetic operations
7. Cancellation
8. Triangles
9. Anagrams

Results

The raw scores for the various tests are in Appendix B. A multiple regression analysis was first performed using all 9 test scores as independent predictor variables, for the first fixed probability and the first variable probability session for all 36 subjects. The dependent variables considered were those that had been utilized in previous analyses i.e. the parametric and nonparametric measures of sensitivity and bias, the overall score obtained by the subjects and the false alarm and correct detection probabilities. An analysis of variance was performed on each multiple regression equation to test its significance.

For the first analysis, high multiple correlation coefficients were obtained, but because of the large number of predictor variables utilized, relative to the number of subjects, none of the regression analyses reached significance, although significant correlations were obtained for some of the individual tests. The analyses were therefore repeated, but prior to each multiple regression, those independent variables which did not account for a significant degree of the total variance were deleted. This procedure, which is a standard one for multiple regression analyses, produced a considerable improvement in the fit of the regression equations, virtually all of which were now significant. The results for the fixed probability session are shown below in Table 7.2. All the multiple

correlation coefficients are significant at $p < 0.05$.

A similar procedure was carried out for the variable probability session, and the best combination of predictor variables found as before. Again, all the multiple correlation coefficients are significant at $p < 0.05$. These results are given in Table 7.3. In both tables, the independent variables used in the regression are set out, and any significant correlations of those variables with the dependent variable are indicated. The significance test employed was a two-tailed t test in this case. The independent variables referred to by number in the tables are the following tests:

2. Group Embedded Figures Test
3. Closure Flexibility
4. Distracting contexts test 1
5. Distracting contexts test 2A
6. Distracting contexts test 2B
7. Arithmetic operations
8. Cancellation
9. Triangles
10. Anagrams

It will be recalled that tests 2 and 3 are related to the field independence variable, tests 4 - 8 to the ability to resist distraction variable, and tests 9 and 10 to the shifting of attention variable.

Discussion

In general the results are encouraging in that they provide reasonable support for the assumption that at least some of the cognitive skills

dependent variable	independent variables included in the regression	multiple correlation
Altham-Hammerton index	3, 4, 5, 7*, 8*	0.636
Pollack-Norman index	2, 5, 7, 8	0.516
P(A)	3, 4, 8, -10	0.554
d'	2, 5, 8	0.481
log beta	2**, -3, -4, 5	0.550
Hodos-Grier bias index	2*, -3, -4	0.503
Score	2*, -7*, 8*, 10	0.500
False alarm prob.	-2*, -3, -8	0.486
Correct detection prob.	2, 3*, 4, 8, -10	0.596

* = $p < 0.05$

** = $p < 0.01$

\- = negative correlation

Table 7.2 Summary of multiple regression results for fixed probability session.

dependent variable	independent variables included in the regression	multiple correlation
Altham-Hammerton index	2*, 5, -7, 8	0.555
Pollack-Norman index	2**, 5, -7*	0.571
P(A)	2*, 5	0.557
d'	2*, 5, -7*	0.525
log beta	2*, -3, -4*, -7, 9	0.587
Hodos-Grier bias index	2*, -3, -4, -7*, -9, 10	0.630
Score	2, -7**, 9, 10*	0.625
False alarm prob.	-2*, 7**, -10	0.639
Correct detection prob.	3*, 5, 9, 10	0.550

* = $p < 0.05$

** = $p < 0.01$

\- = negative correlation

Table 7.3 Summary of multiple regression results for varying probability session.

measured by the tests utilized in this study are related to performance on the simulated inspection task.

If we consider the sensitivity indices for both fixed and varying probability tasks certain general patterns emerge.

In both situations, sensitivity seems to be consistently related to the field independence dimension, as measured by tests 2 and 3 in the first session and test 2 in the second. Some of the tests of distractability also show clear relationships with the various sensitivity indices, tests 5 and 8 being positively related in both fixed and variable probability sessions. However, test 7, the Arithmetic Operations test, shows a positive correlation for the fixed probability session and a negative correlation for the variable probability situation. This will be discussed subsequently (n.b. a high score in the distractibility tests indicates low distractibility). The indices of distractibility apart from variable 7, show the expected relationship, i.e. the less distracted the inspector is by non-signal stimuli, the higher his apparent sensitivity is likely to be.

Examination of the bias measures indicates a negative correlation between the dependent variable and measures of distractibility, this time tests 3 and 4 in the fixed probability session and 3, 4 and 7 in the variable session. Since subjects who have a low criterion will be those who have not raised it to the level appropriate to the overall defect density in both sessions of 0.2, we can infer that the ability to modify the criterion in the appropriate direction seems to be negatively correlated with distractibility as measured by tests 3, 4 and 7. The results for test 7 for the sensitivity indices

can now be explained. In the fixed probability session, the ability to ignore irrelevent stimuli that is characterized by high scores on distractibility measures, leads to the positive correlation of variable 7 with sensitivity. It could be hypothesized that another aspect of low distractibility, at least as measured by test 7, might be an inability to readily change strategies in the light of a change in defect density. This would give rise to the negative correlation with apparent sensitivity in the variable session.

These conclusions are complicated by the intercorrelations that can be expected between the bias and sensitivity measures in the unequal variance case, and it is probably better to concentrate on the performance measures which do not exhibit these complications. Considering the overall score, this can be regarded as a function of both the sensitivity of a subject and the appropriateness of his response strategy. The significant correlations obtained with the Embedded figures test are clearly related to the strong relationship of the field independence measures to sensitivity. The negative correlation of test 7 with the score, confirms the earlier suggestion that a high score on this test implies a lack of flexibility of strategy. This could be due to an inability of the subject to observe the 'irrelevant' evidence of a change in defect probability. The presence of the shifting variables in connection with the score is also presumably related to the ability of the inspector to shift his attention from the primary focus of the task, to the secondary aspect of the defect density, and its possible changes.

The final dependent variables of interest are the correct detection and false alarm probabilities. In both fixed and varying probability

sessions the correct detection probability shows a significant correlation with field independence as measured by the Closure Flexibility Test. In view of the earlier correlation of this cognitive dimension with sensitivity, this result is not surprising. This also accounts for the negative correlation of variable 2 with the false alarm probability. Significantly, the false alarm probability shows a strong positive correlation with the arithmetic operations test 7, which lends weight to the earlier suggestion that a high score on this test is associated with an inflexible strategy.

To summarize these findings, it appears that sensitivity for defects on tasks similar to the one utilized in this study, is predicted fairly well by performance on field independence tests. The ability to modify the criterion optimally seems to be associated with a low score on the arithmetic operations test.

It should be emphasized at this point that the analytical techniques employed in this study have been somewhat crude. For example the selection of variables to include in the multiple regression may not have been optimal, since it was based primarily on the significance of these variables in the first analysis. A better procedure is to use a step-wise multiple regression, which includes variables in the regression in the order that they reduce the sum of squares of the variability. Work is continuing along these lines, but the results could not be included in this thesis. Finally, a much more sophisticated approach such as a factor analysis might have been appropriate. This was considered, but the limited sample size of 36 meant that this technique could not be meaningfully employed.

In spite of the crudity of the methods the degree of significant correlation found in this essentially exploratory study suggest that the approach is a fruitful one, particularly in comparison with the other attempts at predicting inspection performance, as discussed in Chapter 3. The next step involves testing the validity of these findings in a large scale industrial study.

7.4 Conclusions

The overall conclusions that emerge from experiments 6 and 7 are in line with the perceptual learning principles that have been established in laboratory tasks using simple signals. To this extent, the studies have performed a useful function in extending the results to a situation more typical of a real inspection task. Certain previously unreported findings have, however, emerged.

In terms of training an inspector to modify his criterion optimally in a changing defect density situation, the use of a fixed probability during the training session appears to produce the best results. A possible explanation for this is that the fixed defect density gives the trainee sufficient experience of a single probability that he is able to utilize this as an anchor point on the subjective probability continuum. A particular defect density would then be recognized as being greater or less than the anchor probability, and this would presumably facilitate the modification of the criterion in the appropriate direction. This model suggests that the appropriate training technique changes the recognition of the prevailing probability from an absolute to a comparative judgement task. All the available evidence suggests that the latter form of judgement is performed more

effectively than the former, e.g. Guilford (1954)

The finding that the most effective criterial adjustment occurred after training with the feedback + cuing + score condition is a logical consequence of regarding effective training as providing as much experience as possible of an anchor probability. The provision of the score information would in addition enable the trainee to develop appropriate strategies to maximize his payoffs. Such strategies are identical to those which optimize the criterion setting.

The general conclusions from the existing literature, that there is little to choose between cuing and KR as a means of enhancing sensitivity, where these two techniques produce the same information, is generally confirmed by the results. Sensitivity was increased significantly by training in which the cuing or KR was gradually reduced, as found by Abrams and Cook (op.cit.) and predicted by Wallis (op.cit.). The finding that a fixed probability during training is also necessary for an increase in sensitivity is at first sight difficult to account for, since the subject receives the same number of signal and noise samples under both fixed and varying probability conditions. One possibility is that the fixed probability sessions allow a more evenly spaced occurrence of the defect samples, and learning theory in general suggests that spaced practice is preferable to the massed practice that would be represented by the high defect probability blocks during the training sessions.

The results of the correlational study between cognitive skills as measured by certain pencil and paper tests and performance on the inspection task, suggest that this is a promising approach to the

question of selecting inspectors. Overall, the results suggest that sensitivity for defects can be predicted to some extent by tests which measure field independence, the ability to be able to perceive wanted configurations embedded in background noise which contains perceptual elements similar to the signal. This result is in accord with the predictions of SDT, that sensitivity is a function of the distance apart of the signal + noise and the noise distributions, whether this separation is due to external attributes of the signal, or to internal characteristics of the perceiver.

The finding of a consistent negative correlation between a test measuring the dimension of distractibility and the ability to maximize the payoffs in the inspection task was unexpected. If distractibility is regarded as the ability or otherwise to maintain a fixed focus of attention, however, the results become more comprehensible. It seems reasonable that such a quality might correlate negatively with the flexibility of strategy required to change the criterion to match the prevailing defect probability. The rigidity of attention that would be an asset in a situation where many external distractions are present, could be a disadvantage where more subtle aspects of the task needed to be noted, such as changes in defect density.

Although the present results are of considerable interest, they need to be replicated with a larger sample size, and with an industrial task, before the tests could be utilized as part of a selection procedure for inspectors.

CHAPTER 8 GENERAL CONCLUSIONS

8.0 INTRODUCTION

In this chapter the findings from the previous chapters will be presented and overall conclusions drawn. Finally, directions for further research will be discussed.

8.1 The literature review and its application to the analysis of inspection tasks

Chapter 2 provided a general overview of the SDT literature and established basic guidelines for applying SDT to inspection studies. Consideration of the studies that had utilized SDT suggested that the usefulness of SDT in the context of inspection tasks needed to be further investigated. The possibility of using SDT to examine the ability of inspectors to change their strategies emerged as a further experimental goal.

Problems were encountered when attempting to classify the inspection literature as a whole. Originally it had been hoped that the informal inspection model proposed in Chapter 1 would provide the basis for a classification scheme. It was found, however, that although this model provided a useful conceptual summary of the psychological and other areas relevant to inspection, it did not allow the major external factors influencing performance to be readily included.

Although the number of studies on inspection as such was limited, a very large number of theoretical areas were relevant to analysing such tasks. The review was, therefore, divided into two parts. The first part considered the major theoretical areas apart from SDT,

i.e. vigilance and visual search, and the second part utilized a simple classification scheme. This consisted of four main headings, i.e. task characteristics, environmental factors, organizational factors and individual factors. These main categories were divided into sub-categories as indicated below:

1. Task characteristics

a. Pacing and movement of the item being inspected

b. Magnification, lighting and other aids to enhance defect discriminability

c. Complexity

d. Display organization

e. Signal rate

f. Number of inspectors

g. Repeated inspection

2. Environmental factors

a. Heating

b. Lighting

c. Noise

d. Workplace layout

3. Organizational factors

a. Management and social aspects

b. Motivational variables

4. Individual factors

a. Selection

b. Visual abilities

c. Age

d. Sex

Although this scheme was somewhat ad hoc in nature, it encompassed most of the available literature and provided a means of structuring the subsequent analysis of the inspection tasks considered in the case studies.

8.2 The case studies

8.2.1 The data analysis group

The first case study was a re-analysis of an inspection system described by the author in an earlier study. The much broader range of literature available, and the insights obtained from the use of SDT enabled a much more sophisticated analysis of this task to be performed. The first goal of the study was to investigate if the SDT model applied to the data, and to examine the interrelationships between the parametric and non-parametric measures of sensitivity and bias. A number of tests of the SDT model were applied, most of them employing the detection and false alarm data. In general, the model, in its unequal variance form, fitted the results reasonably well. Corroborative evidence was obtained, from a consideration of the latency data, that the inspectors were employing a likelihood ratio criterion.

The parametric and non-parametric sensitivity and bias indices showed a high correlation, although it was demonstrated that they did not produce equivalent results if utilized in statistical analysis. It

was shownthat other commonly used inspection measures were correlated with both sensitivity and bias.

There were very few significant effects of the experimental variables on the SDT indices, apart from log beta, which showed a significant decline with time on task. This effect, which was in the opposite direction to that generally found in vigilance tasks, was explained in terms of the subject's perception of the experiment as a 'risky' situation, inducing an initially abnormally high criterion. With habituation to the experimental situation, the criterion declined to its usual level. No significant main effects were found for the noise variable, although some interactions with subjects occurred.

Analysis of the latency data was not attempted in detail, because scanning and decision times could not be separated. The mean latencies for the various categories of response were however, in accord with a simple extension of the SDT model to the latency situation. The unexpected lack of any significant differences in discriminability between the two types of signal employed, was explained as being due to the self-paced nature of the task. The inspectors were able to take longer to sample more attributes of the inherently less discriminable signals.

In general it was felt that the use of SDT in the analysis of this task provided insights which could not have been obtained by conventional measures.

8.2.2 The Ilford quality control system

The other case study, although it was also concerned with targets on film, presented completely different problems from the first. The difficulty of obtaining accurate estimates of the correct detection and false alarm probabilities limited the application of SDT in this study. A further difficulty was the highly subjective nature of the standards of product acceptability. A decision making model was proposed for the task, which suggested that SDT principles operated at two levels. The first of these determined the quality of the evidence that was obtained from the film in terms of the number of defects that were observed. Although the number of defects occurring was not usually precisely noted, the particular sensitivity and bias being employed by an inspector would determine his subjective estimate of the incidence of defects, which would in turn be used as evidence at the second stage of decision making. It was hypothesized that the inspector utilized a criterion at this stage which determined whether or not the film was acceptable as a whole. This criterion could be set at a different position than that employed at the first stage.

Two experiments were performed in an attempt to apply SDT to the task. In both cases considerable difficulty was experienced for the reasons discussed earlier, and the SDT parameters could only be obtained for the second experiment. Even in this case, the estimation of these quantities involved a number of untested assumptions. The most obvious characteristic of the results were the relatively low detection scores for defects which occurred, particularly in view of the considerable experience of the examiner subjects. This was felt

to be because the examiners normally made global estimates of the acceptability or otherwise of a film sample, rather than noting the incidence of individual defects. No statistically significant effects were obtained from the experiments, apart from subject differences, although a decline in log beta over time in experiment 3 similar to that observed in experiment 1, was noted.

Although the SDT parameters could not be easily determined in this situation, it was felt that the general principles of the SDT model provided useful insights into ways of improving the system. In particular, it was proposed that a form of training which would enable the examiners to modify their criteria in the optimal direction, in the event of a change in defect incidence, was desirable. The indications from management, that the ability to detect a change in quality was important, provided a spur for experimental investigations in this area in Chapter 5. Various suggestions of improvements were made to management as a result of the investigation. In order to facilitate the stabilization and standardization of the definitions of acceptable quality, it was proposed that a library of reference samples of defects be provided, together with examples of entire films that illustrated the required standards of quality. Regular 'calibration sessions' were also suggested to ensure homogeneous standards between different inspectors. It was felt that the relationship between the quality standards utilized by the inspectors and the criteria of acceptability of the customers should also be investigated.

8.3 The laboratory studies

8.3.1 The effects of between and within session defect probability changes

The first group of laboratory studies, comprising experiments 4 and 5, were concerned with the reaction of subjects to changes in the incidence of defects, both within and between sessions.

In the first experiment, performance was compared under four conditions. In the first two conditions, the subjects were required to detect signals which occurred at the same probability as they had been accustomed to during extensive practice sessions. In one case feedback in the form of a summary of performance was provided every hundred trials, whilst feedback was absent under the second condition. The other experimental conditions were similar to the first two, except that a much lower probability of signal was utilized. The results indicated that subjects were able to increase their criterion in the correct direction for the changed probability, but that this increase was only significant where feedback was provided. These results were considered from the standpoint of the amount of information available to the subject concerning the defect density. It was suggested that there were two sources of information available, which allowed the subject to revise his subjective probability estimates of the actual defect probability. These were information from the task and information from external sources. Where the defects were of low discriminability, as in this task, the subjective probability estimate, and hence the ability of the subject to modify the criterion, was primarily dependent on the external evidence available. There was some evidence that under the low probability conditions the subjects were

attempting to maximize their score by adopting a more nearly pure strategy in game theory terms. The latency data were in agreement with the unequal variance SDT model which had been found to fit the data as a whole. The extension of SDT to the latency scores which had been utilized in experiment 1 seemed to provide a reasonable description of the results.

Experiment 5 considered the subjects' reaction to within session changes in defect probability, with and without the provision of prior warning and feedback. It was found that prior warning that a change in defect probability would occur, was more effective in producing a criterion change in the required direction, than feedback from the task. Even in the absence of feedback, the subjects were able to modify their criteria appropriately. Without feedback, the degree of adjustment was more extreme than when feedback was provided.

Both experiments provided general support for the hypothesis that the degree of criterial change produced by a subject in a changing defect density situation was primarily a function of the amount of evidence available, from both within and outside the task, concerning the actual defect probability. Comparison of this study with others suggested that the greater the difficulty of the discrimination required, the more important external evidence became. In industrial situations, the provision of 'feedforward' information, giving the inspector prior warning of defect changes, was emphasized.

8.3.2 Training techniques for inspection

In experiments 6 and 7 two main questions were investigated. The first of these was concerned with testing the applicability of existing research on perceptual training, to tasks employing stimuli more representative of those found in real inspection tasks, than had hitherto been employed. The second objective was to investigate the possibility of devising training techniques which would both enhance the sensitivity of the inspector and also enable him to modify his criterion in a changing defect probability situation.

Both experiments 6 and 7 utilized stimulus material obtained from the Ilford inspection task described in Chapter 5, and the experiments were designed to simulate the critical features of this task.

Experiment 6 was a straightforward comparison between cuing and KR as training techniques, and a control condition. There were no significant changes in sensitivity as measured by the SDT indices, but the cuing training produced a significantly greater increase in the criterion towards the optimum, compared with the control condition. It was not, however, significantly different from the KR condition. The reason for the absence of a change in sensitivity was thought to be due to the complex nature of the signals employed, compared with those utilized in most perceptual training experiments. It seemed likely that many more training trials would be necessary to produce a significant change in sensitivity. The adjustment of the criterion in the appropriate direction after training sessions can be regarded as being due to the additional information on the signal distribution that these sessions provided.

The final experiment was designed to provide more evidence concerning the effectiveness of various types of training, and utilized a considerably greater number of training trials than experiment 6, in the light of the non-significant results obtained in that experiment. Two of the training conditions investigated were identical to those in the previous experiment, and the remainder utilized combinations of cuing, KR and/or the provision of the score information which indicated the subjects overall payoff from the task, given the payoff matrix assigned to the various response possibilities. Two further factors were included in the experimental design. These were the use of a fixed or VARYING defect probability during the training sessions, and either gradually reducing or keeping constant the amount of cuing or KR being provided. The variable probability factor was included to test the hypothesis that the inspector would be better able to adjust his criterion to a variable defect probability situation if he had previously experienced a range of known probabilities. The inclusion of the fading or otherwise condition was intended to establish whether the previous findings of Abrams and Cook (op.cit.), and Wallis (op.cit.), that cues or KR needed to be removed during training to produce a sensitivity increase, applied in this situation. The experiment employed two pre and two post training sessions, one with a fixed and the other with a varying defect probability.

Very few significant effects of training were found with the fixed probability session. No obvious reason was apparent for this finding, apart from the fact that the post training session may have been affected by the absence of the training aids that had been present during the immediately preceding two sessions.

With the variable probability sessions, a number of interesting effects were obtained. The combination of a fixed probability during training, with the training condition that provided both feedback, cuing, and the score information, appeared to produce the most optimal performance in a situation of changing defect density. It was suggested that the fixed probability condition enabled the inspector to build up an accurate perception of a particular defect incidence. This then provided an anchor point on the scale of subjective probabilities, that allowed him to accurately assign other observed probabilities as being greater or less than this probability. Consequently the inspector was able to modify his criterion more accurately in accord with the changing probabilities of defects. The particular training condition found to be most effective, provided the maximum information on the defect density, and also the score information that would allow the inspector to develop the response strategies that would maximize his payoffs.

Considering the sensitivity changes, it was found that, as predicted by the workers cited earlier, significant sensitivity changes were obtained in training sessions where the cuing or feedback was gradually reduced. It was also found that a fixed probability during the training session promoted the greatest increase in sensitivity. A possible explanation for this effect was the more even spacing of the defect samples that would occur during the fixed probability trials.

8.4 Cognitive skills as factors in the selection of inspectors

The scores from experiment 7 were used in conjunction with a number of tests of cognitive skills to determine if such tests could be used in selection procedures for inspectors. In order to make the

subject population more representative of those employed in inspection work, 16 inspectors from the Ilford quality control system performed the first two sessions of experiment 7 and their scores were included in the correlation analysis.

It was found that sensitivity, as measured by a number of parametric and non-parametric indices, seemed to be significantly related to performance on tests of the field independence dimension of cognitive skills. There was some indication that the ability of the subject to modify his strategy in a changing defect density situation was negatively correlated with performance on one of the distractibility tests, the arithmetic operations test. It was emphasized that the correlational study as it stood was exploratory in nature, and that further analysis could be performed on the data. The findings did indicate however, that with proper validation in an industrial context, this particular approach shows promise as a potential selection aid.

8.5 General conclusions

Both the industrial and the laboratory based studies have shown that SDT, as a general conceptual standpoint, offers unique advantages in suggesting ways in which industrial inspection systems can be optimized. It is suggested that a basic initial step in the analysis of any ongoing system is to perform an experimental study of its effectiveness which can be subsequently analysed in SDT terms, as shown in the industrial case studies. A particular strength of the SDT approach is that even where, as in the Ilford situation, precise measurements of the SDT parameters cannot be made, the insights that

can be gained by considering the system from this standpoint are still valuable. SDT has often been applied in a rather casual manner in many previous studies. An attempt has been made to show that the simple equal variance form of the model cannot be applied to any situation without initial tests to ensure that its underlying assumptions are fulfilled. The procedures that need to be observed in applying SDT to industrial situations have been spelt out in some detail. The use of the non-parametric indices of bias and sensitivity using real inspection data, has allowed some assessment of their usefulness. In general it is felt that their main application is in adding weight to conventional measures, provided these have been obtained using the appropriate model.

The survey of the literature produced a simple classification scheme which provides at least an initial approach to structuring the analysis of an inspection system. Used in conjunction with the SDT paradigm, it should provide a useful source of data in the design of new quality control systems. It is hoped, in the long run, to produce a more comprehensive handbook and classification scheme which should prove to be a useful aid to ergonomists working in this area.

The first laboratory experiments emphasized the importance of prior warning that changes were going to occur in the incidence of defects, in allowing the inspectors to adjust their criteria to maintain optimal performance. It could be argued that the function of a quality control system is to monitor sudden changes of this type and that prior warning does not usually occur. Although this is true in some situations, there is very often some indication at the production stage, that a higher incidence of defects can be expected. If this

information is effectively communicated to the inspection system, a far higher detection efficiency can be expected. The existence of 'feedforward' links of this type is of course a function of the organizational structure of which the quality control system is a part. The other result from the first two laboratory experiments was that feedback during the task was an effective means of maintaining an optimal criterion. Although most inspection systems do not provide such knowledge of results, in a direct way, there seems to be good grounds for recommending that re-inspection be carried out far more frequently, by senior inspectors, in order to provide such feedback. An additional advantage of this procedure would be that a greater degree of consensus would be produced between all the inspectors, as to what the appropriate criterion should be for a particular product. This is particularly important in a system such as at Ilford, where the levels of acceptability are essentially subjective, and are determined by a complex combination of factors, such as product, customer, and market conditions.

In the analysis of the experiments under discussion, the possibility that insights from probability learning theory might explain performance in changing defect density situations was put forward. Although the approach in this study has largely been from the stance of SDT, this should not lead us to ignore other relevant viewpoints. There seems to be many common areas between the two orientations that could usefully be explored in an inspection context.

The final experiments confirmed the general principle from previous studies of perceptual learning, that up to a point, the greater the information presented to the subject concerning the characteristics

of the defect and its distribution in time, the more effectively he was able to increase his sensitivity and adapt his criterion to the prevailing defect density situation. The experiments were limited in size and scope and therefore care needs to be taken in generalizing from them to other situations. Nevertheless the suggestion that subjects detect changes in defect probabilities by comparing them with a previously established single subjective probability, seems to be a reasonable one which fits the experimental data. Further work is needed to establish its validity more generally. The same considerations apply to the finding that a gradual reduction in the amount of cuing and feedback during training enhances sensitivity. This result apparently conflicts with the earlier finding concerning the importance of information in enhancing sensitivity. It seems clear that the provision of feedback or cuing is only effective up to a point in increasing sensitivity. Beyond that point, further information is counter productive, because it hinders the development of the internal referents necessary for true learning. The exact point at which the supplementary information should be reduced is probably a function of the individual task concerned, and the provision of general guidelines on this point would require more detailed research.

8.6 Directions for further research

As indicated in the introduction to the thesis, the research philosophy adopted has been to approach the area of quality controm from a number of directions, which were unified by the general orientation of SDT. The outcome of this research has been that a number of results of direct practical applicability have been obtained. The practical utility of the approach has to be weighed

against the fact that some degree of speculation has been necessary, and the size of the experiments has been smaller than might be found in the traditional academic investigation of a single clearly defined research hypothesis. It is argued that the orientation adopted has generated testable hypotheses that have relevance to real world problems and can therefore be verified in an industrial environment. Most of the recommendations for further research are concerned with the validation of these findings.

8.6.1 Validation of the two stage inspection model

In Chapter 5 a model was proposed which postulated two distinct decision making phases which might be expected to occur in inspection situations involving the aggregation of information over time. It is proposed that this model be validated both at Ilford and in other similar inspection systems.

8.6.2 Factors affecting the modification of the criterion

The discussion of Chapter 6 considered the sources of information that the inspector might utilize in modifying his subjective probability estimates, and the degree to which his response strategy actually changed, given that he perceived the on-going defect density accurately. Two experimental investigations are required to clarify these points. The first would involve changing the amount of evidence available from the two sources of information and manipulating its reliability. Studies of this type have been carried out by Ingleby (1974) in the context of auditory detection, but no corresponding studies have been performed for inspection. The second study

would involve performing an experiment similar to that described by Sims (1972) and discussed in detail in Chapter 6. In that experiment subjects were required to give their subjective probability that an item would be defective before inspecting it. It would be of considerable interest to investigate the relationship between this subjective probability, the value of the criterion adopted, and the actual defect density.

8.6.3 Verification of the perceptual training findings

As mentioned earlier, the findings from experiments 6 and 7 although intuitively reasonable, require further validation, using both laboratory simulation and real-life inspection tasks. In this case, it would be necessary to utilize a task which would allow the unambiguous calculation of the SDT parameters, and which would allow extensive trials to be taken.

Ilford have already expressed their willingness to incorporate the results of work on the perceptual training problem in their standard inspector training procedures. This would allow a longitudinal study of their usefulness to be performed.

8.6.4 Further work on the cognitive skills approach to selection

As discussed earlier, the current analyses of the cognitive skills tests and the inspection data must be regarded as provisional. Further analysis is under way, and additional correlational studies using industrial performance data are planned.

8.7 Concluding remarks

Although it is clear that we are only just beginning to be able to specify all the human factors requirements of an optimal inspection system, it is hoped that the data and techniques reported in this research will be of direct applicability in the design and analysis of quality control systems.

In addition to this objective, an attempt has been made to achieve a more general goal. This is to show that given an appropriate research strategy, the theoretical models of experimental psychology, exemplified in this case by SDT, can make a significant contribution to solving practical industrial problems. The approach adopted in this study, of combining field work with simulation and laboratory studies, unified by a common theoretical orientation, seems to be of potentially wide application.

There seems to be a considerable need at the present time for a bridge building operation between the concerns of much research in the behavioural sciences and the practical problems of society and industry. It is hoped that the research methodology presented in this study, whilst far from being a blueprint, will at least suggest some ways in which the construction of such bridges might begin.

REFERENCES

ABRAMS, A.J., & COOK, R.L., (1971) Information feedback: Contributions to learning and Performance in Perceptual Identification Training. Technical Bulletin STB 72-5 Naval Personnel and Training Research Laboratory, San Diego.

ADAMS, S.K., (1975) Decision making in quality control: some perceptual and behavioural considerations. In: Human Reliability and Quality Control (London: Taylor & Francis).

ALPERN, M., & BARR, L., (1962) Durations of the after images of brief light flashes and the theory of the Broca and Sulzer phenomenon. Journal of Optical Society of America, 52, 219-221.

ALTHAM, M.E., (1973) A non-parametric measure of signal discriminability. The British Journal of Mathematical and Statistical Psychology, 26.

ANNETT, J., (1959) Some Aspects of the Acquisition of Simple Sensori-motor Skills (D.Phil. Thesis, Oxford University).

ANNETT, J., (1961) The Role of Knoweedge of Results in learning: A Survey (NAVTRADEVCEN Technical Report No. 342-3, U.S. Naval Training Device Centre, New York).

ANNETT, J., (1966) Training for perceptual skills. Ergonomics, 9, 459-468.

ANNETT, J., (1969) Feedback and Human Behaviour (Harmondsworth: Penguin Books).

ANNETT, J., (1971) Sonar recognition training: An investigation of whole versus part and analytic versus synthetic procedures. (Technical Report No. 67-L-0105-L: U.S. Naval Training Devices Center, Orlando).

ANNETT, J., & CLARKSON, J.K., (1964) The use of cuing in training tasks. (U.S. Naval Training Device Center: Technical Report 3143-1).

ANNETT, J., & PATERSON, L., (1966) The use of cuing in training tasks: Phase II. (U.S. Naval Training Device Center: Technical Report 4119-1).

ANNETT, J., & PATERSON, L., (1967) The use of Cuing in Training Tasks: Phase III (Technical Report, NAVTRADEVCEN 4717-1, Orlando, Florida).

ASTLEY, R.W., & FOX, J.G., (1975) The analysis of an inspection task in the rubber industry. Human Reliability in Quality Control (London: Taylor and Francis).

AYERS, A.W., (1942) A comparison of certain visual factors with the efficiency of textile inspectors. Journal of applied psychology, 26, 812-827.

BADALAMENTE, R.V., & AYOUB, M.M., (1969) A behavioural analysis of an assembly line inspection task. Human Factors, 11, 339-352.

BADDELEY, A.D., & COLQUHOUN, W.P., (1969) Signal probability and vigilance: a reappraisal of the 'signal rate' effect. British Journal of Psychology, 60, 169-178.

BAKAN, P., & MANLEY, R., (1963) Effect of visual deprivation on auditory vigilance. British Journal of Psychology, 54, 115-119.

BAKER, C.H., (1963) Further towards a theory of vigilance. Vigilance: A Symposium (New York: McGraw-Hill) 127-153.

BAKER, E.M., (1975) Signal Detection Theory Analysis of Quality Control Inspector Performance. Journal of Quality Technology, 7(2), 62-71.

BAKER, C.A., MORRIS, D.F., & STEEDMAN, W.C., (1960) Target recognition in complex displays. Human Factors, 2, 51.

BAKER, E.M., & SCHUCK, J.R., (1975) Theoretical Note: Use of Signal Detection Theory to Clarify Problems of Evaluating Performance in Industry. Organisational Behaviour and Human Performance, 13, 307-317.

BANKS, W.P., (1970) Signal Detection Theory and Human Memory. Psychological Bulletin, 74, 81-99.

BELBIN, R.M., (1957) New fields for quality control. British Management Review, 15, 79-89.

BELT, J.A., (1971) The Applicability of Vigilance Laboratory Research to a Simulated Inspection Task. U.S. Government Report No. A.D. 728-490.

BLACKWELL, H.R., (1959) Specification of interior illumination levels. Illuminating Engineering, 43, 906-931.

BLACKWELL, H.R., (1952) Journal of Experimental Psychology, 44, 306.

BLOOMFIELD, J.R., (1970) Visual Search. (Ph.D. Thesis, University of Nottingham).

BLOOMFIELD, J.R., (1975) Theoretical Approaches to Visual Search. Human Reliability and Quality Control, (London: Taylor and Francis)

BLOOMFIELD, J.R., (1975) Classifying studies of visual inspection. Human Reliability and Quality Control (London: Taylor and Francis).

BROADBENT, D.E., & GREGORY, M., (1963) Vigilance considered as a statistical decision. British Journal of Psychology, 54, 309-323.

BROADBENT, D.E., (1971) Decision and Stress (London: Academic Press).

BROCK, J.F., WELLS, R.G., & ABRAMS, M.L., (1974) Development and Validation of an Experimental Radiograph Reading Training Program. Navy Personnel Research and Development Center, San Diego, NPRDC-TR74-33.

BROWNE, R.W., (1965) On-the-Job Training of the Aerospace Nondestructive Test Inspector, Materials Evaluation, October 1965, 489-492.

BUCK, J.R., (1975) Dynamic Visual Inspection, Human Reliability and Quality Control (London: Taylor and Francis).

BUSH, R.R., & MOSTELLER, F., (1955) Stochastic Models for Learning. (New York: Wiley).

CAMPBELL, R.A., (1964) Feedback and noise-signal detection at three performance levels. Journal of Acoustic Society of America, 36, 434-438.

CHANEY, F.B. & HARRIS, D.H., (1966) Human Factors techniques for quality improvement. 20th Annual Conference A.S.Q.C.: Technical Conference Transactions, 400-413, New York.

CHANEY, F.B., & TEEL, K.S., (1967) Improving inspector performance through training and visual aids. Journal of Applied Psychology, 51(4), 311-315.

CHAPMAN, D., & SINCLAIR, M.A., (1975) Applications of Ergonomics in Inspection Tasks in the Food Industry. Human Reliability and Quality Control (London: Taylor and Francis).

COCHRAN, D., PURSWELL, J.L., HOAG, L., (1973) Development of a Prediction Model for Dynamic Visual Inspection Tasks. Proceedings of the Seventeenth Annual Meeting of the Human Factors Society, Santa Monica, California.

COCKRELL, J.T., & SADACCA, R., (1971) Training Individual Image Interpreters Using Team Consensus Feedback. U.S. Army Behaviour and Systems Research Laboratory, Technical Research Report 1171.

COLQUHOUN, W.P., (1960) Temperament, Inspection efficiency, and time of day. Ergonomics, 3, 377-378.

COLQUHOUN, W.P., (1959) The effect of a short rest-pause on inspection efficiency, Ergonomics, 2, 367-372.

COLQUHOUN, W.P., (1961) The effect of unwanted signals on performance in a vigilance task. Ergonomics, 4, 41-5.

COLQUHOUN, W.P., (1967) Sonar target detection as a decision process. Journal of Applied Psychology, 51, 187-190.

COLQUHOUN, W.P., & BADDELEY, A.D., (1964) Role of pretest expectancy in vigilance decrement. Journal of Experimental Psychology, 68. 156-160.

COLQUHOUN, W.P., & BADDELEY, A.D., (1967) Influence of signal probability during pretraining on vigilance decrement. Journal of Experimental Psychology, 73, 153-155.

COOMBS, C.H., DAWES, R.M., & TVERSKY, A., (1970) Mathematical psychology, (New Jersey: Prentice-Hall).

CORNSWEET, T.N., (1970) Visual Perception (New York: Academic Press)

CRAWFORD, W.A., (1960) Perception of Moving Objects, IV, The Accuracy of fixation required in the Perception of Detail in Moving Objects. Air Ministry, Flying Personnel Research Committee Memo ISOD - London.

DAVIES, D.R., & TUNE, G.S., (1970), Human Vigilance Performance (London: Staples Press).

DORFMAN, D.D., & ALF, E. JR., (1968) Maximum likelihood estimation of parameters of signal detection theory - A direct solution. Psychometrika, 33,(1).

DORFMAN, D.D., & ALF, E. JR., (1969) Maximum likelihood estimation of signal detection theory and determination of confidence intervals - rating method data. Journal of Mathematical Psychology, 6, 487-96.

DRURY, C.G., (1975) Inspection of Sheet materials: Model and Data. Human Factors, 17 (3), 257-265).

DRURY, C.G., (1975) Human decision making in quality control. Human reliability and quality control (London: Taylor and Francis).

DRURY, C.G., (1973) The effect of speed of working on industrial inspection accuracy. Applied Ergonomics, 4 (1), 2-7.

DRURY, C.G., (1973) The inspection of sheet materials - model and data. 17th Annual Meeting of the Human Factors Society, Washington, October 1973, 457-464.

DRURY, C.G., & ADDISON, J.L., (1973) An industrial study of the effects of feedback and fault density on inspection performance. Ergonomics, 16, 159-169

DRURY, C.G., & SHEEHAN, J.J. (1969) Ergonomic and economic factors in an industrial inspection task. International Journal of Production Research 7 (4), 333-341.

DUSOIR, A.E., (1975) Treatments of bias in detection and recognition models: A review. Perception & Psychophysics, 17 (7), 167-178.

EDWARDS, W., (1961) Probability learning in 1000 trials. Journal of Experimental Psychology, 62, 385-394.

EDWARDS, W., (1962) Dynamic decision theory and probabilistic information processing. Human Factors, 4, 59-73.

EDWARDS, W. & PHILLIPS, L.D., (1964) Man as a transducer for probabilities in Bayesian Command and control systems. In G.L. BRYAN & M.W. SHELLEY (Editors), Human Judgements and Optimality, (New York: Wiley).

EHLERS, H.W., (1972) Effects of low frequency continuous noise on an inspection task. U.S. Army Logistics Management Center, Texarkana, Texas, Intern Training Center Report No. USAML-IT6-2-72-02.

EILON, S., (1961) Recirculation of products through an inspection station. International Journal of Production Research, August 1961, 39-44

EGAN. J.P., & CLARKE, F.R., (1966) Psychophysics and signal detection . In: Experimental Methods and Instrumentation in Psychology, (New York: McGraw - Hill).

EGAN, J.P., SCHULMAN, I., & GREENBERG, G.Z., (1959) Operating characteristics determined by binary decisions and by ratings. Journal of the Accoustical Society of America, 31, 768-73

ELLIOTT, E., (1960) Perception & alertness. Ergonomics, 3, 357-364

EMBREY, D.E., (1970) The Bubble Chamber Data Analysis Group. A Study of the Organisation and Human Factor Aspects of a Large Research Group. Unpublished B.Sc. Thesis, Department of Physics, University of Birmingham.

EMBREY, D.E., (1975) Training of the Inspectors' Sensitivity and Response Strategy. In: Human Reliability in Quality Control. (Taylor & Francis, London pp 123-131.

ERICKSON, R.A., (1964) Relation between visual search time and peripheral visual acuity. Human Factors, 6, 165-177.

EVANS, R.N., (1951) Training improves micrometer accuracy. Personnel Psychology, 4, 231-242.

ESTES, W.K., & JOHNS, M.D., (1958). Probability learning with ambiguity in the reinforcing stimulus. American Journal of Psychology, 71, 219-228.

FARINA, A.J., & WHEATON, G.R.,(1971) Development of a Taxonomy of Human Performance. The task characteristics approach to performance prediction. American Institutes for Research, Washington Technical Report 7, Contract Nos. F44620-67-0116.

FAULKNER, T.W., & MURPHY, T.J., (1973) Lighting for difficult visual tasks. Human Factors, 15 (2), 149-162.

FECHNER, G.T., (1860) Elemente der Psychophysik (Leipzig, Breitkopf & Hartel).

FLEISHMAN, E.A., & STEPHENSON, R.W., (1970) Development of a Taxonomy of Human Performance: A review of the third years progress. Report No. AIR-726-9/70-TPR, American Institutes for Research, Washington, D.C.

FOX, J.G., (1964) The ergonomics of coin inspection. Quality Engineer, 28, 165-169.

FOX, J.G., (1973) Sustaining vigilance in inspection. Paper at E.R.S. Meeting Human Reliability in Quality Control - Birmingham.

FOX, J.G., & HASELGRAVE, C.M., (1969) Industrial inspection efficiency and the probability of a defect occurring. Ergonomics, 12, 713-721.

FREEMAN, H.A.., FRIEDMAN, M., MOSTELLER, F., & WALLIS, W.A. (Eds.), (1948), Sampling Inspection. (New York, McGraw - Hill).

FREEMAN, P.R., (1973) Tracts for Computers XXX: Table of d′ and Beta. Cambridge: Cambridge University Press).

FROOT, H.A., & DUNKEL, W.E., (1975) Visual inspection of integrated circuits: A case study. Human Reliability in Quality Control (London, Taylor & Francis).

GALE, A., BULL, R., PENFOLD, V., COLES, M., & BARRACLOUGH, R., (1972) Extraversion, time of day, vigilance performance and physiological arousal: failure to replicate traditional findings. Psychonomic Science, 29, 1-5.

GARDNER, R.W., LOHRENZ, L.J., & SCHOEN, R.B., (1968) Cognitive control of differences in the perception of persons and objects. Perceptual and Motor Skills, 26, 311-330.

GARDNER, R.W., & MORIARTY, A., (1968) Personality structure at pre-adolescence..(Seattle:University of Washington Press).

GIBSON, E.G., (1953) Improvement in perceptual judgements as a function of controlled practice or training. Psychological Bulletin, 50, 401-431.

GILLIES, G.J., (1975) Industrial applications in the glass industry. Human Reliability and Quality Control, (London: Taylor & Francis).

GRANT, D.A., & HORNSETH, J.P., (1951) Aquisition and Extinction of a verbal conditional response with differing percentages of reinforcement. Journal of Experimental Psychology, 42, 1-5

GREY, D.R., & MORGAN, B.J.T. (1972) Some aspects of ROC curve-fitting: normal and logistic models. Journal of Mathematical Psychology, Volume 9.

GRIER, J.B., (1971) Nonparametric indexes for sensitivity and bias. Psychological Bulletin Vol. No. 75.

GUILDFORD, J.P., (1954) Psychometric Methods (New York: McGraw-Hill).

HAMMERTON, M., & ALTHAM, P.M.E., (1971) A nonparametric alternative to d'. Nature, 234, 487-488.

HARRIS, D.H. (1964) Development and Validation of an aptitude test for inspectors of electronic equipment. Journal of Industrial Psychology, 2, 29-35.

HARRIS, D.H., (1966) Effect of equipment complexity on inspection performance. Journal of Applied Psychology, 50, 236-237.

HARRIS, D.H., (1968) Effect of defect rate on inspection accuracy. Journal of Applied Psychology, 52, 377-79.

HARRIS, D.H. (1969) The nature of industrial inspection. Human Factors, 11, (2), 139-148.

HARRIS, D.H. & CHANEY, F.B., (1969) Human factors in quality assurance. (New York: Wiley).

HEIMSTRA, N.W., ELLINGSTAD, V.S. & DEKOCK, A.R., (1967) Effects of operator mood on performance in a simulated driving task. Perceptual & Motor Skills, 25, 729-735.

HODOS, W., (1970) Nonparametric index of response bias for use in detection and recognition experiments. Psychological Bulletin, 74, (5).

INGLEBY, J.D., (1974) Further studies of the human observer as statistical decision maker. Organisational Behaviour and Human Performance, 12.

JACOBSON, H.J. (1953) A study of inspector accuracy. Engineering Inspection, 17, (2-10), 1953.

JAMIESON, G.H , (1966) Inspection in the telecommunications industry: A field study of age and other performance variables. Ergonomics, 9, 297-303.

JENKINS, H.M., (1958) The effect of signal rate on performance in visual monitoring. American Journal of Psychology, 71, 647-661.

KAPPAUF, W.E., CROWDER, W.F., McDIARMID, C.G., & DAVIES, J.D., (1955) Performance during prolonged watch-keeping at a visual detection task involving search. (University of Illinois Memor. Report 14-10, Urbana).

KAPPAUF, W.E., & POWE, W.E., (1959) Performance decrement on an audio visual checking task. Journal of Experimental Psychology, 57, 49-56.

KARP, S.A., (1962) A factorial study of overcoming embeddedness in perceptual and intellectual functioning. Unpublished doctoral dissertation, New York University.

KARP, S.A., (1962) Kit of selected distraction tests. Cognitive tests: State University of New York, Downstate Medical Center.

KARP, S.A., (1963) Field dependence and overcoming embeddedness. Journal of Consulting Psychology, 27, (4), 294-302.

KEPPEL, G., (1973) Design and Analysis: A Researcher's Handbook. Englewood Cliffs, New Jersey: Prentice-Hall Inc.).

KIBLER, A.W.,,(1965) The relevance of vigilance research to aerospace monitoring tasks. Human Factors 7, (2), 93-99).

KIRK, R.E., (1968) Experimental design: procedures for the behavioural sciences. (Monterey, California: Brooks/Cole.)

KIRK, R.E., & HECHT, E., (1963) Maintenance of vigilance by programmed noise. Perceptual and Motor Skills, 16, 553-560.

KRKOVIC, A., & SVERKO, B.,(1967) Characteristics of performance on a new variety of vigilance task. Acta Instituti Psychologici Universitatis Zagrebiensis, No. 54.

LAU, A.W., (1966) Doppler Discrimination as a function of variations in dimensions of the sonor echo. U.S. Naval Personnel Research Activity: Technical Bulletin STB 66-25.

LEE, W., (1969) Relationships between Thurstone category scaling and signal detection theory. Psychological Bulletin, 71, 101-7.

LEE, W., (1971) Decision Theory and Human Behaviour. (New York: Wiley).

LEVINE, J.M., ROMASHKO, T., & FLEISHMAN, E.A., (1971) Development of a taxonomy of human performance. Evaluation of an abilities classification system for integrating and generalising research findings. American Institutes for Research, Washington Technical Report 12, Contract No. DAHL 19-71-L-0004.

LINK, H.C., (1920) Employment Psychology, (New York: McMillan).

LION, J.S. (1964) The performance of manipulative and inspection tasks under Tugsten and fluorescent lighting. Ergonomics, 7, 51-61.

LION, J.S., RICHARDSON, E. & BROWNE, R.C., (1968) A study of the performance of industrial inspectors under two kinds of lighting. Ergonomics, Vol. 11, No. 1, 23-34.

LION, J.S., RICHARDSON, E., WEIGHTMAN, D., & BROWNE, R.C., (1975). The influence of the visual arrangement of material and of working singly or in pairs, upon performance at simulated industrial inspection. Ergonomics, 18, (2), 195-204.

LION, J.S., RICHARDSON, E., BROWNE, R.C., WEIGHTMAN, D. (1975) The influence of the visual arrangement of material at simulated industrial inspection. Ergonomics, 11, (1), 23-34.

LUDVIGH, E.J., & MILLER, J.W., (1958) Study of visual acuity during ocular pusuit of moving test objects. Journal of Optical Society of America, 48, 799-802.

LUSTED, L.B., (1971) Signal detectability and medical decision making. Science, 171, 1217.

MACKWORTH, J.F.,(1969) Vigilance and habituation. (Harmondsworth-Penguin Books)

MACKWORTH, J.F., (1970) Vigilance and attention. (Harmondsworth: Penguin Books)

McCann, P.H., (1969) The effects of ambient noise on vigilance performance. Human Factors, 11, 251-256.

MCCORMACK, R.L., (1961) Inspector Accuracy: A study of the literature (SCTM-53-61 (14), Sandia Laboratories, Albuquerque, New Mexico.

MCFARLING, L.H., (1974) Noise effects, sex differences, and task pacing in simulated inspection. (Ph.D., Thesis, University of South Dakota).

MCKENZIE, R.M., (1958) On the accuracy of inspectors. Ergonomics pp 225-272.

MCKENZIE, R.M., & PUGH, D.S., (1957) Some Human Aspects of Inspection Journal of the Institute of Production Engineers Vol. No. 36.

MCNICOL, D., (1972) A primer of signal detection theory. (London, George Allen and Union Ltd.)

MARTINEK, H., & SADACCA, R., (1965) Error Keys as reference aids in image interpretation. U.S. Army Personnel Research Office, Washington, Technical Research Note 153.

MENDELSOHN, G.A., GRISWOLD, B.B., & ANDERSON, M.L., (1966) Individual differences in anagram-solving ability. Psychological Reports, 1966, 19, 799-809.

MILLER, R.B., (1962) Task description and analysis. Psychological Principles in System Development (New York: Holt, Rinehart and Winston).

MILLER, R.B., (1971) Development of a taxonomy of human performance: Design of a systems task vocabulary. Technical Report II American Institutes for Research, Washington.

MITCHELL, J.H., (1935) Subjective stands in inspection for appearance. Human Factor 9, 1935.

MITTEN, L.G., (1957) Research team approach to an inspection operation. In - Introduction to Operations Research, (New York: Wiley).

MORAAL, J., (1975) Analysis of an industrial inspection task. Human Reliability and Quality Control (London, Taylor & Francis).

MORRISSETTE, J.O., HORNSETH, J.P., & SHELLAR, K. (1975) Team organisation and monitoring performance. Human Factors 17, (3), 296-300.

NAVON, D., (1975) A simple method for latency analysis in signal detection tasks. Perception and Psychophysics 18 (1), 61-64.

NEAL, G.L., & PEARSON, R.G., (1966) Comparative effects of age, sex, and drugs, upon two tasks of auditory vigilance. Perceptual and Motor Skills, 23, 957-974.

NELSON, J.B., & BARANY, J.W., (1969) A dynamic visual recognition test for paced inspection tasks. AIIE Transactions, 1, (4), 327-332.

OGILVIE, J.C., & CREELMAN. C.D., (1968) Maximum likelihood estimation of Receiver Operating Characteristics curve parameters. Journal of Mathematical Psychology, 5, 377-91

PASTORE, R.E., & SCHEIRER, C.J., (1974) Signal Detection Theory: Considerations for General Application. Psychological Bulletin, 81, No. 12, 945-958.

PEMBERTON, C.L., (1951) A study of the speed and flexibility of closure factors. Unpublished Ph.D. dissertation, Department of Psychology, University of Chicago.

PERRY, G., (1968) Lighting for Inspection. (Technical Note No. 115., British Glass Industry Research Association).

PIKE, A.R., (1971) The latencies of correct and incorrect responses in discrimination and detection tasks: Their interpretation in terms of a model based on a simple counting. Perception & Psychophysics, 9, pp 455-460.

PINNEO, L., (1966) On noise in the nervous system. Psychological Review, Vol. No. 73, No. 3.

POLLACK, I., & HSIEH, R., (1969) Sampling variability of the area under the ROC curve and of d'. Psychological Bulletin Vol. 71, No. 3.

POLLACK, I,, & NORMAN, D.A., (1964) A nonparametric analysis of recognition experiments. Psychonomic Science, Vol. No. 1.

POLLACK, I., NORMAN, D.A., GALANTER, E., (1964) An efficient non-parametric analysis of recognition memory. Psychonomic Science 1, 327-328.

POULTON, E.C., (1969) Bias in experimental comparisons between equipments due to the order of testing. Ergonomics, 12, (4), 679-687

POULTON, E.C., (1973) The effect of fatigue upon inspection work. Applied Ergonomics, 4.2, 73-83.

POWERS, J.R., BRAINARD, R.W., ABRAM., R.E., & SADACCA, R., (1973) Training techniques for rapid target detection. U.S. Army Research Institute for the Behavioural & Social Sciences, Technical Paper 242.

PURSWELL, J.L., etal. (1972) An inspection task experiment. Proceedings of the 16th Annual Meeting of the Human Factors Society, Vol. 1., 297-300.

RAPHAEL, W.S., (1942) Some Problems of Inspection. Occupational Psychology, 16, (4), 157-163.

RIZZI, A.M., BUCK, J.R., & ANDERSON, V.L., (1974) Effects of some task variables on conver-paced visual inspection accuracy. (Working Paper, School of Industrial Engineering, Purdue University.)

ROETHLISBERGER, F.J., & DICKSON, W.J., (1939) Management and the Worker. (Cambridge: Harvard University Press).

SACK, S.A., & RICE, C.E., (1974) Selectivity, resistance to distraction and shifting as three attentional factors. Psychological Reports, 34, 1003-1012.

SAKGUCHI, T., & NAGAI, H., (1973) Studies on relation between various light sources and visual fatigue. Journal of the Illuminating Engineering Institute of Japan, 57, (5), 4-13.

SANGUILIANO, I.A., (1951) An investigation of the relationship between the perception of the upright in space and several factors in personality organization. Unpublished doctoral dissertation, Fordham University.

SARTAIN, A.Q., (1945) The use of certain standardized tests in the selection of inspectors in an aircraft factory. Journal of Consulting Psychology, 9, 234-235.

SCHLEGEL, R E., BOARDMAN, D.W. & PURSWELL, J.C., (1973) A comparison of single and multiple inspection systems. Proceedings of the 17th Annual Meeting of the Human Factors Society, Santa Monica, California.

SCHOONARD, J.W., & GOULD, J.D. (1973) Field of view and target uncertainty in visual search and inspection. Human Factors, 15, (1), 33-42.

SCHUMAN, J.T., (1945) The value of aptitude tests for factory workers in the aircraft engine and propller industries. Journal of Applied Psychology, 29, 156-163.

SCOTT BLAIR, G W , & COPPEN, F.M.V., (1942) The subjective conception of the firmness of soft materials. American Journal of Psychology, 5, 127-139.

SEALE, S.J., (1972) Some psychometrics in relation to target acquisition. British Aircraft Corporation Ltd., Guided Weapons Division: Bristol Works, Target Acquisition Research Group B15-1-5.

SHEEHAN, J.J., & DRURY, C.G., (1971) The analysis of industrial inspection. Applied Ergonomics, 2, 74-78.

SHEFT, D.J., JONES, M.D., BROWN, R.F., & Ross, S.E. (1970) Radiology, 94, 427.

SIEGEL, S., & GOLDSTEIN, D.A., (1959) Decision-making behavior in a two-choice uncertain outcome situation. Journal of Experimental Psychology, 57, 31-42.

SIEGEL, A.I., & WOLF, J., (1969) Man Machine Simulation Models. (New York: Wiley).

SIMS, L.P., (1972) "An analysis of inspector perception of quality levels". M.Sc. Thesis, Auburn University, Alabama.

SINCLAIR, M., (1971) Paced and unpaced inspection of a bakery product. Paper at E.R.S. Annual Conference.

SLOVIC, P., FISCHHOFF, B., & LICHTENSTEIN, S., (1975) Cognitive Processes and Societal Risk Taking. Oregon Research Institute.

SMITH, G.L. JR., (1975) Signal detection theory and industrial inspection. In, Human Reliability & Quality Control (London: Taylor & Francis)

SMITH, G.L., & ADAMS, S.K. (1971) Magnification and microminiature inspection. Human Factors, 13, (3), 247-254.

SMITH, L.A., & BARANY, J.W., (1970) An elementary model of human performance on paced visual inspection tasks. AIIE Transactions, Volume II, No. 4., 298-308, 1970.

SMITH, L.A., & BARANY, J.W., (1971) An elementary model of human performance on paced visual inspection. A.I.I.E. Transactions, 4, 298-308.

SMITH, R.L., LUCACCINI, L.F., GROTH, H, & LYMAN, J., (1966) Effects of anticipatory alerting signals and a compatible socondary task on vigilance performance. Journal of Applied Psychology, 50, 240-246

SMITH, R.L., & LUCACCINI, L.F., (1969) Vigilance Research: Its application to industrial problems. Human Factors, 11, 149-156.

SOSNOWEY, J.K., (1967) An investigation of the effects of incoming quality and inspection rate on inspector accuracy. (M.S.I.E. Thesis, Texas Technological College).

STAEL von HOLSTEIN, C.A.S., (1971) The effect of learning on the assessment of subjective probability distributions. Organisational Behaviour and Human Performance, 6, 304-315.

SURY, R.J., (1964) An industrial study of paced and unpaced operator performance in a single stage work task. International Journal of Production Research 3, (1), 91-98.

SWAIN, A.D., (1972) Design techniques for improving human performance in production.(London: Industrial and Commercial Techniques Ltd.)

SWETS, J.A., (1963a) Control factors in auditory frequency selectivity. Psychological Bulletin, 60, 429-440.

SWETS, J.A., (1964) Signal Detection and recognition by human observers. (New York: Wiley).

SWETS, J.A., (1973) The relative operating characteristics in Psychology Science, (Vol. 182).

SWETS, J.A., & GREEN, D.M., (1966) Signal Detection Theory and Psychophysics). (New York: Wiley)

SWETS, J.A., MILLMAN, S.H., FLETCHER, W.E., & GREEN, D.M., (1962) Learning to identify non verbal sounds: An application of a computer as a teaching machine. Journal of Acoustical Society of America, 34, 928-935.

SWETS, J.A., TANNER, W.P., & BIRDSALL, T.G., (1961) Decision processes in perception. Psychological Review, 68, 301-340.

TANALSKI, T.G., (1956) The eyes have it - At Convair. Industrial Quality Control, 12, 9-10.

TAYLOR, M.M., (1967) Detectability theory and the interpretation of vigilance data. Acta Psychologica, 27, 390-399.

THEODOR, L.H., (1972) Some comments on 'A table for the calculation of d' & beta'.

THOMAS, L.F., & SEABORNE, A.E.M., (1961) The socio-technical context of industrial inspection. Occupational Psychology Vol. 35, pp 36-43.

THOMAS, L.F., (1962) Perceptual organization in industrial inspectors. Ergonomics, 5, 429-434.

THOMPSON, L.W., OPTON, E., & COHEN, L.D., (1963) Effects of age, presentation speed, and sensory modality on performance of a vigilance task. Journal of Gerontology, 18, 366-369.

THORNTON, C.L., BARRETT, G.V. & DAVIS J.A., (1968) Field dependence and target identification. Human Factors, 10, (5), 493-496.

THURSTONE, L.L., (1944) A Factual Study of Perception. Psychometric Monographs No. 4 (Chicago: University of Chicago Press).

THURSTONE, L.L., & JEFFREY, T.E., (1965) Closure Flexibility (Concealed Figures) Test (Industrial Relations Center, University of Chicago).

TIFFIN, J., & ROGERS, H.B. (1941) The selection and training of inspectors. Personnel, 18, (1), 14-31.

TRIMBY, H. (1959) The development of industrial vision screening in the United States, with special reference to work done with the B & L Ortho-Rater. "The Optician".

TUKEY, J.W., (1949) One degree of freedom for non additivity. Biometrics, 5, 232-242.

VIRSU, V., (1972) Inspection and presentation of radiographs and micrographs: An applied review of basic visual parameters. Reports from the Institute of Psychology, University of Helsinki, Report No. 3.

WAAG, W L., HALCOMB, C.G. & TYLER, D.M., (1973) Sex differences in monitoring performance. Journal of Applied Psychology, 58, 272-274.

WALLACK, P M., (1967) An experimental investigation of industrial inspector accuracy under varying levels of product defectiveness. (Unpublished Doctoral Dissertation, Oklahoma State University, Stillwater, Oklahoma).

WALLACK, P.M., & ADAMS, S.K., (1969) The utility of signa-detection theory in the analysis of industrial inspector accuracy. AIIE Transactions, 1, (1), 33-44.

WALLACK, P.M., & ADAMS, S.K., (1970) A comparison of inspector performance measures. AIIE Transactions, II,(2), 97-105.

WALLIS, D., (1963) Occupational Psychology in the sixties: Some implications of recent studeies of perceptual training and skill. Occupational Psychology, 37, No. 4, 237-253.

WELFORD, N.T., (1952) S.E.T.A.R. A sequentail event timing and recording apparaturs. Journal of Scientific Instruments, 29, 1-4.

WHITFIELD, D., (1975) Man-computer symbiosis: A 1975 review. A.P. Report 57, Applied Psychology Department, Unviersity of Aston in Birmingham.

Whittenberg, J.A., & ROSS, S., (1953) A study of three measures of perceptual efficiency during sustained vigilance (University of Maryland Tech. Rep. 14, College Park.

WIENER, E.L., (1963) Knowledge of results and signal rate in monitoring: A transfer of training approach. Journal of Applied Psychology, 47, 214-222.

WIENER, E.L., (1967) Transfer of training from one monitoring task to another. Ergonomics, 10, 649-58

WIENER, E.L., (1968) Training for vigilance - repeated sessions with knowledge of the results. Ergonomics, 11, 547-556.

WIENER, E.L., (1969) Money and the monitor. Perception and Motor Skills, 29, 627-634.

WIENER, E.L. (1975) Individual and group differences in inspection. In: Human reliability and quality control (London:-Taylor & Francis).

WIENER, E.L., & ATTWOOD, D.A., (1968) Training for vigilance. Combined cueing and knowledge of results. Journal of Applied Psychology, 6, 474-479.

WILKINSON, R.T., (1960) The effect of lack of sleep on visual watch keeping. Quarterly Journal of Experimental Psychology, 12, (1), 36-40.

WILKINSON, R.T., (1964) Artificial 'signals' as an aid to an inspection task. Ergonomics, 7, 63-72.

WILLIAMS, L.G., & BOROW, M.S., (1963) The effect of rate and direction of display movement upon visual search. Human Factors, 5, (2), 139-146.

WILLIGES, R.C. & STREETER, H. (1971) Display Characteristics in Inspection Tasks. Journal of Applied Psychology 55, (2). 123-125.

WINER, B.J., (1962) Statistical Principles in experimental design. (New York: McGraw-Hill).

WITKIN, H.A., (1950) Individual Differences in Ease of Perception of Embedded figures. Journal of Personality 19, 1-15.
Perception

WITKIN, H.A., LEWIS, H.B., HERTZMAN, M., MACHOVER, K., MEISSNER, P.B., & WARNER, S., (1954) Personality through perception. (New York: Harper).

WITKIN, H.A., DYK, R.B., FATERSON, H.F., GOODENOUGH, D.R., & KARP, S.A. (1962) Psychological Differentiation (New York: Wiley).

WITKIN, H.A., Oltman, P.K., RASKIN, E., KARP, S.A., (1971). A manual for the embedded figures test. (Palo Alto: Consulting Psychologists Press).

WYATT, S., & LANGDON, J.N., (1932), Inspection Processes in Industry. (Industrial Health Research Board Report 63, London).

ZUNZANYIKA, X.K. , & DRURY, C.G., (1975) Effects of information on industrial inspection performance. Human Reliability and Quantity Control (London: Taylor & Francis).

Appendices

Appendix A : Statistical Analysis

Experiment 1

" 3

" 4

" 5

" 6

" 7

Appendix B : Data

Appendix C : The experimental apparatus and its associated control program

Appendix D : Analysis programs

Because of space limitations only a selection of the statistical analyses referred to in the text are to be found in Appendix A. Appendix B contains the raw test scores of the cognitive tests. All other data and statistical analyses are available at the Department of Applied Psychology, The University of Aston in Birmingham. For similar reasons, Appendix D only contains the most important programs used.

APPENDIX A : EXPERIMENT 1

ANALYSIS OF VARIANCE. . . . 2ASIN CORRECT DETECTION PROBABILITY

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 1.80137

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	0.20974	2	0.10487	2.3987	2,24	0.11069
P	0.01117	1	0.01117	0.2555	1,24	0.62322
NP	0.06263	2	0.03132	0.7163	2,24	0.50278
T	0.41057	2	0.20529	4.6954	2,24	0.01861*
NT	0.14627	4	0.03657	0.8364	4,24	0.51720
PT	0.07414	2	0.03707	0.8478	2,24	0.44404
NPT	0.18590	4	0.04647	1.0630	4,24	0.39701
S	1.18058	6	0.19676	4.5005	6,24	0.00371*
NS	0.24648	12	0.02054	0.4698	12,24	0.91313
PS	0.57977	6	0.09663	2.2101	6,24	0.07689
NPS	0.30784	12	0.02565	0.5868	12,24	0.83165
TS	0.45288	12	0.03774	0.8632	12,24	0.59187
NTS	0.70647	24	0.02944	0.6733	24,24	0.83041
PTS	0.35515	12	0.02960	0.6769	12,24	0.75703
NPTS	1.04929	24	0.04372	1.0000	24,24	0.50000
TOTAL	5.97888	125				

ANALYSIS OF VARIANCE. . . . A1 INDEX

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 92.92094

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	13.54423	2	6.77212	0.5235	2,24	0.60416
P	2.27998	1	2.27998	0.1762	1,24	0.68052
NP	30.01689	2	15.00844	1.1602	2,24	0.33090
T	13.06806	2	6.53403	0.5051	2,24	0.61490
NT	99.98139	4	24.99535	1.9322	4,24	0.13699
PT	2.92952	2	1.46476	0.1132	2,24	0.88801
NPT	72.09114	4	18.02279	1.3932	4,24	0.26559
S	323.81124	6	53.96854	4.1719	6,24	0.00542**
NS	137.25577	12	11.43798	0.8842	12,24	0.57357
PS	137.40765	6	22.90128	1.7703	6,24	0.14771
NPS	134.20350	12	11.18363	0.8645	12,24	0.59072
TS	326.31057	12	27.19255	2.1020	12,24	0.05846
NTS	337.39571	24	14.05815	1.0867	24,24	0.42016
PTS	107.54167	12	8.96181	0.6928	12,24	0.74328
NPTS	310.46919	24	12.93622	1.0000	24,24	0.50000
TOTAL	2048.30652	125				

ANALYSIS OF VARIANCE. . . D'

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 2.79303

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	0.53628	2	0.26814	0.8619	2,24	0.43815
P	0.02309	1	0.02309	0.0742	1,24	0.77751
NP	0.89781	2	0.44890	1.4430	2,24	0.25518
T	1.00675	2	0.50337	1.6181	2,24	0.21796
NT	1.96700	4	0.49175	1.5807	4,24	0.21096
PT	0.27749	2	0.13874	0.4460	2,24	0.65063
NPT	1.90961	4	0.47740	1.5346	4,24	0.22326
S	8.21668	6	1.36945	4.4021	6,24	0.00415**
NS	2.72701	12	0.22725	0.7305	12,24	0.71006
PS	3.60381	6	0.60064	1.9307	6,24	0.11641
NPS	2.37865	12	0.19822	0.6372	12,24	0.79081
TS	5.48413	12	0.45701	1.4691	12,24	0.20358
NTS	5.37898	24	0.22412	0.7204	24,24	0.78628
PTS	3.05035	12	0.25420	0.8171	12,24	0.63269
NPTS	7.46619	24	0.31109	1.0000	24,24	0.50000
TOTAL	44.92381	125				

ANALYSIS OF VARIANCE....LOG CORRECTED BETA

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 0.72958

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	1.84492	2	0.92246	2.4484	2,24	0.10614
P	0.47117	1	0.47117	1.2506	1,24	0.27405
NP	0.62900	2	0.31450	0.8348	2,24	0.44959
T	3.88696	2	1.94348	5.1584	2,24	0.01352*
NT	1.32725	4	0.33181	0.8807	4,24	0.49187
PT	0.49871	2	0.24936	0.6618	2,24	0.52934
NPT	0.82255	4	0.20564	0.5458	4,24	0.70646
S	1.17037	6	0.19506	0.5177	6,24	0.79019
NS	5.09648	12	0.42471	1.1273	12,24	0.38445
PS	3.29944	6	0.54991	1.4596	6,24	0.23354
NPS	5.47149	12	0.45596	1.2102	12,24	0.33115
TS	15.63257	12	1.30271	3.4577	12,24	0.00492**
NTS	16.02485	24	0.66770	1.7722	24,24	0.08409
PTS	5.01980	12	0.41832	1.1103	12,24	0.39613
NPTS	9.04219	24	0.37676	1.0000	24,24	0.50000
TOTAL	70.23776	125				

ANALYSIS OF VARIANCE. . . . ZFA (BIAS INDEX)

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 1.76842

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	0.04747	2	0.02373	0.2792	2,24	0.76165
P	0.13298	1	0.13298	1.5645	1,24	0.22108
NP	0.26647	2	0.13323	1.5675	2,24	0.22807
T	0.24574	2	0.12287	1.4456	2,24	0.25458
NT	0.58374	4	0.14594	1.7169	4,24	0.17844
PT	0.02311	2	0.01155	0.1359	2,24	0.86971
NPT	0.46622	4	0.11655	1.3713	4,24	0.27284
S	0.83669	6	0.13945	1.6406	6,24	0.17898
NS	1.45440	12	0.12120	1.4259	12,24	0.22113
PS	0.46697	6	0.07783	0.9157	6,24	0.50165
NPS	1.32573	12	0.11048	1.2998	12,24	0.28067
TS	2.78422	12	0.23202	2.7297	12,24	0.01743
NTS	3.38231	24	0.14093	1.6580	24,24	0.11143
PTS	1.29286	12	0.10774	1.2675	12,24	0.29801
NPTS	2.03995	24	0.08500	1.0000	24,24	0.50000
TOTAL	15.34886	125				

APPENDIX A : EXPERIMENT 3

ANALYSIS OF VARIANCE. . . D'

LEVELS OF FACTORS

T	4
R	2
S	3

GRAND MEAN 0.95118

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.69767	3	0.23256	0.9934	3, 6	0.45849
R	0.02642	1	0.02642	0.1128	1, 6	0.74237
TR	0.25841	3	0.08614	0.3679	3, 6	0.78026
S	8.43197	2	4.21599	18.0092	2, 6	0.00363**
TS	2.61940	6	0.43657	1.8649	6, 6	0.23313
RS	0.46714	2	0.23357	0.9977	2, 6	0.42457
TRS	1.40461	6	0.23410	1.0000	6, 6	0.50000
TOTAL	13.90561	23				

ANALYSIS OF VARIANCE. . . LOG BETA

LEVELS OF FACTORS

T	4
R	2
S	3

GRAND MEAN 0.61572

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.04966	3	0.01655	0.0763	3, 6	0.96596
R	0.00219	1	0.00219	0.0101	1, 6	0.88544
TR	0.95239	3	0.31746	1.4628	3, 6	0.31571
S	2.76177	2	1.38089	6.3627	2, 6	0.03306
TS	0.86914	6	0.14486	0.6675	6, 6	0.68284
RS	0.20630	2	0.10315	0.4753	2, 6	0.64663
TRS	1.30217	6	0.21703	1.0000	6, 6	0.50000
TOTAL	6.14362	23				

APPENDIX A : EXPERIMENT 4

ANALYSIS OF VARIANCE. . . C. D. PROBABILITY

LEVELS OF FACTORS

T	5
F	2
P	2
S	3

GRAND MEAN 0.62367

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.19623	4	0.04906	1.9985	4, 8	0.18752
F	0.30817	1	0.30817	37.5508	1, 2	0.02196*
TF	0.03593	4	0.00898	1.4207	4, 8	0.31071
P	0.04374	1	0.04374	0.2790	1, 2	0.42384
TP	0.05609	4	0.01402	0.7575	4, 8	0.46190
FP	0.00417	1	0.00417	0.0328	1, 2	0.42384
TFP	0.02327	4	0.00582	0.8738	4, 8	0.46190
S	0.37529	2	0.18765	1.0000	2, 2	0.50000
TS	0.19637	8	0.02455	1.0000	8, 8	0.50000
FS	0.01641	2	0.00821	1.0000	2, 2	0.50000
TFS	0.05059	8	0.00632	1.0000	8, 8	0.50000
PS	0.31356	2	0.15678	1.0000	2, 2	0.50000
TPS	0.14811	8	0.01851	1.0000	8, 8	0.50000
FPS	0.25441	2	0.12721	1.0000	2, 2	0.50000
TFPS	0.05325	8	0.00666	1.0000	8, 8	0.50000
TOTAL	2.07559	59				

ANALYSIS OF VARIANCE....F.A. PROBABILITY

LEVELS OF FACTORS

T	5
F	2
P	2
S	3

GRAND MEAN 0.14552

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.04076	4	0.01019	3.7911	4, 8	0.05149
F	0.01157	1	0.01157	0.1123	1, 2	0.42384
TF	0.02068	4	0.00517	1.5042	4, 8	0.28793
P	0.30944	1	0.30944	28.0542	1, 2	0.03033
TP	0.01193	4	0.00298	0.9859	4, 8	0.46190
FP	0.01157	1	0.01157	0.5401	1, 2	0.42384
TFP	0.00927	4	0.00232	0.6158	4, 8	0.46190
S	0.04283	2	0.02141	1.0000	2, 2	0.50000
TS	0.02150	8	0.00269	1.0000	8, 8	0.50000
FS	0.20605	2	0.10303	1.0000	2, 2	0.50000
TFS	0.02749	8	0.00344	1.0000	8, 8	0.50000
PS	0.02206	2	0.01103	1.0000	2, 2	0.50000
TPS	0.02420	8	0.00302	1.0000	8, 8	0.50000
FPS	0.04286	2	0.02143	1.0000	2, 2	0.50000
TFPS	0.03010	8	0.00376	1.0000	8, 8	0.50000
TOTAL	0.83231	59				

ANALYSIS OF VARIANCE. . . LOG BETA

LEVELS OF FACTORS

T	5
F	2
P	2
S	3

GRAND MEAN 0.69509

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	3.10849	4	0.77712	7.1302	4, 8	0.00997**
F	0.15828	1	0.15828	0.0600	1, 2	0.42384
TF	0.94028	4	0.23507	1.6606	4, 8	0.25042
P	10.40656	1	10.40656	9.0236	1, 2	0.09470
TP	1.56394	4	0.39098	2.1261	4, 8	0.16883
FP	2.06128	1	2.06128	42.9889	1, 2	0.01885*
TFP	0.13338	4	0.03334	0.2125	4, 8	0.46190
S	1.55469	2	0.77734	1.0000	2, 2	0.50000
TS	0.87192	8	0.10899	.1.0000	8, 8	0.50000
FS	5.27277	2	2.63639	1.0000	2, 2	0.50000
TFS	1.13248	8	0.14156	1.0000	8, 8	0.50000
PS	2.30652	2	1.15326	1.0000	2, 2	0.50000
TPS	1.47118	8	0.18390	1.0000	8, 8	0.50000
FPS	0.09590	2	0.04795	1.0000	2, 2	0.50000
TFPS	1.25505	8	0.15688	1.0000	8, 8	0.50000
TOTAL	32.33271	59				

ANALYSIS OF VARIANCE. . . . SCORE

LEVELS OF FACTORS

T	5
F	2
P	2
S	3

GRAND MEAN 174.98333

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	6213.06665	4	1553.26666	2.2969	4, 8	0.14720
F	13771.35010	1	13771.35010	0.7560	1, 2	0.42384
TF	3266.40002	4	816.60001	0.9977	4, 8	0.46190
P	*****.*****	1	*****.*****	156.2018	1, 2	0.00445**
TP	4696.40002	4	1174.10001	1.7897	4, 8	0.22378
FP	1118.01666	1	1118.01666	0.1868	1, 2	0.42384
TFP	293.06667	4	73.26667	0.0715	4, 8	0.46190
S	21105.73340	2	10552.86670	1.0000	2, 2	0.50000
TS	5409.93335	8	676.24167	1.0000	8, 8	0.50000
FS	36433.20019	2	18216.60010	1.0000	2, 2	0.50000
TFS	6547.80005	8	818.47501	1.0000	8, 8	0.50000
PS	2172.93335	2	1086.46667	1.0000	2, 2	0.50000
TPS	5248.40002	8	656.05000	1.0000	8, 8	0.50000
FPS	11970.13330	2	5985.06665	1.0000	2, 2	0.50000
TFPS	8192.53345	8	1024.06668	1.0000	8, 8	0.50000
TOTAL	296146.96873	59				

APPENDIX A : EXPERIMENT 5

ANALYSIS OF VARIANCE. . . . ASIN C. D. PROBABILITY

LEVELS OF FACTORS

T	5
C	3
S	3

GRAND MEAN 1.40293

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.28675	4	0.07169	5.0032	4, 8	0.02585*
C	0.02973	2	0.01486	0.0709	2, 4	0.44610
TC	0.21039	8	0.02630	1.8576	8,16	0.13881
S	0.91364	2	0.45682	1.0000	2, 2	0.50000
TS	0.11463	8	0.01433	1.0000	8, 8	0.50000
CS	0.83896	4	0.20974	1.0000	4, 4	0.50000
TCS	0.22653	16	0.01416	1.0000	16,16	0.50000
TOTAL	2.62062	44				

ANALYSIS OF VARIANCE....ASIN F.A. PROBABILITY

LEVELS OF FACTORS

T	5
C	3
S	3

GRAND MEAN 0.70090

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.62792	4	0.15698	7.0379	4, 8	0.01034*
C	3.26910	2	1.63455	15.5388	2, 4	0.01487*
TC	0.73618	8	0.09202	7.5128	8,16	0.00055***
S	0.34964	2	0.17482	1.0000	2, 2	0.50000
TS	0.17844	8	0.02231	1.0000	8, 8	0.50000
CS	0.42077	4	0.10519	1.0000	4, 4	0.50000
TCS	0.19598	16	0.01225	1.0000	16,16	0.50000
TOTAL	5.77802	44				

ANALYSIS OF VARIANCE....LOG BETA

LEVELS OF FACTORS

T	5
C	3
S	3

GRAND MEAN 0.80606

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	6.25840	4	1.56460	12.2851	4, 8	0.00222**
C	18.14703	2	9.07351	86.1783	2, 4	0.00157**
TC	6.71521	8	0.83940	9.9151	8,16	0.00017**
S	3.01589	2	1.50795	1.0000	2, 2	0.50000
TS	1.01886	8	0.12736	1.0000	8, 8	0.50000
CS	0.42115	4	0.10529	1.0000	4, 4	0.50000
TCS	1.35454	16	0.08466	1.0000	16,16	0.50000
TOTAL	36.93109	44				

APPENDIX A : EXPERIMENT 6

1.63339	2.73116	1.53032	2.56538	1.43255
0.09577	0.09063	-0.09354	0.56225	0.26783
2.59809	0.77063	-0.88734	-1.02837	1.93695
-0.09924	0.05910	-0.18079	-0.03360	0.10574
0.66481	1.50698	0.37168	0.67528	0.25862
0.60346	0.20698	0.84227	0.57748	-0.02517
9.95465	2.65741	0.40815	-13.80570	-2.33217
10.73728	12.98660	4.96505	11.47474	0.68046
0.37835	0.02644	-0.78374	0.61847	0.21812
1.02444	1.51353	2.15374	1.15869	1.60965
5.39828	0.55667	9.32736	8.66278	-2.94146
1.34264	0.71353	4.44813	6.12654	10.02842
-0.29103	0.78550	1.15864	0.20424	-0.99834
-0.07581	-0.33091	0.64082	-0.51076	-0.13069
1.59155	2.77815	5.40200	10.80911	3.08859
0.83843	0.50670	1.18631	0.90100	2.73391
0.04041	0.68924	0.39752	1.82370	-0.05261
0.56836	0.73176	0.75698	1.80200	1.14269
0.23882	-0.02714	0.13314	-1.07175	0.17225
0.19806	1.72183	-0.71142	0.37445	1.32868
1.50988	1.72619	-3.69463	1.09295	1.35023

ANALYSIS OF VARIANCE. . . BETA

LEVELS OF FACTORS.

A	5
S	7
C	3

GRAND MEAN 1.44357

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
A	9.69997	4	2.42499	0.2546	4,24	0.42801
S	257.41335	6	42.90222	1.0000	6, 6	0.50000
AS	228.62827	24	9.52618	1.0000	24,24	0.50000
C	86.12684	2	43.06342	3.9802	2,12	0.04639
AC	84.40780	8	10.55098	1.4074	8,48	0.21727
SC	129.83274	12	10.81940	1.0000	12,12	0.50000
ASC	359.84792	48	7.49683	1.0000	48,48	0.50000
TOTAL	1155.95691	104				

APPENDIX A : EXPERIMENT 7

-0.16563	0.32076	0.63364	0.10020
-0.13822	0.46137	0.86600	0.72724
0.35593	0.86594	0.44707	-0.27972
0.26099	-0.00374	1.42711	0.18099
0.46161	0.41297	0.12714	0.13994
-1.18051	-2.29181	-1.85298	-2.43281
-0.85628	-1.06143	-1.33867	-1.75247
-1.23795	-1.41369	-2.88128	-1.93901
-0.48175	-2.57972	-2.63432	-0.76164
-2.03946	-1.21843	-1.66506	-0.45850

ANALYSIS OF VARIANCE. . . D'

LEVELS OF FACTORS.

A	2
B	2
C	5

GRAND MEAN 0.36008

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
A	0.09108	1	0.09108	1.4337	1, 4	0.29760
B	0.11822	1	0.11822	1.8608	1, 4	0.24387
AB	0.76633	1	0.76633	12.0629	1, 4	0.02625*
C	0.20077	4	0.05019	0.7901	4, 4	0.50000
AC	0.54530	4	0.13632	2.1459	4, 4	0.23830
BC	1.19526	4	0.29881	4.7036	4, 4	0.08273
ABC	0.25411	4	0.06353	1.0000	4, 4	0.50000
TOTAL	3.17108	19				

```
  -28.00000    62.00000   440.00000    31.00000
  161.00000   193.00000   325.00000   405.00000
  270.00000   320.00000  -103.00000   172.00000
  437.00000    34.00000   631.00000    77.00000
  -52.00000   190.00000   135.00000   111.00000
 -973.00000-1218.00000-1076.00000-1182.00000
 -855.00000 -751.00000 -878.00000-1105.00000
 -609.00000 -904.00000-1091.00000-1119.00000
 -723.00000-1197.00000-1144.00000 -616.00000
 -644.00000 -578.00000 -874.00000 -539.00000

ANALYSIS OF VARIANCE. . . SCORE


LEVELS OF FACTORS
  A            2
  B            2
  C            5


GRAND MEAN              190.55000


SOURCE OF      SUMS OF    DEGREES OF    MEAN      F-RATIO    DF     P
VARIATION      SQUARES     FREEDOM     SQUARES

 A          19282.04980          1   19282.04980   0.9863   1, 4 0.3758
 B          20288.45019          1   20288.45019   1.0378   1, 4 0.3676
 AB         20672.45019          1   20672.45019   1.0574   1, 4 0.3635
 C          *****.*****          4   31069.67480   1.5893   4, 4 0.3314
 AC         *****.*****          4   69135.92577   3.5364   4, 4 0.1250
 BC         *****.*****          4   36904.07519   1.8877   4, 4 0.2759
 ABC        78199.30078          4   19549.82519   1.0000   4, 4 0.5000

 TOTAL     686880.95312         19
```

-0.09000	0.08000	-0.07000	0.04000
-0.29000	0.00000	0.06000	-0.01000
-0.06000	0.15000	-0.09000	-0.09000
-0.22000	0.04000	-0.24000	0.00000
-0.08000	-0.17000	-0.03000	-0.03000
-0.43000	-0.63000	-0.64000	-0.50000
-0.37000	-0.51000	-0.48000	-0.57000
-0.58000	-0.51000	-0.62000	-0.72000
-0.22000	-0.60000	-0.58000	-0.45000
-0.59000	-0.36000	-0.56000	-0.38000

ANALYSIS OF VARIANCE. . . C.D. PROBABILITY

LEVELS OF FACTORS.

A	2
B	2
C	5

GRAND MEAN -0.05500

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
A	0.06272	1	0.06272	8.1534	1, 4	0.04634*
B	0.00162	1	0.00162	0.2106	1, 4	0.37583
AB	0.01568	1	0.01568	2.0383	1, 4	0.22603
C	0.02445	4	0.00611	0.7946	4, 4	0.50000
AC	0.04453	4	0.01113	1.4472	4, 4	0.36379
BC	0.05553	4	0.01388	1.8047	4, 4	0.28987
ABC	0.03077	4	0.00769	1.0000	4, 4	0.50000
TOTAL	0.23530	19				

-0.01000	-0.00750	-0.18500	0.00000
-0.16250	-0.13250	-0.16750	-0.14750
-0.15000	-0.10000	-0.02500	-0.00250
-0.25000	0.00250	-0.37750	-0.05250
-0.09250	-0.17500	-0.03000	-0.07000
-0.08750	-0.02500	-0.06750	-0.00750
-0.11750	-0.15000	-0.08250	-0.05750
-0.15000	-0.08250	-0.00500	-0.08750
-0.10500	-0.01000	-0.00750	-0.18750
-0.03500	-0.05750	-0.06500	-0.22250

ANALYSIS OF VARIANCE. . . F. A. PROBABILITY

LEVELS OF FACTORS .

A	2
B	2
C	5

GRAND MEAN -0.10675

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
A	0.02926	1	0.02926	16.8988	1, 4	0.01583*
B	0.00002	1	0.00002	0.0116	1, 4	0.37583
AB	0.00338	1	0.00338	1.9520	1, 4	0.23444
C	0.04313	4	0.01078	6.2274	4, 4	0.05387
AC	0.06860	4	0.01715	9.9036	4, 4	0.02592*
BC	0.03481	4	0.00870	5.0260	4, 4	0.07488
ABC	0.00693	4	0.00173	1.0000	4, 4	0.50000
TOTAL	0.18613	19				

APPENDIX B

Scores from cognitive tests, main subjects

SUBJECTS	GEFT	CLOSURE FLEX.	DCT 1	DCT 2A	DCT 2B	ARITH. OPS.	CANC.	Δ's	ANAG.
1	16	124	20	13	13	16	58	47	20
2	14	86	28	13	13	14	64	69	8
3	12	74	19	11	9	16	59	60	7
4	18	139	32	13	12	23	81	55	20
5	8	74	20	10	12	17	77	49	2
6	15	102	32	13	12	21	69	55	6
7	14	85	26	13	9	12	38	42	6
8	14	92	32	12	12	18	65	66	5
9	16	100	26	13	11	13	46	47	3
10	6	60	24	12	11	20	54	35	8
11	16	95	28	13	10	16	42	64	7
12	9	58	28	13	12	17	87	74	2
13	17	128	32	13	13	24*	62	66	9
14	14	65	28	13	11	17	56	41	8
15	9	63	26	9	12	13	36	41	4
16	14	104	31	13	13	21	63	50	12
17	16	126	24	13	12	22*	58	56	18

Scores from cognitive tests Ilford subjects

SUBJECTS	GEFT	CLOSURE FLEX.	DCT 1	DCT 2A	DCT 2B	ARITH. OPS.	CANC.	Δ's	ANAG.
1	14	58	19	13	11	11	57	32	9
2	17	74	19	12	11	13	50	30	16
3	16	92	31	13	13	21	62	32	20
4	17	106	28	12	13	18	62	44	14
5	5	20	26	13	11	8	29	44	10
6	6	30	24	13	12	2	46	26	4
7	13	72	16	13	7	12	52	44	11
8	17	88	29	13	11	17	48	64	7
9	13	38	24	5	10	10	58	35	11
10	2	52	17	11	9	14	58	28	13
11	5	57	18	11	12	13	52	35	13
12	5	70	24	13	9	3	60	35	2
13	15	76	29	13	10	6	62	38	3
14	8	42	16	11	6	9	31	20	11
15	17	86	27	13	12	14	66	47	14

APPENDIX C

APPENDIX C

The experimental apparatus and its associated control program

General description

The experimental apparatus provided an extremely flexible means of performing a wide range of experiments in which visual stimuli were presented to subjects by means of slides.

As described in Chapter 6, the apparatus consisted basically of two Kodak Carousel projectors, a Compur magnetic shutter and a bank of press buttons for responses. A total of 12 response buttons were available, although only six were utilized in this study.

The whole experiment was controlled by a PDP-8E computer, which presented the stimuli and outputted the number of the button pressed and the subjects' response latency.

The basic disadvantage of utilizing stimuli on slides is that only eighty slides can be presented sequentially, after which the experiment has to be stopped to change the magazine. The control program, EXPON 5, allows two alternative solutions to this problem.

One system, utilizing alternate stimuli and non-stimuli slides, has already been described. A steering tape is read in by the teletype, and the magazine either remains on a non-stimulus slide, or changes to a stimulus slide, depending on whether the tape contains a zero or a 1. The other option utilizes a series of magazines containing the

stimulus and non-stimulus slides in the desired order. After 80 slides have been shown, the computer automatically changes to the other projector. Thus, by changing the magazines in one projector whilst the second one is in operation, any number of slides can be projected without interrupting the experiment.

The control program

Only a brief outline will be given of the operation of the control program, since a detailed description of the logic would involve considerable space.

The program utilizes a piece of internal hardware known as a DB8E 12 bit i/o buffer, which is connected to the interrupt system of the PDP-8E. Also employed is the KD8-EP programmable real time clock.

At the beginning of an experimental session, when the program has been started, the computer reads in a control tape. If the character * occurs on the tape, or is typed in the teletype, the program reads all subsequent characters as parameters which define the conditions and the program facilities that will be utilized during the coming experimental session. A ! on the tape indicates that there are no further conditions to be read in, and starts the experiment. One of the great advantages of the program is that all the experimental conditions can be changed on-line if required. All that is necessary to do, is to include on the steering tape an asterisk followed by the new condition parameters and a !. Options allow all the experimental counters to be either reset to zero or left at their previous values as required.

Example of program operation

The following description is an outline of the sequence of operations which occur at the beginning of a typical experiment.

a. When the program is started, at address 200, the interrupt system is turned on, the DB 8E buffer enabled and the program stays in a loop, waiting for an interrupt.

b. An * is typed into the keyboard. This raises an interrupt and causes a branch via location zero to the service routine SERV. This tests to see if the interrupt originated from the clock, buffer, keyboard or teletype, and in this case branches to KB, the keyboard interrupt handling routine. After clearing the keyboard flag, the routine checks to see if the character typed was a control character (* or !). If neither, it assumes that the keypress was a response, and branches to EVTIM, the timing routine. Receipt of the * causes a branch to the initialization routine, SETUP.

c. SETUP reads the various experimental options and parameters from the steering tape. When entered at the beginning of the experiment, the following parameters are read in:

1. Projector mode 1 = one projector 0 = 2 projectors (see earlier description).
2. Shutter delay. This is set to the opening time of the shutter in 100th's of a second.
3. Cuing delay. Time in 100th's of a second between presentation of cuing information to the subject and onset of the cued slide.

4. Delay after summary information in hundreths of a second. This is used when the experiment is in continuous mode. (The summary information is presented at the end of a block of 100 responses, a rest period of 30 seconds occurs and then the next slide is presented). This parameter is the length of time (in 100th's of a second) between the presentation of the KR and the first slide in the next block.

5. Dummy (unused).

6. Initial value of subjects cumulative score. The above parameters are read in at the beginning of the experiment. The following can be changed on-line, but must be preceded by an * on the steering tape.

7. Interstimulus interval (in 100th's of a second). When the experiment is in a paced mode, this is the time between stimuli presentations. In self-paced mode, this is the time allowed to the subject to make his decision, up to a maximum of 20 seconds. If he does not respond in that time the next slide is presented and a zero channel on time is punched out.

8. Pace option 1 for self, zero for external pacing.

9. Modifier for feedback. This quantity allows the number of trials between feedback trials to be varied on a trial by trial basis. For example if this quantity is 1, the number of trials between feedback is successively 1, 2, 3 etc. The quantity only applies to feedback and can be negative.

10. Number of trials between feedback. If zero, no feedback, if 1 alternate trials, etc. This quantity can be changed by the modifier 9.

11. Same as 10.

12. Modifier for cuing (as 9).

13. Number of trials between cued trials.

14. As 13.

15. Number of trials before summary feedback of score.

16. New value of 15 after summary feedback has been given.

17. Summary feedback option 1 = summary, 0 = no summary.

d. After reading in these parameters, the program goes to EXIT if SETUP was entered at the beginning of the session, or CLK if entered on-line. After clearing the various flags, the last control character ! is read in, the program presents the first slide and awaits the subjects response.

APPENDIX D

```
C
C MAIN PROGRAM FOR ANALYSIS OF D.A DATA
C USES SR MEANY,AVDA
C
      DOUBLE INTEGER FILIN(2)
      DIMENSION NCD(252),NFA(252),NOM(252),NCR(252),CDL(252)
     1,CRL(252),FAL(252),OML(252)
     1,PCD(126),PCR(126),POM(126),PFA(126)
      COMMON IDUM(8),WORK1(386)
      EQUIVALENCE (PCD(1),CDL(127)),(PCR(1),CRL(127)),(PFA(1),
     1FAL(127)),(POM(1),OML(127))
      WRITE(4,600)
600   FORMAT(' TYPE INPUT FILE NAME')
      READ(3,100) FILIN(1),FILIN(2)
100   FORMAT(A5,A4)
      CALL SEEK(2,FILIN)
C
C READ INPUT DATA
C
      DO 2 I=1,252
2     READ(2,) NCD(I),NCR(I),NFA(I),NOM(I),CDL(I),CRL
     1(I),FAL(I),OML(I)
C
C COLLAPSE 6 TO 3 TIME INTERVALS
C
      I=1
      DO 3 J1=1,252,12
      J2=J1+5
      DO 3 J=J1,J2
      NA=NB=NC=ND=2
      NCD(I)=NCD(J)+NCD(J+6)
      NCR(I)=NCR(J)+NCR(J+6)
      NFA(I)=NFA(J)+NFA(J+6)
      NOM(I)=NOM(J)+NOM(J+6)
      IF(CDL(J).LE.0.5.OR.CDL(J+6).LE.0.5) NA=1
      IF(CRL(J).LE.0.5.OR.CDL(J+6).LE.0.5) NB=1
      IF(FAL(J).LE.0.5.OR.FAL(J+6).LE.0.5) NC=1
      IF(OML(J).LE.0.5.OR.OML(J+6).LE.0.5) ND=1
      CDL(I)=(CDL(J)+CDL(J+6))/NA
      CRL(I)=(CRL(J)+CRL(J+6))/NB
      FAL(I)=(FAL(J)+FAL(J+6))/NC
      OML(I)=(OML(J)+OML(J+6))/ND
3     I=I+1
C
C A PRIORI PROBS
C
      DO 39 J=1,126
39    PCD(J)=FLOAT(NCD(J)+NOM(J))/FLOAT(NCD(J)+NOM(J)+NCR(J)+NFA(J
      WRITE(11,41)
41    FORMAT(//30X,'A PRIORI SIGNAL PROB'/)
      CALL COPY (PCD)
      CALL MEANY
      CALL AVDA
      WRITE(11,40)
40    FORMAT(//30X,'ASIN TRAN SIGNAL PROB'/)
      CALL TRANS(2)
      CALL MEANY
```

```
          CALL AVDA
C
C CALC. PROBS. USE ESTIMATE IF FA=0 OR CD PROB =1
C
          DO 4 J=1,126
          IF(NFA(J).GT.0) GO TO 500
          PFA(J)=1.-((0.5)**(1./FLOAT(NCR(J))))
          GO TO 501
500       PFA(J)=FLOAT(NFA(J))/FLOAT((NFA(J)+NCR(J)))
501       IF(NOM(J).GT.0) GO TO 502
          POM(J)=1.-((0.5)**(1./FLOAT(NCD(J))))
          PCD(J)=1.0-POM(J)
          GO TO 503
502       PCD(J)=FLOAT(NCD(J))/FLOAT((NCD(J)+NOM(J)))
          POM(J)=1.0-PCD(J)
503       PCR(J)=1.0-PFA(J)
4         CONTINUE
C
C OUTPUT ZCD FOR PLOT
C
          DO 207 J=1,126
          ACD=PCD(J)
          CALL NDI(ACD,ZCD,DFA,IE)
207       WORK1(J)=ZCD
          WRITE(11,208)
208       FORMAT(//30X,'ZCD'/)
          CALL MEANY
          CALL AVDA
C
C BEGIN MAIN ANALYSIS
C
C PROBS
C
          CALL COPY(PCD)
          WRITE(11,9)
9         FORMAT(//30X,'CORRECT DETECTION PROBABILITY'/)
          CALL MEANY
          CALL AVDA
          CALL TRANS(2)
          WRITE(11,14)
14        FORMAT(//30X,'2ASIN TRANS OF CD PROB'/)
          CALL MEANY
          CALL AVDA
          CALL COPY(PFA)
          WRITE(11,11)
11        FORMAT(//30X,'FALSE ALARM PROBABILITY'/)
          CALL MEANY
          CALL AVDA
          WRITE(11,15)
15        FORMAT(//30X,'2ASIN TRANS OF FA PROB'/)
          CALL TRANS(2)
          CALL MEANY
          CALL AVDA
C
C INSPECTION PERF. INDICES
C
          DO 300 J=1,126
```

```
          NTOT=NFA(J)+NOM(J)+NCD(J)+NCR(J)
300       WORK1(J)=(FLOAT(NCD(J)+NCR(J))/NTOT)*100.0
          WRITE(11,301)
301       FORMAT(//30X,'EFFICIENCY INDEX A1')
          CALL MEANY
          CALL AVDA
          DO 302 J=1,126
302       WORK1(J)=(FLOAT(NCR(J))/(NFA(J)+NCR(J)))*100.0
          WRITE(11,303)
303       FORMAT(//30X,'EFFICIENCY INDEX A2'/)
          CALL MEANY
          CALL AVDA
          DO 304 J=1,126
304       WORK1(J)=(FLOAT(NCD(J))/(NCD(J)+NOM(J)))*100.0
          WRITE(11,305)
305       FORMAT(//30X,'EFFICIENCY INDEX A3'/)
          CALL MEANY
          CALL AVDA
          DO 307 J=1,126
          NIL=NCD(J)+NOM(J)-NFA(J)
          IF(NIL.EQ.0) GO TO 306
          WORK1(J)=(FLOAT(NCD(J)-NFA(J))/NIL)*100.0
          GO TO 307
306       WORK1(J)=0.0
307       CONTINUE
          WRITE(11,308)
308       FORMAT(//30X,'EFFICIENCY INDEX A4')
          CALL MEANY
          CALL AVDA
C
C SDT PARAMETERS
C
          ISW=1
50        DO 10 J=1,126
          AFA=PFA(J)
          ACD=PCD(J)
          CALL NDI(AFA,ZFA,DFA,IE)
          CALL NDI(ACD,ZCD,DCD,IE)
          IF(ISW.EQ.0) GO TO 6
          WORK1(J)=ZFA-ZCD
          GO TO 10
6         WORK1(J)=DCD/DFA
10        CONTINUE
          IF(ISW.EQ.0) GO TO 8
          WRITE(11,17)
17        FORMAT(//30X,'D',' SENS.  INDEX'/)
          CALL MEANY
          CALL AVDA
          ISW=0
          GO TO 50
8         WRITE(11,18)
18        FORMAT(//30X,'BETA'/)
          CALL MEANY
          CALL AVDA
          WRITE(11,19)
19        FORMAT(//30X,'LOG BETA'/)
          CALL TRANS(1)
```

```
        CALL MEANY
        CALL AVDA
C
C NONPARAMETRIC SENS. MEAS.
C
C POLLACK-NORMAN INDEX A'
C
        DO 20 J=1,126
20      WORK1(J)=0.5+((PCD(J)-PFA(J))*(1.0+PCD(J)-PFA(J)))
        1/(4.0*PCD(J)*(1.-PFA(J)))
        WRITE(11,21)
21      FORMAT(//30X,'POLLACK-NORMAN SENS. INDEX A''')
        CALL MEANY
        CALL AVDA
C POLLACK INDEX
        DO 22 J=1,126
22      WORK1(J)=1.0/(3.0-2.0*WORK1(J))
        WRITE(11,23)
23      FORMAT(//30X,'POLLACK-HSIEH SENS. INDEX P(I)')
        CALL MEANY
        CALL AVDA
C
C LATENCY SENS MEASURE (NAVON 1975)
C
        DO 29 J=1,126
29      WORK1(J)=((OML(J)+FAL(J))/2.)-((CDL(J)+CRL(J))/2.)
        WRITE(11,30)
30      FORMAT(//30X,'LATENCY SENS. INDEX')
        CALL MEANY
        CALL AVDA
C
C BIAS MEASURES
C
C
C HODOS-GRIER INDEX BHG
C
        DO 24 J=1,126
24      WORK1(J)=((PCD(J)*(1.-PCD(J)))-(PFA(J)*(1.-PFA(J))))
        1/((PCD(J)*(1.-PCD(J)))+(PFA(J)*(1.-PFA(J))))
        WRITE(11,25)
25      FORMAT(//30X,'HODOS-GRIER BIAS INDEX B''')
        CALL MEANY
        CALL AVDA
C
C LR VERSION OF B'
C
        DO 26 J=1,126
        AD=WORK1(J)
        PFT=PFA(J)
        AC=3.0-4.0*(PFT+AD*(1.0-PFT))
        WORK1(J)=2.0*(1.-AD)-(0.5*(5.0-6.0*PFT-14.0*AD+16.0*AD*PF
        1+8.0*AD*AD*(1.0-PFT)))/SQRT(PFT*(1.-PFT)+((AC*AC)/4.0))
26      CONTINUE
        WRITE(11,35)
35      FORMAT(//30X,'L R OF B'' BIAS INDEX')
        CALL MEANY
        CALL AVDA
```

```
C
C ZFA AS BIAS INDEX
C
          DO 27 J=1,126
          AFA=PFA(J)
          CALL NDI(AFA,ZFA,DFA,IE)
27        WORK1(J)=ZFA
          WRITE(11,28)
28        FORMAT(//30X,'ZFA AS BIAS INDEX'/)
          CALL MEANY
          CALL AVDA
C
C LATENCY DATA
C
          CALL COPY(CDL)
          WRITE(11,45)
45        FORMAT(//30X,'CORRECT DETECTION LATENCY'/)
          CALL MEANY
          CALL AVDA
          CALL TRANS(1)
          WRITE(11,60)
60        FORMAT(//30X,'LOG C.D. LATENCY'/)
          CALL MEANY
          CALL AVDA
          CALL COPY(CRL)
          WRITE(11,31)
31        FORMAT(//30X,'CORRECT REJECTION LATENCY'/)
          CALL MEANY
          CALL AVDA
          CALL TRANS(1)
          WRITE(11,61)
61        FORMAT(//30X,'LOG C.R. LATENCY'/)
          CALL MEANY
          CALL AVDA
          CALL COPY(OML)
          WRITE(11,32)
32        FORMAT(//30X,'OMISSION LATENCY'/)
          CALL MEANY
          CALL AVDA
          CALL TRANS(1)
          WRITE(11,62)
62        FORMAT(30X,'LOG OMM. LATENCY'/)
          CALL MEANY
          CALL AVDA
          CALL COPY(FAL)
          WRITE(11,33)
33        FORMAT(//30X,'FALSE ALARM LATENCY'/)
          CALL MEANY
          CALL AVDA
          CALL TRANS(1)
          WRITE(11,63)
63        FORMAT(//30X,'LOG F.A. LATENCY'/)
          CALL MEANY
          CALL AVDA
          DO 400 J=1,126
          ND=4
          IF(CDL(J).LE.0.5) ND=ND-1
```

```
            IF(CRL(J).LE.0.5) ND=ND-1
            IF(OML(J).LE.0.5) ND=ND-1
            IF(FAL(J).LE.0.5) ND=ND-1
400         WORK1(J)=(CDL(J)+CRL(J)+OML(J)+FAL(J))/ND
            WRITE(11,64)
64          FORMAT(//30X,'TOTAL RESPONSE LATENCY'/)
            CALL MEANY
            CALL AVDA
            CALL TRANS(1)
            WRITE(11,65)
65          FORMAT(//30X,'LOG TOTAL RESPONSE LATENCY'/)
            CALL MEANY
            CALL AVDA
            CALL CLOSE(2)
            STOP
            END
```

```
      SUBROUTINE SEMAIN(IDATA)
      INTEGER R(2,10),N(2),RR(2,10),IDATA(12)
      REAL P(2,10),AM(11,11),Z(9),G(2),ZZ(9)
      REAL FF(2,9),FG(2,9),DDZ(9),BR(2,9),D(2,10)
      COMMON IFLAG,AP,BP,RATIO,AGS,CHI,ND,ZP(9),BETAS(9),DUMM(18)
C
C
C THIS PROGRAM PERFORMS THE 'NORMAL' ANALYSIS DESCRIBED
C IN, 'SOME ASPECTS OF R.O.C. CURVE-FITTING : NORMAL AND
C LOGISTIC MODELS' BY GREY AND MORGAN, J. MATH. PSYCHOL.,
C CA., 1972. THE PROGRAM IS WRITTEN NAIVELY AND HAS NO PRETENSIONS
C TO BEING OPTIMAL ! BUT IT DOES THE JOB REQUIRED OF IT.
C VERSIONS DO EXIST INCORPORATING LOGISTIC RATHER THAN NORMAL
C DISTRIBUTIONS AND THESE WILL BE QUICKER (SEE ABOVE PAPER); BUT
C THEY ARE NOT AS SOPHISTICATED AS THIS PROGRAM. MODIFICATION
C TO GIVE LOGISTIC AS AN OPTION IS TRIVIAL; ONE MUST
C CHANGE THE FUNCTION, F, AND ALSO THE DEFINITION OF
C THE MATRIC FG (AGAIN SEE THE PAPER).
C
C DATA MUST BE READ IN AS FOLLOWS :
C  'N' (THE NUMBER OF CATEGORIES OF RESPONSE) AND THEN THE
C RESPONSE MATRIX, .
C FOR EXAMPLE, FOR THE RESPONSE MATRIX,
C
C          NO                YES
C  N        25        17      5
C  SN        2        18     30
C
C THE INPUT DATA IS OF THE FORM,
C
C       3
C      25   17    5
C       2   18   30
C
C
C FURTHER RESPONSE MATRICES OF THE SAME SIZE MAY THEN FOLLOW
C ON BELOW.  TO CHANGE THE SIZE OF THE RESPONSE MATRIX
C THE LAST MATRIX OF A GIVEN SIZE MUST BE FOLLOWED BY
C A SINGLE '/' IN THE PLACE OF A NUMBER, WHICH IS THEN
C FOLLOWED BY THE NEW NUMBER OF CATEGORIES AND THE NEW
C RESPONSE MATRIX.  DATA IS INPUT IN 'FREE FORMAT', IE.,
C IT SUFFICES TO SEPARATE NUMBERS BY A TAB OR TWO OR
C MORE SPACES [ IT IS HERE THAT THE PROGRAM DEPARTS FROM
C THE USASI SPECIFICATIONS].
C IMPORTANT : THE NUMBER OF CATEGORIES MUST BE
C AT LEAST THREE.  OCCASIONALLY, FOR OUTRAGEOUSLY POOR
C DATA THE PROGRAM MAY PROTEST.
      ND=ND-1
      CAP=0
      L=L+1
      WRITE(11,500) L
500   FORMAT(//1X,'BLOCK NO.',I3/)
C
C  UP TO 100 RESPONSE MATRICES OF A GIVEN SIZE ARE PERMITTED. ALL RUN
C  WITH UNSEEMLY EXHAUSTION OF INPUT
C
C
```

```
          IF (CAP.EQ.1.) GOTO 104
          JA=ND+1
          K=0
          DO 970 I=1,2
          DO 970 J=1,JA
          K=K+1
970       R(I,J)=IDATA(K)
99        FORMAT(12(I3,1X))
104      WRITE(8,73)
73       FORMAT(11H0INPUT DATA)
          WRITE(11,76)((R(I,J),J=1,JA),I=1,2)
76        FORMAT(1X,'N',2X,6I4/1X,'SN',1X,6I4/)
         NDS=0
         MO=ND
          MO1=MO+1
         DO 90  J=1,MO1
         IF (R(1,J)) 91,91,90
91        IF(R(2,J)) 90,92,90
92       ND=ND-1
          JA=JA-1
         NDS=NDS+1
         R(1,J) = 7000
          NNDS=ND+NDS
C
C
C THIS COLLAPSES OVER UNUSED CATEGORIES
C
C
90       CONTINUE
         DO 93 J=1,NNDS
95       IF (R(1,J)-7000) 93, 94,94
94       DO 96 K=J,NNDS
         R(2,K) =R(2,K+1)
96       R(1,K)=R(1,K+1)
         R(2,ND+NDS+1)=0
         R(1,ND+NDS+1)=0
         GO TO 95
93       CONTINUE
C
C
C FIRST OF ONE TRIES THE CONSISTENT INITIAL START
C (PROCEDURE 0)
C
C
          DO 3 I=1,2
          IS=0
          DO 7 J=1,JA
7        IS=IS+R(I,J)
          N(I)=IS
          G(I)=FLOAT(IS)
          DO 8 J=1,JA
8        P(I,J)=FLOAT(R(I,J))/G(I)
3        CONTINUE
          NA=0
          SUM1=0
          SUM2=0
          SUM3=0
```

```
          SUM4=0
          SUM5=0
          SUM6=0
          SUM7=0
          SUM8=0
          DO 110 J=1,JA
          SUM1=SUM1+FLOAT(R(1,J))
110       SUM2=SUM2+FLOAT(R(2,J))
          DO 111 J=1,ND
          SUM3=SUM3+FLOAT(R(1,J))
          SUM4=SUM4+FLOAT(R(2,J))
          Z(J)=GG(SUM3/SUM1)
111       ZZ(J)=GG(SUM4/SUM2)
          DO 112 J=1,ND
          SUM5=SUM5+Z(J)
          SUM6=SUM6+ZZ(J)
          SUM7=SUM7+Z(J)*ZZ(J)
112       SUM8=SUM8+Z(J)**2
          AXE=FLOAT(ND)
          CAD=(SUM7-SUM5*SUM6/AXE)/(SUM8*AXE-SUM5**2)
          A=CAD*SUM5-SUM6/AXE
          B=AXE*CAD
          GOTO 62
178       IX=-1
179       DO 77 ID=1,2
          DO 77 IA=1,6
C
C
C THESE LOOPS DETERMINE THE INITIAL STARTING POINTS
C WHICH ARE TRIED SUCCESSIVELY, IF THE CONSISTENT START
C FAILS, IN SEARCH OF A CONVERGENT PROCEDURE.
C
C
          A=FLOAT(IA)/2.
          B=1.
          NA=NA+1
          DO 4 J=1,ND
4        Z(J)=FLOAT(J-1+IX)/(2.*FLOAT(ID))
C
C
C ONE NOW SETS UP THE INITIAL A MATRIX
C
C
62        DO 13 J=1,ND
           FF(2,J)=F(B*Z(J)-A)
           FG(2,J)=0.39894*EXP(-0.5*(B*Z(J)-A)**2)
           FF(1,J)=F(Z(J))
13        FG(1,J)=0.39894*EXP(-0.5*Z(J)**2)
           DO 15 I=1,2
           D(I,1)=FF(I,1)
           DO 14 J=2,ND
14        D(I,J)=FF(I,J)-FF(I,J-1)
           D(I,JA)=1.-FF(I,ND)
           DO 15 J=1,JA
           IF (D(I,J).GT.1E-5) GOTO 15
           D(I,J)=1E-5
15        CONTINUE
```

```
      DO 30 I=1,2
      DO 30 J=1,ND
30    BR(I,J)=P(I,J)/D(I,J)-P(I,J+1)/D(I,J+1)
      DDA=0.
      DDB=0.
      DO 32 J=1,ND
      DDA=DDA+FG(2,J)*BR(2,J)
      DDB=DDB+FG(2,J)*Z(J)*BR(2,J)
32    DDZ(J)=G(2)*FG(2,J)*B*BR(2,J)+G(1)*FG(1,J)*BR(1,J)
      DDA=-G(2)*DDA
      DDB=G(2)*DDB
      BR(2,1)=FG(2,1)/D(2,1)-(FG(2,2)-FG(2,1))/D(2,2)
      NDM1=ND-1
      DO 33 J=2,NDM1
33    BR(2,J)=(FG(2,J)-FG(2,J-1))/D(2,J)-(FG(2,J+1)-FG(2,J))/D(2,J+
      BR(2,ND)=(FG(2,ND)-FG(2,ND-1))/D(2,ND)+FG(2,ND)/D(2,JA)
      NDP2=ND+2
      DO 34 I=1,NDP2
      DO 34 J=1,NDP2
34    AM(I,J)=0.
      DO 35 J=1,ND
      AM(1,1)=AM(1,1)+FG(2,J)*BR(2,J)
      AM(1,2)=AM(1,2)+FG(2,J)*Z(J)*BR(2,J)
      AM(1,J+2)=G(2)*FG(2,J)*B*BR(2,J)
35    AM(J+2,1)=AM(1,J+2)
      AM(1,1)=-G(2)*AM(1,1)
      AM(1,2)=G(2)*AM(1,2)
      AM(2,1)=AM(1,2)
      BR(2,1)=FG(2,1)*Z(1)/D(2,1)
      DO 36 J=2,ND
      BR(2,J)=(FG(2,J)*Z(J)-FG(2,J-1)*Z(J-1))/D(2,J)
      AM(2,2)=AM(2,2)+FG(2,J-1)*Z(J-1)*(BR(2,J-1)-BR(2,J))
      AM(2,J+1)=-G(2)*FG(2,J-1)*B*(BR(2,J-1)-BR(2,J))
36    AM(J+1,2)=AM(2,J+1)
      K=ND
      AM(2,2)=AM(2,2)+FG(2,K)*Z(K)*(BR(2,K)+FG(2,K)*Z(K)/D(2,K+1))
      AM(2,2)=-G(2)*AM(2,2)
      AM(2,K+2)=-G(2)*FG(2,K)*B*(BR(2,K)+FG(2,K)*Z(K)/D(2,K+1))
      AM(K+2,2)=AM(2,K+2)
      DO 37 J=1,ND
      K=J+2
      AM(K,K)=-G(2)*FG(2,J)*B*B*(FG(2,J)/D(2,J)+FG(2,J)/D(2,J+1))
37    AM(K,K)=AM(K,K)-G(1)*FG(1,J)*(FG(1,J)/D(1,J)+FG(1,J)/D(1,J+1))
      DO 38 J=2,ND
      AM(J+2,J+1)=G(2)*B*B*FG(2,J)*FG(2,J-1)/D(2,J)
      AM(J+2,J+1)=AM(J+2,J+1)+G(1)*FG(1,J)*FG(1,J-1)/D(1,J)
38    AM(J+1,J+2)=AM(J+2,J+1)
      CALL MB01A(AM,ND+2,11)
C                              N+1
C
C SUBROUTINE AT END OF PROGRAM TO INVERT MATRIX
C
C
      AP=A-AM(1,1)*DDA-AM(1,2)*DDB
      BP=B-AM(2,1)*DDA-AM(2,2)*DDB
      DO 50 J=1,ND
50    ZP(J)=Z(J)-AM(J+2,1)*DDA-AM(J+2,2)*DDB
```

```
          DO 51 K=1,ND
          AP=AP-AM(1,K+2)*DDZ(K)
          BP=BP-AM(2,K+2)*DDZ(K)
          DO 52 J=1,ND
52       ZP(J)=ZP(J)-AM(J+2,K+2)*DDZ(K)
51       CONTINUE
          IF (ABS(A-AP).GT.1E-3) GOTO 70
          IF (ABS(B-BP).GT.1E-3) GOTO 70
          DO 56 J=1,ND
         IF (ABS(Z(J)-ZP(J)).GT.1E-3) GOTO 70
56        CONTINUE
C
C
C OUTPUT FORMAT IS AS IN ABOVE PAPER BUT
C ALSO INCLUDES LIKELIHOOD RATIOS (BETA) AT
C ALL CUT-OFF POINTS.
C
C
58       FORMAT(43H0MAXIMUM LIKELIHOOD ESTIMATES OF PARAMETERS/)
59       FORMAT(5H0A = ,F8.3)
60       FORMAT(1X,4HB = ,F8.3/)
61        FORMAT(1X,1HZ,I1,3H = ,F8.3,4X,'BETA=',F8.3,4X,'LOG BETA',F8.3
          IF (BP.LT.0.) GOTO 77
          DO 67 J=2,ND
         IF (ZP(J).LT.ZP(J-1)) GOTO 77
67        CONTINUE
C
C
C ITERATION HAS CONVERGED TO A MEANINGFUL SOLUTION
C
C
68       FORMAT(13H0PROCEDURE NO,I3,11H SUCCESSFUL)
          WRITE (11,68) NA
          WRITE (11,58)
          WRITE (11,59) AP
          WRITE (11,60) BP
          RATIO=1.0/BP
          WRITE(11,600) RATIO
600       FORMAT(1X,'RATIO OF VARIANCES,SN TO N =',F8.3/)
          DO 65 J=1,ND
          BETA=(BP)*EXP((ZP(J)*ZP(J)-((ZP(J)*BP-AP)*(ZP(J)*BP-AP)))/2.)
          EBETA=ALOG(BETA)
          BETAS(J)=BETA
65       WRITE (11,61) J,ZP(J),BETA,EBETA
69       FORMAT(/1X,8HDDASH = ,F8.3/)
          AGS=AP/SQRT(BP)
          WRITE(11,69) AGS
          GO TO 777
776       WRITE(11,22)
22       FORMAT(40H0VARIANCE-COVARIANCE MATRIX OF ESTIMATES)
          NDP2=ND+2
         DO 230 I=1,NDP2
         WRITE(11,24)
         DO 230 J=1,NDP2
          CAM=-AM(I,J)
23       FORMAT(1X,7F8.3/)
          WRITE(11,23) CAM
```

```
230       CONTINUE
777       CONTINUE
24       FORMAT(1H )
         CHI=0
         SUM1=0
         SUM2=0
         DO 25 J=1,JA
         SUM1=SUM1+FLOAT(R(1,J))
25       SUM2=SUM2+FLOAT(R(2,J))
         DO 27 I=1,2
         IF (I-1) 28,28,29
29        DO 40 J=1,ND
40         ZP(J)=(ZP(J)-AP/BP)*BP
          SUM1=SUM2
28        A=F(ZP(1))
         B=(A*SUM1-FLOAT(R(I,1)))**2
         B=B/(A*SUM1)
         CHI=CHI+B
         DO 26 J=2,ND
         B=F(ZP(J))-F(ZP(J-1))
         C=(B*SUM1-FLOAT(R(I,J)))**2
         C=C/(B*SUM1)
26       CHI=CHI+C
         A=1.-F(ZP(ND))
         B=(A*SUM1-FLOAT(R(I,JA)))**2
         B=B/(A*SUM1)
27       CHI=CHI+B
         WRITE(11,41) CHI
41       FORMAT(39H0CHI-SQUARE GOODNESS OF FIT OF MODEL = ,F12.3)
          NDM2=ND-2
          ANDM2=NDM2
          CALL ZTFCHI(CHI,ANDM2,30000.0,2,PR)
         WRITE(11,42) NDM2,PR
42       FORMAT(6H DF = ,I3,5X,'P= ',F7.5)
          IFLAG=0
          FND=ND
          GOTO 78
70       IF (ABS(AP).GT.8.) GOTO 177
          IF (ABS(BP).GT.5.) GOTO 177
          DO 72 J=1,ND
         IF (ABS(ZP(J)).GT.8.) GOTO 177
72        CONTINUE
          A=AP
          B=BP
          DO 71 J=1,ND
71       Z(J)=ZP(J)
C
C
C THE ITERATION CONTINUES
C
C
          GOTO 62
177       IF(NA.EQ.0) GOTO 178
77       CONTINUE
79       FORMAT(24H0NO PROCEDURE SUCCESSFUL)
          WRITE (11,79)
          DO 899 KL=1,18
```

```
899       DUMM(KL)=0.0
          IFLAG=1
          GO TO 78
971       IF (CAP.EQ.1.) GOTO 78
          CAP=1.
          ND=ND+NDS
          DO 106 J=1,JA
106       RR(1,J)=R(1,J)
          DO 107 J=1,JA
107       R(1,J)=R(2,ND+2-J)
          DO 108 J=1,JA
108       R(2,J)=RR(1,ND+2-J)
C
C
C IF ALL ELSE FAILS THIS INVERTS THE DATA AND STARTS AGAIN
C (SEE PAPER)
C
C
          GO TO 104
78        CONTINUE
          RETURN
          END
```

SIGNAL DETECTION THEORY IN THE ANALYSIS AND OPTIMISATION OF INDUSTRIAL INSPECTION TASKS

A Thesis submitted for the degree of

Doctor of Philosophy

by David Edward Embrey BSc.

Department of Applied Psychology

University of Aston in Birmingham

September 1976

SUMMARY

The overall aim of this study was to investigate the applicability of Signal Detection Theory (SDT) to a number of problems in the area of industrial inspection, including training and selection. Two industrial studies comprising three experiments are presented, together with four laboratory experiments and a correlational study.

The first three chapters of the thesis comprised a comprehensive review of SDT and the literature of inspection.

Chapter 2 described an industrial case study, the inspection of photographs of nuclear particles, designed to test the applicability of SDT in an applied setting. The variables of auditory noise, defect complexity and time on task were also considered. SDT in its unequal variance form was found to fit the data. The next case study attempted to apply SDT to the inspection of photographic film. A two stage decision making model was proposed to describe performance in this task.

The first two laboratory studies investigated the effect on performance of within and between session changes in defect probability. It was found that the subject could adjust his criterion appropriately to between session changes if feedback was provided, and to within session changes if he received prior warning of the change.

The final laboratory studies were concerned with training the inspector's ability to modify his criterion, and the enhancement of his sensitivity. The first experiment replicated previous work in

perceptual training, and the second utilized a wide range of differing training techniques. It was found that certain combinations of conditions were significantly superior in achieving the training goals.

A correlational study was conducted utilizing the results of the previous two experiments and tests of various cognitive skills. Significant correlations were found between certain groups of test scores and performance on the task. These tests were proposed as potential selection techniques for inspectors.

"Quality is shapeless, formless, indescribable. To see shapes and forms is to intellectualize. Quality is independent of any such shapes and forms. The names, the shapes and forms we give Quality depend only partly on the Quality. They also depend partly on the a priori images we have accumulated in our memory. We constantly seek to find, in the Quality event, analogues to our previous experiences. If we didn't we'd be unable to act. We build up our language in terms of these analogues. We build up our whole culture in terms of these analogues."

Pirsig R.M., ZEN and the ART of MOTORCYCLE Maintenance.

ACKNOWLEDGEMENTS

I would like to express my gratitude to everybody who has helped and encouraged me during the long gestation period of this thesis. David Whitfield, my supervisor, was always ready to offer assistance, despite his heavy timetable. Mrs. Christine Maddison deserves my thanks for both deciphering my handwriting and for producing accurate typescripts at remarkable speed.

Acknowledgements are due to Professor W. T. Singleton for providing Departmental facilities and to the technical staff including, Les Bagnall, Paul Bernard and Colin Mason, who assembled the experimental equipment. The case studies were made possible by the generosity of Professor D. C. Colley of the Physics Department, University of Birmingham, and Mr. R. F. Salmon of Ilford Ltd. All the subjects deserve my thanks for the often arduous experimental sessions they endured.

Last, but not least, I must thank all my close friends, whose apparently unlimited willingness to work hard on my behalf made everything possible.

Contents

CHAPTER 1 INTRODUCTION

1.0 INTRODUCTION

In preparing this study, an attempt has been made to satisfy a need that has become increasingly apparent to ergonomics practitioners: the provision of data from experiments that have direct relevance to situations encountered in real world tasks. Although ergonomics is essentially an applied science, the orientation of much research has been towards purely laboratory based studies that provide little in the way of information which is readily applicable in an industrial context. Chapanis (1967) discusses this problem in detail. In an emerging science this bias is perhaps understandable. However, ergonomics and human factors have now been in existence for nearly thirty years and should by this time be producing such data on a large scale. In order to achieve this aim, it seems clear that research work needs to change its orientation from the formalized consideration of the effects of a few selected variables in highly specialized laboratory studies, to a more pragmatic approach yielding results which can be utilized more readily in real-life applications. This approach does not mean that compromises necessarily have to be made in standards of experimentation. It is a question of broadening the experimental focus rather than of producing work that has no theoretical significance.

An attempt has been made to make this study sensitive to these needs in a number of ways. Rather than considering a narrow research topic in some depth and then attempting to show that it also has practical relevance, the starting point of this study is a whole area of application within the industrial sector, that of quality control. Since this study is primarily concerned with human factors, the main consideration will be given to the inspection aspect of quality control.

The approach has been to show how data from a number of research areas contribute towards the goal of optimizing those aspects of a quality control system in which human beings play a predominant role. The utility of this approach is that it brings together data from a number of disparate areas in such a way that their effect on the inspection function can be clearly seen and some insights gained into the ways they may interact. The aim is to provide a usable body of information for the practitioner on the factors known to affect inspection. Of course, such a review also serves the more usual function of a literature survey in that it suggests potential areas of further research.

From the practitioner's standpoint, the data available in such a review will be most useful if it is classified in a manner clearly related to attributes common to a large proportion of real inspection tasks.

Another way in which this study attempts to maintain relevance to real world problems is in the experimental phase. The impetus for the laboratory based work is provided by field studies which clarify the nature of the important variables which need further consideration in a more controlled environment. The laboratory studies then seek to simulate the critical features of the real life task to provide directly relevant data. At the same time the experimental studies provide a vehicle in which a number of more theoretical ideas can be investigated.

In addition to the areas outlined above, the study attempts to assess the utility of considering the area of inspection from a particular theoretical standpoint - that of Signal Detection Theory, which serves as a unifying concept for a wide range of experimental work relevant to inspection.

The general aim then, can be summed up as an attempt to advance knowledge on a broad front, in a manner related to practical needs.

1.1 The importance of industrial inspection as an area of study

As the increasing application of technology to manufacturing industries reduces the importance of purely manual, motor skills, attention is being increasingly focussed on areas of industry in which higher level perceptual abilities of the operator, such as pattern recognition and decision making, are employed. Industrial inspection is such an area.

The reason why the human operator continues to be important in the quality control area is that these higher level functions cannot readily be performed by machines. Although automated pattern recognition is possible with sufficiently simple patterns, and computer aided decision making is being utilized in certain situations, (see Whitfield (1975) for a review of this area), the extremely high cost of such devices, and their inflexibility relative to a human operator, means that they are unlikely to find wide application in the industrial inspection area. The attribute of quality has a multidimensional nature which can encompass both simple parameters such as specification of size, as well as complex aesthetic judgements. The possibility of building machines to handle the whole range of quality judgements seems remote, although it seems feasible, and indeed desirable, to use automated techniques where very simple discriminations such as size and weight are required between acceptable and non-acceptable articles.

It is found in industry that even when simple judgements of this type are required, it is common to inspect manually. In fact Fox (1973)

states that 90% of all inspection in the United Kingdom is dependent on the unaided human operator. The reasons for this are usually straightforward cost-effectiveness considerations. The volume of the product to be inspected may not justify the design and manufacture of an automatic device to monitor quality. Alternatively, quality specifications may change frequently and the human operator is more readily 'reprogrammed' than his machine counterpart.

It seems clear that inspection is likely to remain a labour intensive area, and as such, ergonomics and human factors will continue to be able to make important contributions in optimizing the performance of the human operator.

Another important reason for studying inspection is that research in this area is not restricted in its usefulness purely to the industrial sector. Many other important tasks have characteristics very similar to those found in industrial inspection. For example the surveillance of radar screens in both military and civil applications can be seen to contain many elements common to inspection tasks. The operator is continuously monitoring an information source which provides signals which are complex in nature, infrequently occurring and unpredictable both in time and space. In the medical sphere, the examination of medical X-ray plates and cervical cancer smears are examples of almost classical inspection tasks which employ vast resources of trained manpower. A similar inspection problem occurs in high energy nuclear physics research, in which millions of bubble chamber photographs are scanned for patterns of tracks. This task will be considered in detail in chapter 4.

A final important reason for applying human factors and ergonomics principles to the optimization of inspection systems can be found from considerations of system reliability. The inspection phase of system development is a vital link between the manufacture of components and their incorporation in a total system. Many disastrous system failures can be traced to inadequate inspection procedures. Meister (1971) discusses the importance of inspection from a system reliability standpoint.

From the above discussion we can conclude that although inspection as an area of study has received relatively little attention compared with tasks in industry involving mainly motor skills, it is an area of considerable intrinsic interest and importance. With growing automation it is likely to grow in importance, by comparison with traditional manufacturing tasks. Additionally, inspection-like tasks are found in many spheres distinct from the industrial sector, and hence any research findings are likely to be widely applicable.

1.2 Characteristics of inspection tasks

One of the most obvious characteristics of inspection tasks is their diversity. Virtually any item which is manufactured is likely to be inspected at some stage of manufacture or assembly. Although the visual modality is the usual one employed in inspection, other sensory inputs are sometimes employed. Thomas (1962) describes how vacuum cleaners were inspected for mechanical faults by the tester actually listening to the sound the cleaner made in operation. Several categories of fault could be distinguished by this method. Needles are inspected for straightness by gently rolling them under the palms on a flat table. Frequently combinations of sensory modalities are employed, as when an

article is inspected for surface finish both by its appearance and by its tactile characteristics.

Perhaps the most common form of inspection is when the inspector sequentially examines a series of items to determine whether the characteristics of the items fall within the quality specification. Even in this case a more detailed task description rapidly leads to complications. For example the items may be on a conveyor belt, moving at a range of speeds or the task may be completely self paced. They may be large items, in which case some form of scanning may be required, or they may be small enough to be examined in a single fixation. The defects may be visible with the naked eye or may require enhancement by means such as magnification or lighting. They may be distinguishable from non-defects by an extremely large range of variables including shape, size, colour, texture, weight and any combination of these and other parameters. This is an additional reason why automated inspection is often impracticable.

Another common type of inspection is the examination of materials in a continuous form. Examples of this type of task are the examination of cloth, steel and glass strip and cine film, which will be discussed as a detailed study in chapter 5 of this thesis.

In chapter 3 the various characteristics of inspection tasks will be considered in a more systematic manner. For the moment it is useful simply to be aware of their diversity and to attempt to describe a common behavioural structure which applies to the majority of inspection tasks. This will be considered in the next section.

1.3 An informal model of the inspection task

As was implied in the last section, the types of inspection task that occur show such diversity that any specific task cannot be representative of the whole range of situations encountered. The majority of inspection tasks do, however, share certain common characteristics, and in this section an informal structure will be described which facilitates the consideration of the tasks from a psychological and ergonomics standpoint.

The model to be described is not predictive in nature, although it could provide the basis for a simulation approach to the predictive modelling similar to that described by Siegel and Wolf (1969) in the context of man-machine systems. For our present purposes, the model will serve to indicate the areas of knowledge relevant to inspection situations in general.

It is convenient to consider inspection as consisting of four broad phases (Figure 1).

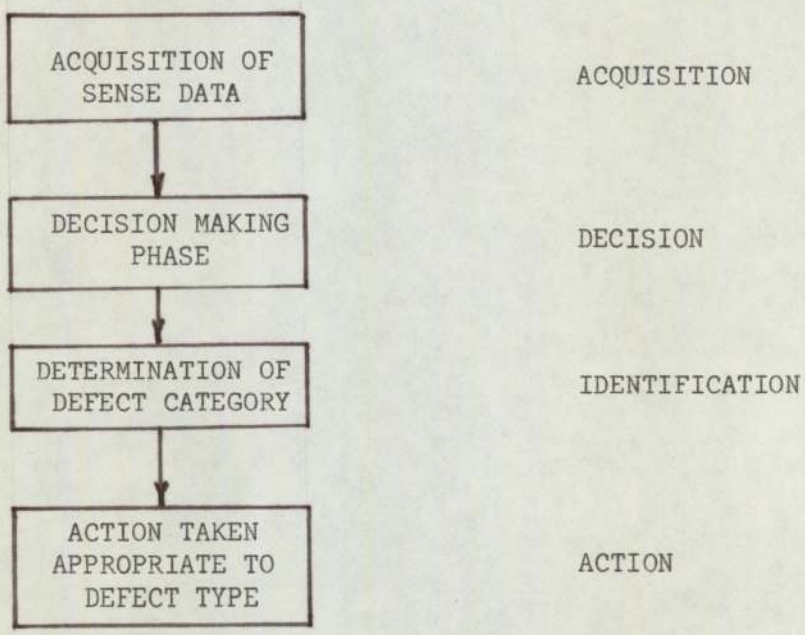


Figure 1. Phases of inspection.

a. Acquisition of sense data. In order that the inspector can decide on the presence or otherwise of a defect, he requires sensory evidence from the item being examined to provide inputs for the decision making phase.

b. Decision making. This aspect of the model refers to the process whereby the inspector assigns the item to the general categories of defect or non-defect, without further sub-categorization.

c. Identification. This is clearly also a decision making process which can be regarded as one stage higher in the decision hierarchy. It involves the classification of the defect into one of a number of sub-categories, if more than one exists

d. Action. Once the nature of the defect has been ascertained the inspector performs the action appropriate to that class of defect, e.g. rejects the item, returns it for reworking etc.

Each of the phases considered can be analysed in terms of the psychological and physiological factors which affect performance at each phase. Examples of these are set out in Figure 2.

1.3.1 Acquisition of the data

The acquisition phase has been considered as being affected by situational, physiological and psychological factors, although many of these could be included under more than one heading.

Situational factors are those which are external to the inspector and include environmental variables as well as the specific attributes of the task itself. The physiological factors affecting target acquisition are primarily visual, reflecting the preponderance of this modality in inspection tasks. Environmental conditions are included under

Figure 2. Examples of psychological and physiological factors affecting various phases of inspection.

1. ACQUISITION FACTORS

(a) Situational

I Paced or unpaced presentation, rate of pacing.
II Enhancement of discriminability of defect, e.g. X-rays, ultrasonics, magnification, lighting.
III Inherent discriminability of defect.

(b) Physiological

I Visual acuity, static or dynamic.
II Visual skills in general, e.g. colour vision.
III Environmental conditions affecting inspection performance e.g. heat, noise, lighting levels.
IV Visual fatigue

(c) Psychological

I Perceptual 'set', i.e. ability to recognize cues characteristic of defects as opposed to other configurations occurring in both defective and perfect product.
II Visual search strategies.
III Vigilance and attentional variables.
IV Organismic variables such as pattern recognition skills, field dependence and distractability.

2. DECISION MAKING FACTORS

I Expected incidence of defects.
II Costs associated with missing defects and rejecting good products.
III Social factors

3. IDENTIFICATION FACTORS

I Training and experience.
II Provision of reference standards.
III Expectancies concerning type of defect likely to occur.
IV Number of categories of defect.

4. ACTION FACTORS

I Existence of clearly defined actions to be taken for various types of defects.
II Consequences of action.
III Social factors.

physiological factors even though they may affect functioning via decrements in psychological skills. The psychological factors are intended to represent the intra-subject variables which influence target acquisition. The question of the perceptual skills which influence the probability of target detection will be discussed in some detail in this study. In including perceptual skills within the phase of acquisition we are referring to the inspector's sensitivity for cues which identify the sample being examined as belonging to the categories defect or non-defect. These cues are often indirect and the inspector may have to infer the existence of a hidden defect from the external evidence available. It is important to note that we are referring to a concept of sensitivity which is independent of the inspector's tendency to respond 'defect' or 'non-defect' as a result of prior knowledge of the probability of a particular item being defective, or because he will be heavily penalized if a good item is incorrectly rejected. These latter factors are considered to be decision making variables.

The concept of intrinsic sensitivity refers to the ability of the operator to effect categorization utilizing the evidence available in the sample. As has been noted earlier, we regard this facility as being uncontaminated by a bias to respond defect or non-defect due to data other than that available from the sample. Whether it is possible to separate sensitivity from 'response bias' due to other factors is a question we shall pursue at length in chapter 2.

It is clear that the inspector's ability to distinguish good from bad products will depend partly on his knowledge of the cues which indicate defectiveness and partly on the amount of data that he can acquire from the sample. A third, slightly more controversial factor, concerns his intrinsic ability, independent of training, to isolate a particular

configuration of cues embedded in a confusing background.

The utilization of cues present in the sample is clearly a function of training and also the provision of reference aids defining defective items. This topic will also be explored more fully later. The inspector's ability to acquire information from the sample is strongly influenced by physiological factors such as visual acuity discussed earlier. The area of visual search has been included in the psychological factors although it has a strong physiological element. Visual search skills consider the ability of the inspector to search an area exhaustively, efficiently and rapidly. Many of the studies of visual search which will be considered later have been concerned with a subject's ability to economically scan large areas for a target. To this extent the prime interest in search strategies is in the area of the inspection of sheet materials for defects. Clearly there is a high degree of interaction between some of the factors being considered. For example if the situational factor of pacing is very high the question of the time taken by the inspector to scan an item becomes highly important.

The relevance of research on vigilance tasks to industrial inspection is a function of the degree to which the task under consideration approaches that of the classical vigilance decrement situation. If an inspection task is conducted for prolonged periods in an unstimulating environment, with low probability, irregularly occurring defects, then vigilance effects might be expected. As will become apparent in the literature survey, there is some controversy as to the applicability of much vigilance research to industrial situations.

The possibility that an inspector's detection skill is related to innate abilities such as field dependence and pattern recognition is an

intriguing one that has not yet been explored by workers in the inspection field. The whole area of individual differences is one which has been neglected, particularly in relation to the selection of inspectors. It is hoped that this study will provide an impetus for research in this area.

1.3.2 Decision making in inspection

Many of the issues relevant to the decision making aspects of inspection skills cannot be discussed at this point because they are an integral part of the Signal Detection Theory orientation to the area which will be considered in detail later. At this stage it is sufficient to point out that one would expect an inspector's decision about whether or not a sample is defective to be influenced not only by the evidence available from the sample but by the expected incidence of defects in the whole series of samples. Similarly it is reasonable to expect the consequences of a particular decision to influence the inspector's judgement. For example, if the item being inspected is a critical part for a spacecraft, then the inspector will be far more likely to reject it if there is even a suspicion that it is defective, than if it were a non-critical item. These commonsense notions will be related to a theoretical structure in a later section.

Social factors can be seen to affect decision making in a situation where a worker is able to exert social pressure on an inspector if too large a proportion of his work is being rejected. Such pressures can be quite subtle and overt threats are not necessary to influence the judgement of the inspector.

1.3.3 Identification factors

Identification is distinguishable from acquisition in that the inspector is attempting to decide between different categories of defect after having decided that a particular signal is in fact a defect. Most of the factors operative at the decision making phase are also important here, although it is essentially a multiple categorization problem rather than a binary decision making process, at least where there are several types of defect. As before, training and experience will be important in allowing accurate differentation between types of defect. Of particular importance is the provision of reference standards in order to provide examples of the distinguishing characteristics between defects. As in the decision making phase, expectancies concerning the type of defect likely to occur will influence the categorization process.

1.3.4 Action factors

The absence of clearly defined actions to be taken in the event of the various types of defect occurring can lead to a considerable degradation in the efficiency of the inspection system. Some defects may indicate the presence of certain manufacturing malfunctions and hence necessitate a rapid feedback to the production section of the factory. Other defects may require reworking rather than rejection. Certain types of defective items may be acceptable to some customers who may be selling at discount to a less discriminating market. The consequences of certain actions may affect the inspector's behaviour in the same way as during the decision making phase of inspection.

Social factors need to be considered again in this context. The prevailing employment situation in a factory might, for example, influence

an inspector in deciding whether a particular item was reworkable or should be scrapped. The action phase of inspection has received little attention in the literature, and should perhaps be considered more explicitly in the analysis of inspection systems.

1.4 Conclusion

The consideration of an informal model of the inspection task has provided an overview of many of the topics which will be considered in more detail in subsequent chapters. If quantitative estimates were available for the effects of the variables considered on inspection performance, then it would be possible to use a model of this type for predictive purposes. As will become apparent during the subsequent review chapters, however, we have a considerable way to go before we can realistically assess the combined effects of some of the variables considered in the model, on inspection performance in general. Nevertheless the predictive modelling of the performance of human operators in an inspection system should be regarded as a desirable long term objective.

The validity of industrial inspection as an area of study for the behavioural sciences was also established in this chapter. It was pointed out that inspection requires a wide range of cognitive, decision making and pattern recognition skills which cannot readily be automated, and hence quality control is likely to remain a labour intensive area of industry. Data from inspection studies can be readily generalized to a number of other important areas such as radar screen surveillance and any situation where prolonged monitoring or repeated perceptual decision making takes place. Finally the quality of an inspection system has a considerable effect on the overall

reliability of a system, which is becoming increasingly important as larger and more complex systems are produced.

CHAPTER 2 SIGNAL DETECTION THEORY AND ITS APPLICATION TO INSPECTION

2.0 INTRODUCTION

In the development which follows, Signal Detection Theory, (hereinafter referred to as SDT) will first be briefly considered from an historical standpoint, emphasizing the psychophysical ideas from which it emerged. Next the important concept of response bias will be considered, with particular reference to inspection and other industrial applications.

The theory will then be developed in its simplest form, the equal variance model, together with the experimental evidence for this position, and then the more general form will be considered. Methods of analysing situations in which the simplified assumptions do not apply will be discussed in detail, particularly from the point of view of applying SDT to real as opposed to laboratory tasks.

The existing studies in which SDT has been applied to inspection tasks will be reviewed in detail and general rules for applying SDT to this area will be set out. Finally potentially rewarding areas of applicability of the theory to inspection will be summarized and directions for further research proposed.

In line with the applied nature of this study, no attempt will be made to produce a mathematically sophisticated analysis of SDT, and therefore detailed mathematical justifications of various points will be referred to standard texts.

2.0.1 Historical background

SDT was developed in its modern form by two groups of researchers working independently at Michigan and Harvard Universities on the

problems of detecting signals in noisy channels, particularly in the context of radar (Swets, 1963). In fact its origins can be traced back as far as Fechner (1801-1887). Fechner observed "the great variability of sensitivity due to individual differences, time, and innumerable internal and external conditions" (Fechner, 1860 p.44), that was found with human detection performance. He hypothesized the existence of a physiological threshold above which a stimulus of a particular intensity could be perceived and below which detection could not occur.

However, evidence began to accumulate that the position of the threshold was influenced by external factors such as the probability of occurrence of the signal over a series of trials, assumed by the subject. Subjects who had been trained to expect a high incidence of stimuli invariably had a lower threshold than those who expected a low probability of stimuli occurrence. In classical psychophysical terms, the subjects were committing the 'stimulus error' by basing their reports on external characteristics of the stimulus rather than their perceived sensations. Another problem was the question of false alarms. If a fixed threshold existed, the observer must either be in a 'detect' or 'non-detect' state. For this reason the occurrence of a 'stimulus' response on a null trial was difficult to account for. In early experiments subjects producing false alarms were simply admonished by the experimenter to take more care. This of course had the effect of biasing them to respond negatively if uncertain.

Later experimenters such as Thurstone sidestepped the problem of response bias by using paired comparison or forced choice techniques. With these methods, a series of pairs of trials are presented, one of each pair containing the stimulus, and the subject simply has to

indicate which one of each pair is the signal trial. However, in many real detection situations the forced choice paradigm is inappropriate.

Blackwell (1952) was concerned with the large scale determination of absolute visual thresholds. By this time it had been realized that even during a trial in which no stimulus was presented, the nervous system of the observer was continually active. Random firing of neurons occurs, exercising a tonic effect on brain functioning, (Pinneo, 1966). It seems likely therefore, that in the case where near threshold signals were to be detected, then sometimes the magnitude of this 'noise' distribution, even on non-signal trials, might be sufficiently great to be construed as a signal. Blackwell's assumption (known as 'high threshold theory') was that the observer's threshold was sufficiently high such that the magnitude of the background noise would never exceed it and give rise to a spurious signal report ('false alarm'). How then, could the undoubted occurrence of false alarms in this situation be accounted for? Blackwell assumed that on a certain proportion of the non-signal trials the subjects simply guessed, incorrectly, that a signal had been presented. A 'guessing correction' was therefore applied to the data to take into account this tendency of subjects.

This assumption, in retrospect, seems unsatisfactory on a number of counts. It is rather arbitrary and still provides no satisfactory account of the effects of, and the reasons for, judgemental bias on the part of the observer.

It seems then, that by the beginning of the fifties the problems of the satisfactory analysis of detection experiments had not been solved and the time was ripe for the introduction of a new theory. Before considering this theory however, it is useful to consider the concept of

observer bias in more detail.

2.0.2 The nature of response bias

Judgemental bias is an all-pervasive aspect of any human choice or decision making situation and is not confined to the rather narrow area of psychophysical judgements that we have just been considering.

A doctor looking, listening or feeling for signs of a disease may far prefer a 'false alarm' to a missed signal, particularly if the disease is serious. In a control room of a nuclear power station the operator may be required to shut down the reactor in the event of certain evidence from the instrumentation, which may be ambiguous, that a dangerous condition has occurred. Shutting down the reactor is usually an expensive business and may cost tens of thousands of pounds in terms of lost output. On the other hand not shutting down the reactor may be even more expensive in terms of damage to plant or even loss of life. The operator's decision will clearly be influenced by these cost factors and also by the relative probability that a dangerous condition is really likely to occur. If it is during the commissioning phase, and similar incidents have been quite common, then the operator will have no hesitation in shutting down the reactor. If on the other hand the plant has been in operation for some years and such an incident has never before occurred, the operator may well defer his decision for some time before he makes the critical control action.

In the area of inspection we can see that the inspector's criterion is going to be influenced by the on-going level of defects present. If it is known that a particular batch has been produced by a 'rogue machine', then the inspector will be far more likely to reject an article that is

a borderline case. A similar effect could occur where cost factors were involved. If the manufacture of an article involved an extremely expensive series of processes, the inspector would be biased to reject the article only if he was absolutely certain that it was defective. Similarly, if the manufacturing process were very cheap but the product was destined for a highly discriminating market, then the inspector would reject if there were any doubt at all that the item was not perfect.

It is important to note that we are referring to subjective probabilities in this discussion, and the costs and values of the observer's decisions are in fact personal utilities. These quantities may or may not be the same as the objective probabilities and payoffs which occur in a real situation. Viewed in these terms we can see that many factors which are known to affect the judgemental process can be mapped on to the dimensions of subjective probability or perceived utility. For example social pressures from co-workers in an industrial situation could be seen to affect judgement via the mechanism of altering the personal payoffs of a particular decision. If an inspector is examining the work of a colleague and is aware that too many rejections could lead to his dismissal, then he may be influenced, consciously or otherwise, to accept a higher proportion of borderline items than would be the case if they arrived from an anonymous source. The subjective estimates of the probability of a defect occurring would clearly be highly dependent on the degree of feedback given to the inspector.

Up to this point we have demonstrated that in detection situations, unless forced choice procedures are adopted, any measure of the observers sensitivity is inevitably contaminated by the factors of response bias discussed in this section. One factor that has not been mentioned up to

this point is the question of quantifying the degree of observer bias. The attitude of classical psychophysics has been that observer bias is a 'nuisance variable' that one should attempt to eliminate as far as possible. However, it is clear that the degree of response bias that occurs in a situation is in itself a quantity of some interest. Given that judgemental bias is a fact of life in any real discrimination task, it is clearly of interest to quantitatively measure the amount of bias, in order to answer such questions as the degree to which some of the factors discussed up to this point affects it.

In subsequent sections we will see that SDT provides the means for quantifying aspects of response bias that have been discussed in this section.

2.1 SDT - general considerations and the basic model

A number of general reviews of SDT now exist, e.g. Pastore and Scheirer (1974), Swets et al. (1961), Egan and Clarke (1966), Coombs et al. (1970), Lee (1971) and Swets (1973). Two textbooks and a book of collected papers have also been produced: Swets and Green (1966), McNicol (1972) and Swets (1964).

Signal Detection Theory has applications in a very wide range of situations in which an observer has to make a discrimination or choice on the basis of equivocal or 'noisy' evidence.

Noise is a central concept in SDT and can be regarded as any random process which tends to interfere with discrimination. We can consider two analogous situations, the detection of a very faint signal in a background which may tend to degrade the signal, and discrimination between

two very similar signals which may also be embedded in a background which tends to obscure the differences between them.

Another source of noise might be the inevitable slight variations in the physical nature of the stimulus, particularly if its precise characteristics are ill-defined. These sources of noise are all external to the observer, but noise is also added internally due to the random firing of neurons in the nervous system (Pinneo, op. cit.).

Whenever a subject makes an observation, the sensory effect that occurs can be represented as a point y in an n-dimensional space, the n dimensions representing the n possible characteristics of the response of the sensory system. For example an inspector may be examining coins for defects by their visual appearance, weight and the sound they produce when dropped. In this case we might expect the sensory evidence x to consist of visual parameters such as size, shape, colour etc., tactile evidence including weight and texture, and auditory information such as the slight difference in sound of a 'dud' coin. It is clear that each coin examined will produce a slightly different point y due both to slight variations within the categories 'perfect' and defective coins and also to the other internal and external sources of noise discussed earlier. It is obvious that the distribution of y values due to good coins, $f_N(y)$, will have a mean less than that due to faulty ones, $f_{SN}(y)$, and that the two distributions will overlap. (The subscript N refers to the fact that the former distribution is regarded as being due to noise only, and SN as the latter being due to signal + noise, the signal in this case being the attributes of a defective coin). If we regard $f_N(y)$ and $f_{SN}(y)$ as being the sensory evidence on which the inspector bases his decision, the question arises as to how he decides whether a perfect or defective item has been presented on a particular

trial.

Signal Detection Theory assumes that as a result of experience on the discrimination task the observer is able to compare the probability that a particular sensory effect has arisen, given that a perfect coin was presented, compared with the probability that it arose, given that a defect was presented. To do this, he forms a likelihood ratio from the ordinates of the two density functions $f_N(y)$ and $f_{SN}(y)$ corresponding to the particular value of the sensory evidence on that trial, i.e.:

$$\frac{f_{SN}(y)}{f_N(y)} = \lambda(y), \text{ the likelihood ratio}$$

The likelihood ratio, then, represents the likelihood that the point y arose from the defect (SN) distribution relative to the likelihood it arose from the perfect (N) distribution. Since any point in the space, i.e. any sensory datum, may thus be represented as a real, non-zero number, all sensory data can be regarded as lying along a single axis. Any observation x can therefore be identified with a particular value of the likelihood ratio $\lambda(y)$. It is convenient if x is identified with a transformation of $\lambda(y)$, i.e. $\lambda(x)$, such that Gaussian density functions of the sensory effects of signal and noise, $f_N(x)$ and $f_{SN}(x)$, result. This is produced by using a logarithmic transform of $\lambda(y)$. A further assumption will be made, although this will be modified later, that the distributions are of equal variance. This gives rise to the familiar SDT diagram, Figure (2.1). The normality assumption can be justified on the basis of the Central Limit Theorem, in that if observations are independent, then the distribution of the sums of the noise and signal plus noise distributions each approach normality for reasonably sized samples.

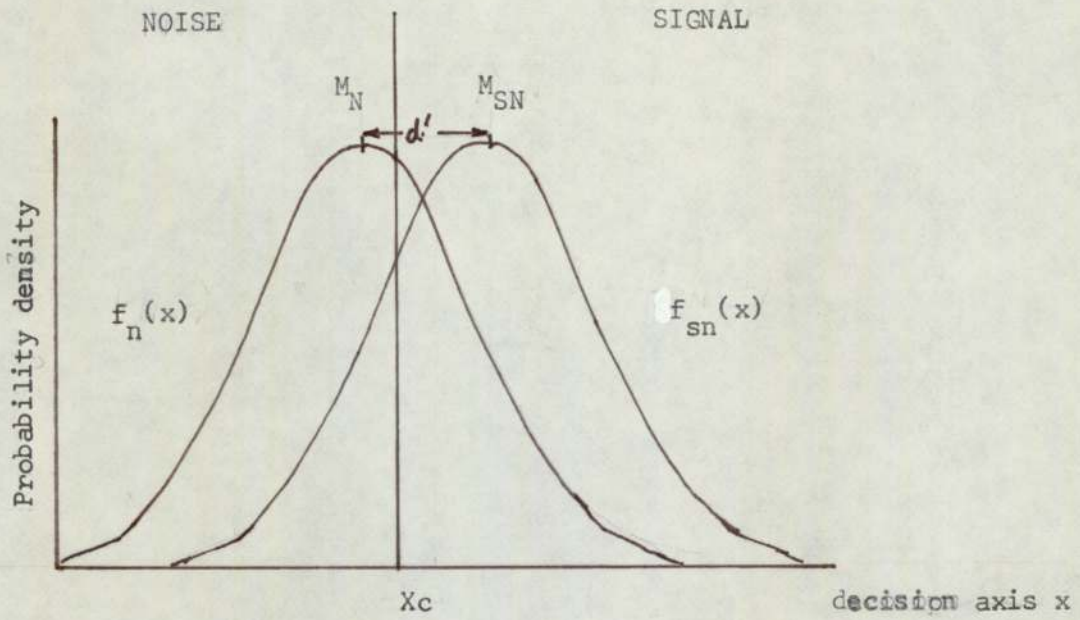


Figure 2.1 Equal variance probability density functions from log transform of likelihood ratio

2.1.1 Measurement of sensitivity and bias

An intuitively reasonable measure of sensitivity is immediately apparent from Figure 2.1 . This is d', the distance apart of the two distributions, scaled in terms of their common standard deviation σ.

$$\text{i.e. } d' = \frac{M_{SN} - M_N}{\sigma}$$

Clearly the further apart the distributions, the more discriminable are the signal and noise. If the two distributions overlapped completely, discrimination would be impossible, and if they were a considerable distance apart, the overlap, and hence the ambiguity involved, would be vanishingly small.

Having mapped the effects of the noise and signal on to a one dimensional axis, the question arises as to how the decision is made. SDT assumes that the likelihood ratio axis is a decision axis, and that the observer establishes a cutoff value of x, x_C, (Fig.2.1), such that if the transformed likelihood ratio exceeds x_C he responds signal, and if less than x_C he responds noise, i.e. non-signal. The position of x_C is defined by the ratio of the ordinates of $f_{SN}(x)$ to $f_N(x)$ at x_C, known as the criterion β (beta).

$$\text{i.e. } \beta = \frac{f_{SN}(x)}{f_N(x)} \Bigg|_{x = x_C}$$

The quantities d' and beta can readily be calculated from an experiment in which the probabilities of correct detections and false alarms can be estimated, using the equal variance assumptions. The results of such an experiment are shown below.

Observer's Decision

state of the world	NOISE N	SIGNAL S
NOISE n	correct 'no signal' response probability p(N/n)	false alarm probability p(S/n)
SIGNAL s	missed signal prob. p(N/s)	correct detection probability p(S/s)

Figure 2.2 Result of hypothetical detection experiment.

The probabilities are estimated from the relative frequencies of the various types of response. From Figure 2.1 it can be seen that a knowledge of P(S/n), the false alarm probability, given by the part of $f_N(x)$ above x_C, and the missed signal probability, given by the portion

of $f_{SN}(x)$ below x_c, together with a table of the areas under the normal curve, will enable d', the distance between the distribution means to be calculated. Similarly a table of the ordinates of the normal distribution will enable $f_N(x)$ and $f_{SN}(x)$ at x_c to be obtained and hence beta.

2.2.2 Some characteristics of beta

SDT assumes that the observer is able to place his criterion at any point on the decision axis x. If the criterion is placed far to the right this will lead to performance characteristic of the cautious observer. Only sensory data falling in the extreme right hand tail of the signal plus noise distribution will elicit a signal response. Performance will be characterized by a low false alarm probability but also by a low signal detection rate. Conversely a criterion placed far to the left (a 'lax' criterion) will produce a high detection rate accompanied by a high false alarm rate. A criterion near the centre of the decision axis will give an intermediate level of both false alarms and correct detections.

We can see that SDT takes a quite different approach to the classical psychophysical theories. Instead of a fixed threshold the criterion is infinitely variable and is independent of the sensitivity index d'.

2.2.3 The position of the criterion as a decision rule

We have not yet considered the factors which influence the position of the criterion. Some insights into this question were gained in the earlier discussion as to the nature of judgemental bias. SDT assumes that the position of the criterion is determined by two factors: the a

priori probability of a signal occurring relative to that of noise, and the costs and values associated with various decision alternatives, e.g. false alarms, correct detections.

The basic assumption is that the position of the criterion chosen by the observer represents the application of a decision rule designed to maximize the payoff of the series of decisions made by the observer during the signal detection tasks. It can be shown (Green and Swets (op.cit.) p.20) that the application of a likelihood ratio decision rule to maximize the expected value (in a decision theory sense) of the observers decisions will also maximize the payoff for a variety of other criteria. It can be shown (Coombs et al., 1970) that the expected value will be maximized if the decision rule is taken such that:

$$\beta_{OPT} \text{ (optimal criterion)} = \frac{P(n)}{P(s)} \cdot \frac{(V_nN+C_nS)}{(V_sS+CsN)}$$

where P(n) = a priori probability of noise

P(s) = " " " " signal

V_sS = value of making a correct detection

VnN = " " " " " rejection

CsN = cost of missing a signal

C_nS = " " making a false alarm

2.3 The ROC curve

The ROC curve, (Receiver Operating Characteristic from the electrical engineering origins of SDT) is a very useful way of representing performance using the SDT approach.

The ROC curve consists of a plot of the probability of correct detections against the probability of false alarms in a detection experiment. As implied earlier, these quantities are estimated from the frequencies of 'hits' and 'false alarms' observed. The pair of probability estimates obtained from an experiment in which the criterion remains fixed and the signal discriminability remains constant produces a single point on the ROC curve. In order to generate a complete curve, the subject has to be induced to alter his criterion to produce a series of points of differing x_c but constant sensitivity. This can be done either by varying the pay-offs for the various decision alternatives or by altering the a priori probabilities. Another method is to employ a rating scale technique, (Egan et al., 1959), whereby the subject has to make ratings as to his degree of confidence that a signal is present on a particular trial, e.g. definitely signal, possibly signal, possibly non-signal etc. This is equivalent to conducting several experiments simultaneously using different criteria, and is regarded as the most efficient way to generate an ROC curve, Green and Swets (op.cit.), McNicol (op.cit.). This technique will be employed extensively in the experimental part of this study. The ROC curves generated by these procedures are shown in Figure 2.3.

Each of the curves corresponds to a signal of different discriminability, and where the equal variance assumptions hold true the curves are symmetrical with respect to the negative diagonal. The constant d' curves are sometimes referred to as isosensitivity curves. A ROC can be regarded as being generated from right to left when the criterion is swept from left to right ('lax' to 'strict') across the decision axis. The positive diagonal represents 'chance' performance and an ROC curve can only be produced below the diagonal by 'malingering', i.e. by

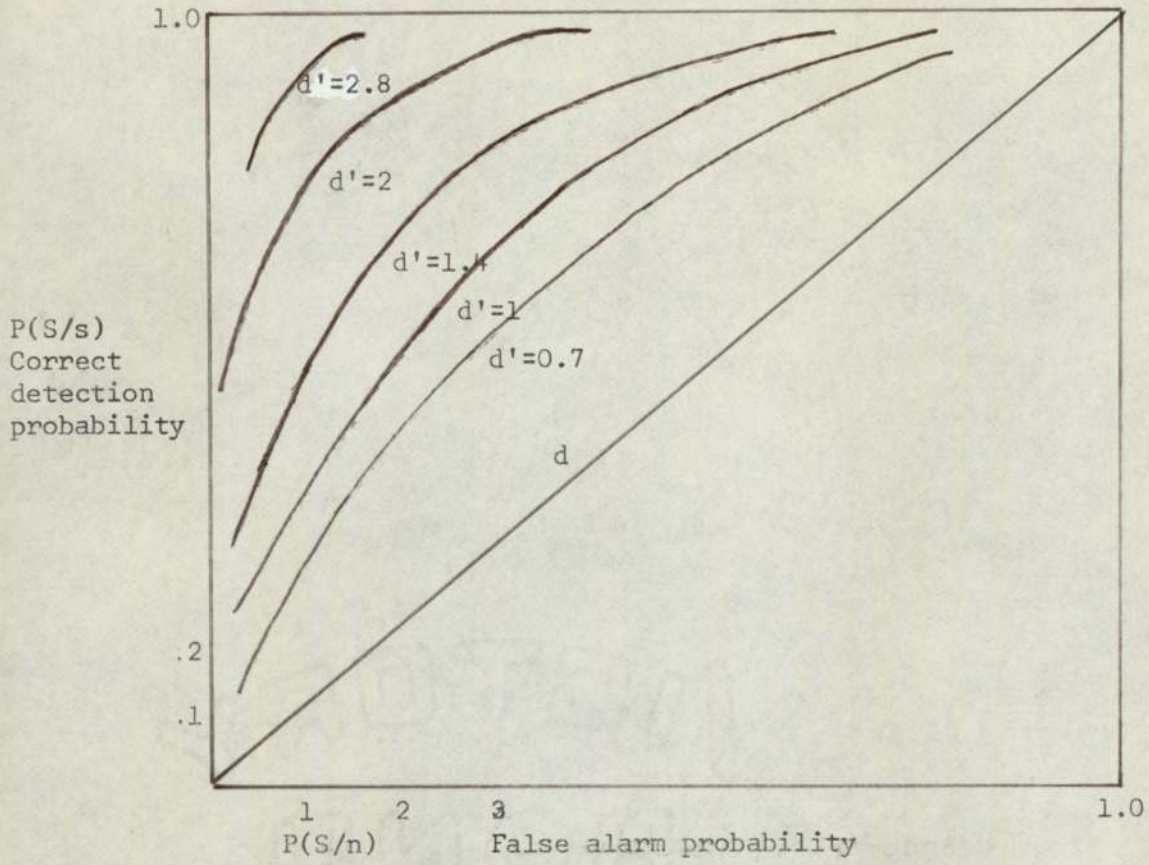


Figure 2.3 ROC curves generated with d' as the parameter

deliberately trying to perform badly. Considerable use is made of normalized ROC curves, sometimes referred to as Z-ROC curves. These are produced by plotting the Z-Scores corresponding to the probability data or by using double-probability paper with axes scaled in terms of the normal deviate. Both of these methods produce straight lines corresponding to the ROC curves of constant d'. In the case of equality of the variances of the underlying distributions these lines are parallel to the positive diagonal, i.e. have slope equal to 1.0. A further property is that the difference between the normalized co-ordinates of any point along the ROC curve is equal to the sensitivity d'.

2.4 Unequal variance model

2.4.1 Experimental consequences

Although many of the earlier experiments involving the detection of

auditory signals in white noise were satisfactorily fitted by the equal variance model described above, it soon became clear that in general an assumption of inequality in the underlying variances provided a better fit to the experimental facts. In many cases no checks were made on the data to find out the appropriate model in a particular experimental situation. A large proportion of the early experiments which employed SDT as a means of separating sensitivity from bias, simply used single estimates of correct detections and false alarms to generate d' and beta measures. These can be obtained without the effort of resorting to normal probability tables, by using tabulated values of d' and beta, e.g. Freeman (1973). As will be discussed subsequently, such a procedure can lead to large errors of estimation in the SDT parameters if the assumption of equal variance underlying distributions is incorrect.

Taylor (1967) has discussed why there is likely to be an asymmetry between the underlying distributions. A signal is normally thought of as something added to the non-signal. The subject normally knows quite well what the non-signal would be, were it not obscured by noise. He does not know so well, however, what the signal would be without the noise. There is an essential asymmetry between the signal event class and the non-signal event class, in that the subject usually knows less about what is a valid example of the signal class than the corresponding non-signal class. It is of interest to note that unequal variance distributions are much more frequently observed with visual signals, which often have complex attributes which cannot be specified exactly, than with auditory signals which can be specified precisely in terms of phase, duration and amplitude.

On purely practical grounds one would expect the variances of the noise and signal + noise distributions to be asymmetrical. In most detection

experiments there are usually far fewer false alarms than correct detections. The variance associated with estimating false alarm probability from a low false alarm frequency is intrinsically greater than with correct detections, where a greater sample size is available.

Whenever a subject knows less about a signal than a non-signal, the same effect will be produced. The ROC curve will cling to the left hand edge of the ROC space longer than it does to the top. The less the observer knows about the exact characteristics of the signal, the more skewed the curve will be.

2.4.2 ROC curve analysis of the unequal variance model

The unequal variance model assumes underlying noise and signal variances of σ_n^2 and σ_s^2. This means that an additional parameter has to be added to the basic SDT model in order to specify the shape of the ROC curve. This is the ratio of the respective standard deviations of the noise and signal plus noise distributions, i.e. $\sigma n/\sigma s$. It can be shown, Green and Swets (op.cit.) p.64 that the slope of the resulting Z-ROC is given by $\sigma n/\sigma s$. As $\sigma s/\sigma n$ increases the slope of the line decreases. A comparison of the Z-ROC's for the equal and unequal variance case is given overleaf. (Figure 2.4).

2.4.3 Measures of sensitivity

One important consequence of the unequal variance situation is that the measure of sensitivity d' is correlated with the criterion position chosen and that this correlation increases as the Z-ROC line becomes less parallel to the positive diagonal. If we consider line A, the

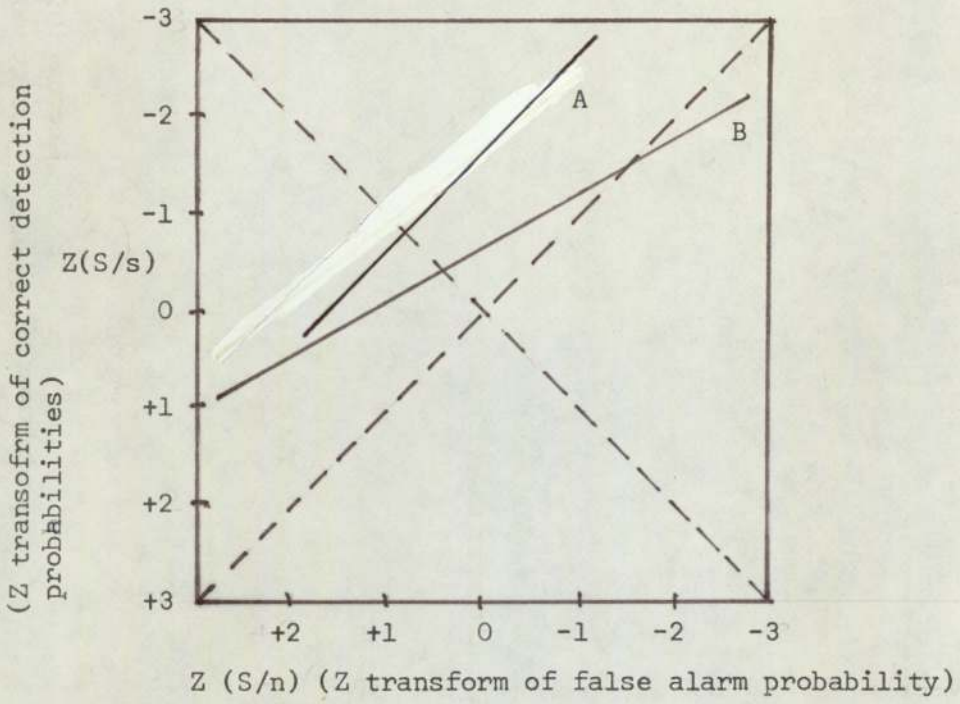


Figure 2.4 Normalized ROC curves representing equal (A) and unequal (B) underlying variances.

equal variance situation, it is obvious that d', (Figure 2.4) the difference between the corresponding Z-scores will remain constant. On the other hand different points on B, the unequal variance Z-ROC line, will indicate different values of the corresponding Z-Scores. It will be recalled that different points on the Z-ROC lines represent different degrees of bias and hence the non-independence of d' is clear. In this situation some decision has to be made about where on the ROC curve the sensitivity measure is to be read. Two sensitivity measures are commonly used. The first of these, Δ m, is the distance between the means of the signal and noise distributions measured in standard deviation units of the noise distribution. It is equal to Z(S/n) at the point on the ROC curve where Z(S/s) = 0.

Another measure of sensitivity is d'e, also called d_s, due to Egan and Clarke, (op.cit.). d'e is defined as twice the value of Z(S/s) or

Z(S/n), ignoring signs, at the point where the Z-ROC curve intersects the negative diagonal. One reason for using d'e is that Z(S/s) and Z(S/n) are equal where the Z-ROC line meets the negative diagonal and hence it gives equal weight to the signal and noise distributions. Since Δ m is scaled in units of the noise distribution, it is the appropriate measure if the noise variance is expected to remain constant over a series of experimental treatments, but the signal variance may change. If both variances are likely to change then d'e would be a more stable measure. Also if we expect signal variance to remain constant and noise variance to change, the most appropriate sensitivity measure would be the value of Z(S/s) at the point on the Z-ROC line where Z(S/n) = 0. This analogous measure to Δ m is scaled in units of the signal distribution and is employed in Thurstonian Category Scaling, (Lee, 1969).

Two final arguments have been advanced in favour of d'e. The first is that the point that it is read from the ROC curve generally falls within the range of responses made by observers. Hence extrapolation is not necessary. Secondly, Egan and Clarke (op.cit.) report that the changes in slope of the ROC curve observed from session to session within the same observer tend to alter the value of Δ m more than d'e, which thus appears to represent a more stable measure. In any event one measure can readily be converted into the other by means of the conversion formula provided by Green and Swets (op.cit.).

$$\text{i.e.} \quad d'e = 2 \Delta m \left(\frac{S}{1+S}\right)$$

where S is the slope of the Z-ROC curve

Theodor (1972) illustrates how incorrect assumptions can lead to erroneous conclusions. The table below represents data for a single subject under three conditions of an experiment with d' calculated

under the assumption that $\sigma s/\sigma n = 1$ and $\sigma s/\sigma n = 2$.

condition	P(correct detections)	P(false alarms)	Z (C.D)	Z (F.A)	d' $\frac{\sigma s}{\sigma n} = 1$	d' $\frac{\sigma s}{\sigma n} = 2$
A	.5000	.0668	0	-1.5	1.5	1.5
B	.6915	.3085	.5	-0.5	1.0	1.5
C	.8413	.6915	1.0	0.5	0.5	1.5

The assumption of equal variances ($\frac{\sigma s}{\sigma n} = 1$) leads to the interpretation that the points are on three different ROC curves of different sensitivity, whereas $\frac{\sigma s}{\sigma n} = 2$ gives the impression that all the points are on the same ROC curve. Unless $\frac{\sigma s}{\sigma n}$ is known, there is no way of telling which hypothesis is true.

2.4.4 Measures of bias

A number of problems arise when measurements of response bias are considered in the unequal variance case. Unfortunately in this situation there is no longer a simple monotonic relationship between the likelihood ratio scale and the underlying evidence variable. For example, in the case where the signal variance exceeds the noise variance, up to the first intersection of the distributions β will be greater than one, after the intersection it will be less than one, and after the second intersection it will be greater than one again. Therefore, there will in general be two values of the likelihood ratio which maximize the expected value of the observer's decision. In actual practice, the second cross-over point will generally occur in the extreme tails of the distributions and in general the observer behaves as if he places his cutoff at a particular value of beta. Even in the equal variance case

there are problems in comparing changes in beta between experimental conditions. If d' is not constant, then apparent changes in beta may be due to changes in d'. Baker (1975) points out that the likelihood ratio can vary between 0.0 and 10.0 when d' = 1.0 and can be as high as 100 when d' = 3.00.

One measure of bias which does overcome this difficulty has been proposed by Banks (1970). This is C, the distance along the likelihood axis from the noise distribution mean to the criterion scaled in Z-units of σ_n. The range of C is not a function of the separation of the distributions and it is always monotonic with the likelihood ratio axis. C for any point on the Z-ROC curve can be determined from the Z score on the false alarm axis. The other bias measurement often used is log β, this being a monotonic function of the evidence variable, i.e. the magnitude of the sensory evidence which the observer uses as input to the decision making mechanism.

Some of the problems of assessing response bias using beta are discussed in McNicol (1972) p. 119. One of the difficulties in using a single index of bias is that in a multiple criteria situation such as occurs in rating experiments, there is no indication if the observer has moved all his criteria up or down the x-axis, or has spaced them closer or wider apart, as found in Broadbent and Gregory (1963) for example. It is of course open to question whether the observer actually uses a likelihood ratio criterion in setting his cutoff. Although such a criterion achieves the broadest range of objectives, there is no guarantee that the observer will use the most rational criterion. It is not impossible that the response criterion be based purely on the evidence variable, the sensory effect produced by the stimulus. In this

case the observer will use a value of y as his criterion, and if the sensory evidence is less than y he will respond 'noise' and if greater than y, 'signal'. In the equal variance case this would produce no anomalies because the likelihood ratio is monotonic with the evidence variable. In the unequal variance case discrepancies would occur. Ingleby (1974) presents persuasive evidence that the decision maker actually does employ a likelihood ratio criterion, however, in auditory detection experiments in which the observer's criterion was systematically varied.

Dusoir (1975) critically reviews the attempts that have been made to measure observer bias using a wide variety of models in addition to SDT. He concludes that none of the existing indices of bias account for all the experimental evidence, and that it may be futile to search for such an index which remains invariant under all types of task, subject and experimental conditions. He also points out that the form of the iso-bias (ROC) curves needs to be established before experimentally manipulating such factors as the a priori signal probability or the payoffs in an attempt to modify the observers bias over experimental conditions. Finally he suggests that inferences about the degree of bias change in groups of subjects should not be used until it is clear that the iso-bias curves are all of the same family for different subjects. However, in spite of these criticisms he does not suggest a usable alternative bias measure, which somewhat weakens the validity of his criticism.

It must be emphasized that this study is not concerned primarily with the theoretical issues considered by Dusoir. As will be discussed in detail subsequently, the intention is rather to investigate the utility of detection theory ideas as a tool in a practical setting. The

intention is to adopt a more rigorous approach that has hitherto been used in real world applications of SDT in order to establish the validity and usefulness of the model in this context. From this standpoint we shall assume that the theoretical validity of the SDT model has been sufficiently well established by the body of evidence in existence. The data provided in this study will serve to provide further confirmatory or otherwise evidence for the theory in the context of real-life tasks.

2.5 Non-parametric indices of sensitivity and bias

As will have been gathered from the preceding sections, the question of the nature of the underlying distributions creates problems when using SDT to measure changes in sensitivity and bias. Fortunately there exist several 'non parametric' measures of sensitivity and bias that require fewer assumptions concerning the nature of the underlying distributions than the parametric indexes d' and beta.

The first of the sensitivity indices to be considered is P(A), the area under the ROC curve. As shown in Figure 2.3, as the index d' increases, the ROC curve becomes closer to the top left-hand corner of the unit square (within which the ROC curve is drawn). Green and Swets (op.cit.) p.45 show that the area under the curve is a measure of sensitivity independent of the shape of the underlying distributions. P(A) lies between 0.5 and 1.0. If an ROC curve is generated using the rating procedure mentioned earlier, a numerical integration technique such as the trapezoidal rule can be used to estimate the area under the curve. McNicol (op.cit.) p.114 gives an example of the technique. Pollack and Hsieh (1969) investigated the sampling distributions of both d'_e

(discussed earlier) and P(A). The expression they obtained for the standard deviation of P(A) is useful in statistical analysis using these indices. Where only a single point on the ROC curve is available another measure of sensitivity is available, A', due to Pollack, Norman and Galanter (1964) and Pollack and Norman (1964). This index is an approximation based on the measure P(A) discussed earlier. It is derived by considering the maximum and minimum values that P(A) can take. Grier (1971) provides a convenient computing formula for this measure and also for P(I), another sensitivity index suggested in Pollack and Hsieh (op. cit.) which is related to A'. Another index which utilizes rating scale data is AH, due to Hammerton and Altham (1971) and Altham (1973). This index has been criticized by Dusoir (op.cit.) because although it makes no assumptions about the underlying variances, it does assume that the observer employs a likelihood ratio criterion. **Navon (1975) has produced a sensitivity index derived from response latency measures.** It is clear that there is a considerable choice of indices available to measure sensitivity. The same cannot, however, be said about bias. Hodos (1970) developed a non-parametric measure of bias B, based on the fact that the negative diagonal of the unit square represents the locus of points where the subject would be equally likely to respond signal or noise given ambiguity. The measure reflects the degree to which a data point deviates from the negative diagonal relative to the maximum possible deviation. A computational formula for B is given by Grier (op.cit.). Hodos' measure is however criticized by Dusoir (op.cit.) as not being 'non-parametric' according to his definition and as being simply an arbitrary parameter that makes no specific reference to any sensitivity parameter.

Apart from B the only other measures of bias available within the SDT model are beta, log beta and C, as discussed earlier, although

McNicol (op.cit.) p.123 presents a procedure for deriving a non-parametric measure of bias from a rating experiment where there is insufficient data to obtain a beta value for each criterion. Only a single overall measure of bias is produced by this technique and it is, at best, a somewhat crude estimate.

2.6 Signal detection theory and inspection

It will be recalled that the basic ideas of the SDT model were developed using an example from the inspection area. The advantages of using the SDT approach in examining inspection tasks are numerous.

One of the most important applications of the SDT model is to provide an index of inspection performance. This is an important issue,because if a variety of different methods are employed to measure inspection performance,it is extremely difficult to compare inspection studies to assess the effects of differing factors on performance, as we shall see during the literature survey. The most common performance index employed in industry is,of course,the percentage of defects detected. McCornack (1961) discusses a number of other indices that have been employed and suggests that the suitability of a particular performance index depends on the objectives of the inspection system. For example different indices are appropriate if the object is to maximize correct detections regardless of false alarms, or whether false alarms should be minimized. Sosnowy (1967) also considers some of the performance indices available and shows that the ranking of inspectors in terms of efficiency can vary drastically depending on which of the performance measures is employed.

Signal Detection Theory, with its separation between sensitivity and bias, offers unique advantages as an inspection performance index. We can identify some of the requirements of an ideal index as below:

1. Should provide insights into why performance is good or bad in a particular case.
2. It should allow quantitative costs and values to be assigned to the various types of errors and correct decisions possible.
3. The index should separate the aspects of performance due to the inspector's sensitivity, from his response bias.
4. The performance index should enable inspection performance to be related to general theories of human performance.

It is clear that the signal detection indices of sensitivity and bias fulfill these requirements.

The likelihood ratio concept is particularly useful, in that it suggests an index of performance that has a theoretical optimum. We can thus compare the actual performance of the inspector with the optimum to see how it needs to be changed to produce a more efficient inspection system. The separation of performance into sensitivity and bias variables also provides insights into the way in which performance can be improved. Sensitivity, for example, can be improved by training the inspector to recognize the whole range of attributes that characterize good and bad products. The ability of observers to alter their response criterion on the basis of instructions suggests that inspectors can be induced to alter their response strategies towards the optimum. This area will be discussed in more detail subsequently.

When using SDT parameters as indices of inspection performance a certain amount of confusion is possible due to the differing definitions of what constitutes a 'signal' in detection experiments and in inspection situations. The two matrices in Figure 2.5 contrasting a detection and an inspection experiment, will make the distinction clear.

observer's decision

state of the world	NOISE N	SIGNAL S
NOISE n	correct 'no signal' response P(N/n)	false alarm probability P(S/n)
SIGNAL s	missed signal probability P(N/s)	correct detection probability P(S/s)

inspector's decision

state of product	perfect N	defect S
perfect n	correct 'product perfect' decision P(N/n)	false alarm or good product called bad P(S/n)
defective s	missed defect P(N/s)	defect correctly detected P(S/s)

Figure 2.5 Comparison of signal detection and inspection situations.

2.7 Relation between the SDT model and acceptance sampling

In industrial quality control, considerable use is made of various types of statistical sampling plans. It is useful to clarify the relationship between SDT concepts and those of statistical quality control (S.Q.C.).

SQC adopts the standard statistical usage of referring to false alarms and misses as Type I and Type II errors respectively. The missed signal probability $P(N/s)$ corresponds to failure to reject a false null hypothesis and is referred to as β in SQC (not to be confused with the entirely different usage of β in SDT). It is also known as the consumer's risk in acceptance sampling inspection, i.e. the risk of accepting a bad

lot. Similarly the probability of false alarms, P(S/n), is analogous to the quantity α in SQC, known as the producer's risk, the risk of rejecting a good lot.

The SDT criterion x_c is analogous to C in sampling inspection where C is the number of sampled defective items which must be exceeded in order to reject the entire lot as defective.

Thus the SDT conceptualization of the inspector as an inferential decision maker is very similar to the theories of inferential acceptance sampling as practised by quality control engineers and statisticians. This adds further weight to its adoption as a conceptual paradigm in inspection.

2.8 Applications of SDT in inspection studies

Considering the very obvious advantages of the SDT approach, it is surprising how infrequently it has actually been applied in the inspection area. This is probably a reflection of the fact that inspection in general has been an under-researched area as far as human factors is concerned. Even where SDT has been applied to inspection this has often been in the context of laboratory studies rather than real life tasks. The methodological issues of applying SDT to inspection situations are dealt with explicitly in most of the papers to be considered in this section, mainly because SDT is still relatively unfamiliar to ergonomists with an applied orientation. A paper by Baker (op.cit.) comprehensively reviews the whole area of SDT applicable to inspection tasks, without being orientated towards a specific study. Other review papers are Drury (1975), which also considers Information Theory and the application of Bayes' theorem to inspection, and Adams (1975). The

growing interest in the application of SDT to inspection was very evident at the 1974 symposium on human reliability in quality control held in Buffalo, U.S.A., where virtually every paper contained some reference to the theory.

If we exclude studies in such areas as sonar detection, where SDT was applied as early as 1967 (Colquhoun, 1967), the first published paper employing SDT in an inspection context was Wallack and Adams (1969), although it reported earlier unpublished work by Wallack (1967). The Wallack and Adams study will be considered in some detail, because its methodological shortcomings serve to exemplify the problems which ensue when SDT is used without due regard for the underlying assumptions of the model. In this study, inspectors were required to examine samples of 260 electrical cables for examples where the conductors had been nicked or abraded in a wire stripping operation. Four sample lots were inspected containing 5, 15, 25 and 35 percent defectives. The inspectors were trained with a scheme which presented them with samples containing 80, 60, 15 and 35 percent of defects respectively, the higher incidence of defects sample being used as a teaching aid in which each wire inspected was discussed with the trainee to provide complete feedback. A payoff matrix was then assigned such that correct acceptances and rejections had a value of 1 unit and missed defects and false alarms were associated with a cost of 3 units. These values were purely abstract - the inspector did not receive any concrete rewards or payoffs for his performance. The inspectors were given further practice until they were familiar with the payoffs, the training terminating when they were able to achieve a particular level of payoff. The experiment proper was conducted such that the different incidence of defects samples were inspected at random times during the week. The

SDT parameters d' and beta were calculated from the incidence of correct detections and false alarms, as outlined earlier. Also calculated was the distance of the criterion from the noise distribution. The results were interpreted by the authors as follows. The obtained mean beta values for each of the different probability of defects samples were not in general equal to the theoretically optimal values predicted by SDT, and the discrepancy was greatest with the lowest defect probability. The obtained betas did not even give the same rank order as the optimal ones. Two reasons were suggested for this. One was that the inspectors did not employ the payoffs assigned, and the other that they were not a homogeneous group, and could be divided into two sub-groups with differing degrees of bias.

The major general criticism that can be levelled at this study, was that no attempt was made to examine ROC curves for the experiment, to check if the data did conform to the SDT model. Without such evidence, any conclusions drawn from the SDT parameters must remain highly questionable. The failure to draw ROC curves is particularly strange, in that data to do this are published in the paper in the form of correct detections and false alarms for each incidence of defects for each inspector. According to the SDT paradigm, the differing a priori probability of defects should induce a series of different criteria in the inspector and hence generate an ROC curve. The ROC curves, when plotted, show that although two of the seven inspectors appear to produce a straight line Z-ROC, the slope does not support the equal variance assumption and hence d' and an uncorrected value of beta do not provide meaningful measures of bias and sensitivity. The fact that some of the results are not fitted by the SDT model does not mean that it is necessarily inappropriate. The fact that the inspectors did not receive any training with some of the defect levels employed could

account for the results for example. In view of these considerations, attempts to compare obtained values of beta with the theoretical optima are obviously misplaced. Another methodological point concerns the instances in the data where the false alarm probability is zero or the hit probability is unity. In these situations, the corresponding z-scores tend to plus or minus infinity and hence neither d' nor beta can be calculated. In spite of this, values of d' do appear in the results table at these places. This is because the authors have used one of the approximations that can be employed in these situations and which will be discussed in detail later. If such approximations are used it is essential that they be made explicit. In the data under consideration they produce inflated values of d' that are incorporated in the group means. It is obvious that SDT cannot be applied in such a casual manner if meaningful results are to be obtained. In SDT terms, the only concrete facts to emerge from this experiment are that some inspectors' performance can be described by the unequal variance model, and that inspectors do not appear to change their criteria according to the a priori probabilities.

The next published paper to use SDT in an inspection context was Embrey (1970). This concerned the inspection of bubble chamber photographs, produced in high energy physics investigations, for the occurrence of particular configurations of tracks. SDT was employed to ascertain whether differences in detection efficiency were due to the differing discriminability of the two configurations employed in the experiment or to differences in response bias on the part of the inspectors. The effects of differing levels of ambient noise, and time on the task were also considered. Although the results were of interest, the study was again a naive application of SDT in that ROC curves could not be plotted and the basic assumptions remained unverified. Wallack and Adams (1970)

reanalyse the data of their earlier paper using the measures of McCornack (1961), discussed earlier. A Bayesian measure of performance is also calculated, and a measure due to Freeman et al. (1948), which is related to statistical quality control considerations. It is pointed out that SDT measures of inspection efficiency are the only ones available that consider the effects of the costs of the various decision alternatives.

Lusted (1971) used SDT to analyse the performance of radiologists evaluating X-ray photographs. He found that the lack of agreement between radiologists on the diagnosis of the photographs could be explained in terms of differing criteria rather than differing sensitivities. ROC curves plotted for this task indicated that the unequal variance SDT model gave an excellent fit to the data. Lusted used the sensitivity parameter d'e to compare the effects of alternative presentation methods. ROC curves were also used to show that paramedical personnel had a lower sensitivity than radiologists in this task. Interestingly enough, the ROC plot for this experiment also indicates that the less experienced paramedical group showed a greater σ_s/σ_n ratio than the radiologists, presumably because they were less familiar with the characteristics of the signal. The data from these studies are not presented in detail but the use of d'e, an appropriate sensitivity model for the unequal variance case, gives one far more confidence in the author's conclusions. This study is one of the very few practical applications of SDT in which such a sophisticated approach has been adopted. Another study in the radiology area, Sheft et al. (1970), used the measure d'e to show that the detection performance of X-ray technicians who had received five months training was indistinguishable from that of senior consultants.

The next industrial application of SDT to appear was Sheehan and Drury (1971), the detailed results of which were published earlier, Drury and Sheehan (1969), but not analysed in SDT terms. Inspectors with a wide spread of age and visual acuity, but judged to be of equal competence by the company, inspected 6 batches of steel hooks containing 20 percent of items with a single defect and 5 batches each containing a similar proportion of defective items, but each item contained two defects. From the point of view of SDT, one of the main results of interest was that when the usual probabilities were plotted on an ROC curve, most of the inspectors results seemed to be moderately well fitted. However, one inspector's data seemed completely discrepant. Subsequently it was found out that she was rejecting a large number of acceptable hooks because of a surface blemish which should actually have been ignored. The authors suggested that a particular advantage of the SDT approach was that it enabled such an error to be more readily detected than using other indices of detection performance. The ROC curve also indicated that all of the inspectors were employing a stringent criterion, leading to a low level of false alarms but a relatively high incidence of missed defects. In this particular inspection situation, this was an inappropriate strategy, since missed defects were 'expensive'. Using tables from Swets (1964), (which assume equal variance underlying distributions), values of d' were calculated and the results used to show a negative correlation between age and sensitivity. d' was also used in a second experiment to show that prior knowledge as to which type of defect was going to occur, enhanced sensitivity. This study was considerably more satisfactory than the first one considered, in that an ROC curve was drawn, and no attempt was made to calculate beta with inadequate information. However, there were still several methodologically suspect aspects. Rather than attempting the difficult task of fitting ROC

curves to the data points by eye it is far more sensible to convert them to z-scores or use double probability paper to produce a straight line z-ROC plot, thus facilitating an objective test of the fit of the data to the equal variance assumptions. The usual criticism of the use of d' without testing the assumptions applies. One of the nonparametric measures would have been more appropriate. The paper discusses the ways in which the inspector might be induced to vary his criterion by rapid feedback of the results of his inspection.

The next paper, Drury and Addison (1973) represents the best and most extensive available application of SDT to inspection tasks. There are a number of reasons for this. The study used data from an on-going industrial task rather than a simulation or a laboratory study. Techniques were presented for obtaining estimates of SDT parameters from the data normally available in inspection situations. The data was tested carefully for its conformity with the SDT assumptions, and the theory was used to gain new insights into the way in which the inspectors performed their task. In the experiment, inspectors in a glassworks, 100% examined certain unnamed expensive glass products (actually colour television tubes). There was a sample inspection by special examiners of the items classified as good and faulty by the 100% inspectors. During the period of measurement the special examiners were moved to a point closely following the 100% inspection and their results made available to the inspectors much more rapidly than before. Twelve consecutive weeks performance before and including the change were measured and eleven weeks after the change. Each week's data consisted of the percentage of items classified as good by the inspectors together with the special examiners reports of the percentage of the rejected items that were faulty, and the percentage of the rejected items

that were good. The latter two quantities are subject to variable degrees of sampling error, due to the variable size of the sample and the fault percentages involved. It also seems likely that the standards of the sample examiners were probably subject to the same limitations as the 100% inspectors, e.g. pressures due to the varying nature of customer's requirements.

The authors point out that although, a priori, the data seem unlikely to conform to the SDT model, the situation does have some resemblence to the classical SDT experiment where some uncontrolled variation still exists. They emphasize that they are attempting to analyse group performance in SDT terms, and that the situation is analogous to the usual SDT experiment.

By using a decision tree approach some extremely useful expressions are derived for obtaining estimates of the quantities P(S/s), P(S/n) and the a priori probability of a defect occurring, from the percentage of items classified as good by the inspectors and the special examiners estimates of the percentage of Type II and Type I errors. These expressions are of considerable general utility, in that they are often available in real life inspection situations, where the actual a priori probability of defects is not usually known of course. The expressions are given below:

P_1 = probability that a good item will be accepted

$$= \frac{(1 - y)\, P_A}{x(1 - P_A) + (1 - y)\, P_A}$$

P_2 = probability that a faulty item will be rejected

$$= \frac{(1 - x)\,(1 - P_A)}{yP_A + (1 - x)\,(1 - P_A)}$$

P_G = a priori probability of a defect = $x (1 - P_A) + (1 - y) P_A$

where P_A = proportion of items inspected which are accepted as good.

x = proportion of items rejected by 100 % inspector which are in fact good.

y = proportion of items accepted by 100% inspector which are in fact faulty.

x and y are both obtained from the special examiners' sample inspection.

The results were first fitted by Z-ROC curves and it was found that the pre and post feedback data (subsequently referred to as 'before' and 'after' data) could be fitted by two straight lines of slope not significantly different from one. The good fit of the data to the equal variance SDT model greatly facilitated subsequent analysis. The fitting of straight lines to this type of data using least squares techniques cannot normally be recommended since there are errors in both variables. Least squares techniques assume that one of the variables is independent. Maximum likelihood fitting techniques have been advocated in this situation by Dorfmann and Alf (1968, 1969), Ogilvie and Creelman (1968) and Grey and Morgan (1972). In view of the very high correlation coefficients obtained for the two lines (0.803 and 0.802), perhaps this criticism is unnecessarily purist, but it could become important in situations where the scatter of the points were greater.

It was shown that the detectability of the defects, as measured by d' increased significantly after the introduction of feedback. The reasons for this are not entirely clear unless the examiners gave the inspectors insights into the nature of the critical defects. A plot of fault density against probability of detecting a defect showed the decreasing relationship that has often been observed in inspection studies, e.g. Fox and Haslegrave (1969).

As pointed out in section 2.2.3, SDT suggests that if an inspector uses a likelihood ratio criterion to maximize the expected value of his decisions, then (using the terminology of this paper):

$$\beta \text{ optimal} = \frac{P_G}{1 - P_G} \times (\text{relative cost factor})$$

Therefore a plot of beta against $P_G/(1 - P_G)$ will be a straight line through the origin. Drury and Addison's data confirm this prediction only for the data from the feedback conditions. In other words, the inspectors were only able to adjust their criteria optimally to the incoming glass quality when feedback was provided. This result might be expected from the assumption that the inspector requires some means of adjusting his subjective estimate of the defect incidence in order to adjust his criterion. This is provided more effectively by the direct feedback from the special examiners than by the intrinsic feedback present in the task.

The paper also suggests that the inspectors change their criteria to keep their outgoing type II error (y) constant, since this is one of the important factors on which their performance is judged. The data presented seem to lend support to this hypothesis.

This paper shows how SDT measures can be used to gain insight into a wide range of inspector behaviour. There are certain criticisms that can be made. One is the use of least-squares techniques in fitting ROC curves, as already mentioned. Another concerns the use of the aggregated measures of the number of inspectors. Although the overall performance of the inspection system is well described by this approach, it masks the considerable inter subject variation that must be present. In order to generate an ROC curve at all, the inspections must be utilizing a very wide of criteria. In view of the fact that there was

no systematic large change in the aspects of the task likely to affect the criterion, the wide variation in this parameter is surprising. Also the use of grouped data conceals the idiosyncrasies of individual performance which would need to be investigated in order to improve the overall efficiency of the inspection system. For example it is possible that a small proportion of inspectors with low sensitivity and inappropriate criteria may be adversely affecting the system. In spite of these criticisms, this study illustrates the insights that can be gained into an inspection system by the use of SDT.

Other studies employing SDT have paid less attention to the underlying assumptions of the theory. Smith (1975) used d' in a study on the optimal magnification levels for microminiature inspection. Buck (1975) discusses the applications of SDT in the dynamic visual inspection area. He concludes that the theory is only likely to be useful in situations where the relationships between task parameters and their effects on SDT variables can be established. A study by Zunzanyika and Drury (1975) used the rating scale technique, where the inspectors were required to place inspected items in boxes labelled 'definitely accept', 'probably accept', 'probably reject' and 'definitely reject'. The experiment was designed to investigate the effects of various types of information on inspection performance. The three conditions considered were feedforward, where the inspectors were given prior knowledge of the type of defect likely to occur, feedback, where knowledge of results were provided, and a combination of these two information sources. Batches used in the study consisted of 10, 20 or 30% defects. The control group inspected the same batches as the experimental group but without the information conditions. For some reason, the rating part of the experiment was not used to generate an ROC curve, but was instead used to

provide three estimates of d' from the boundaries of the different categories employed. The hypothesis was that if SDT applied then d' would remain constant for the different criteria generated by the rating procedure. The results suggested that SDT did apply to this study, since there were no significant differences in d' across criteria. The differences between the various information conditions was no greater than that due purely to learning in the control group. Surprisingly, the authors state that criterion information is lost with studies of this type. In fact it is simple to calculate the position of the series of criteria that result from rating scale experiments. The applicability of SDT was further tested in a similar manner to the Drury and Addison study, by plotting the beta values obtained from the different a priori defect levels, against the ratio of the probability of a defect occurring to that of the probability of perfect product. The experimental group was fitted by such a straight line, but not the control group. In spite of this partial confirmation that SDT was applicable, both the sensitivity and criterion were affected by the changes in the a priori probability of a defect, an effect that was statistically significant. It is not easy to draw any definite conclusions from this study regarding the applicability of SDT, partly because of the equivocal nature of the results and partly because no ROC curve was drawn and hence we cannot make any inferences about the nature of the underlying distributions.

Most of the other applications of SDT have used the theory merely to provide the quantities d' and beta, using the equal variance assumptions and without any particular theoretical discussion of the applicability of the theory. Examples of such approaches have been Moraal (1975), in the inspection of steel sheets and Chapman and Sinclair (1975),

concerning the inspection of food products.

2.8.1 General discussion of the literature

At first sight SDT appears to be ideally suited for application in the area of industrial inspection. The separation of sensitivity and response bias, the existence of normative standards for the criterion, and the possibility of incorporating the costs and values of the inspector's decision making into the model are powerful arguments for its use. The analysis of the preceding papers has suggested that up to now, however, SDT has been used in a somewhat naive manner in the inspection area. As has been suggested throughout this section, unless the basic assumptions of the theory are shown to be applicable to the data under consideration, then any conclusions drawn from the use of d' and beta must be viewed with considerable reservations. Some workers in the applied area, primarily Drury and Lusted,have taken the precaution of checking the SDT assumptions before using the theory. Also there has been some recognition by these workers of the difficulty of applying the parametric measure of bias, beta, in situations where the SDT equal variance assumptions do not hold. Even where the ROC curve has been drawn, there has been a tendency to assume that an approximate fit of the data to freehand curves is sufficient evidence that SDT can be applied. A far more accurate procedure is to transform the probabilities to Z-Scores to give a Z-ROC plot which can be readily fitted by a straight line either by eye or by the maximum likelihood techniques outlined in Grey and Morgan (1972). Such a procedure allows the slope of the line to be evaluated such that decisions can be made as to which form of the SDT model is appropriate.

One aspect of industrial experiments that is of interest concerns the way that the ROC curve is generated. In all industrial experiments to date, no explicit attempt has been made to generate the curves by the procedures used in laboratory experiments on SDT, i.e. by manipulating the a priori probability of signals or by varying the payoffs. The general practice has been to simply plot the probabilities of correct detections and false alarms (or their Z-Scores) and by good fortune there has been sufficient random variability in the response bias to generate an ROC curve. It is certainly of interest to see the wide variability of criteria that exist in data taken from industrial situations. From the Z-ROC curves in Drury and Addison (1973) it can be calculated that the criteria vary between 1.02 to 5.03 in terms of beta. These data were, of course, produced by a wide variety of personnel working in conditions where there might well be frequent changes in criterion due to alterations in customer standards. In other industrial situations, it is possible to foresee problems if the inspectors' criteria were to be very homogeneous and stable. This might lead to an inadequate spread of points to establish the ROC curve. In this situation it might be necessary to perform supplementary experiments using a rating scale approach, in order to provide the range of criteria necessary. When one is attempting to apply SDT in industrial situations it is essential that the subjects are thoroughly experienced in the task. Otherwise they cannot be expected to have acquired the necessary knowledge of the conditional probability distributions to make decisions in accord with the SDT model.

There is so little data available that it is difficult to say whether all the predictions of the model have been verified in the area of quality control. Certainly, the Drury and Sheehan, Drury and Addison and Lusted data, lend support to the basic precepts of the model

concerning the underlying normality of the sensory distributions. In all cases the ROC curve provided an acceptable fit, although only Drury and Addison provided a test of statistical significance for this, and their fit was obtained using an inappropriate least squares procedure. As far as the other aspects of the model are concerned, which predict that the inspector will attempt to maximize the expected value of his decisions by modifying his criterion in terms of the equation given in section 2.2.3 i.e. beta = (prob. good/prob. defect) x (relative/cost factor), the evidence is more equivocal.

Drury and Addison plotted beta against (probability of good product/ probability of bad). In terms of the above equation, this should produce a straight line through the origin. This was found to be the case for the sessions where feedback had been provided but not before. This can be accounted for by assuming that one of the results of feedback was to give the inspectors a more accurate subjective estimate of the proportion of defects present, thus facilitating the optimization of their criteria. An alternative explanation is that the effect of feedback was to stabilize the payoff matrix by the special examiners providing fixed standards for the relative costs of the different types of error. The authors show that yet another possibility is that the inspectors modified their criteria to maintain a constant output proportion of defectives, this being one of the factors on which their performance is judged and which affects their relationships with the customers. In fact, the data suggest that this hypothesis is also reasonable. In terms of the SDT model, the inspectors could be regarded as altering their payoff matrix to keep the output proportion of defects constant, via a modification in the criterion.

It seems then, that on the modest amount of evidence available, SDT does provide useful insights into inspection performance. However, the situation is complex, and needs to be investigated further, particularly from the standpoint of the effects on the criterion of various aspects of the task.

Very little work has been done on the inspector's ability to inspect to a specific set of payoffs.

In summary the following requirements are important in order to meaningfully apply SDT in inspection situations:

1. The inspectors should be well practised.
2. The data should be subjected to an ROC curve analysis to ascertain if the underlying assumptions are correct.
3. If the spread of the criteria are insufficient to define an ROC curve, then some form of off-line experiment may be necessary employing the rating technique or some other means of generating a range of criteria.
4. Straight line Z-ROC plots are to be preferred to attempting to fit complex ROC curves by hand to the data. An accurate Z-ROC enables the ratio of the variances of the underlying distributions to be established. If data of the right form is available it is best to fit the data points using a maximum likelihood ratio technique.
5. Data for groups of subjects should not be combined unless it has been established that they all conform to the SDT model.

2.9 Directions for research into inspection using Signal Detection Theory

Two main areas of research work can be identified in which further SDT

orientated studies would be of interest. The first of these is concerned with the application of the theory to on-going inspection situations. As has been repeatedly emphasized in the last section, the amount of data available from real life studies which has been examined in a rigorous manner, using SDT techniques, is very small. Further studies are necessary to establish the range of application of the theory. Additionally, it would be useful to investigate the usefulness of the various non-parametric measures of sensitivity and bias in applied situations. If it could be established that at least some of these measures were relatively insensitive to variations in parameters such as the slope of the ROC curve, then they might be more readily used in real-life inspection situations than the corresponding parametric measures.

The other main area of research work concerns the use of SDT to investigate a number of more specific aspects of the effects of various taks parameters on inspectors' performance. Perhaps the most interesting of these, and one which is particularly suited to an SDT approach, is the question of the effects of changes in fault density in the incoming items to the inspector. We have seen from the Drury and Addison study that it is not at all clear whether the inspector actually employs the optimal criterion suggested by SDT or whether he bases his criterion on other considerations such as the outgoing percentage of defectives. Further studies are needed to establish in greater detail how the inspector responds to the a priori probability of a defect. Another aspect of this question concerns the inspectors response to change in defect probability within the task itself. Studies in both the signal detection area and inspection have tended to concentrate on static tasks. If one considers an inspector examining products in a continuous flow system, for example, it is possible that one of the

manufacturing units may develop a fault which suddenly increases the incidence of defects. An analogous situation would be if an inspector was transferred from a situation where a low incidence of defects were the norm to an inspection line where a more inherently faulty product were being inspected. Equally the fault density might change from high to low in the latter situation. It would be of interest to consider the inspectors reaction to change in these cases. Would they, for example, move their criteria in the optimal direction predicted by SDT?

In terms of the SDT model, where the defect probability increases, the inspector should lower his criterion to one appropriate to the new defect probability and make a greater number of 'defect' responses. But such an adjustment presupposes that the inspector has a perfect knowledge of the new defect probability and that he is able to act on this knowledge by adjusting his criterion. It is clear that the inspector will only have a limited sample on which to base his estimate of the new a priori defect probability. An accurate estimate would only be available if the inspector were able to discriminate perfectly between defects and non-defects, or if complete knowledge of results were available. In the absence of this, his subjective estimate of the degree of change in fault density will be a function of his intrinsic sensitivity for defects. We know that human beings are very conservative interpreters of evidence as far as the revision of subjective probabilities is concerned (Edwards, 1962, Slovic et al., 1975). Bayes' theorem indicates the optimal degree to which subjective probability estimates should be revised on the basis of evidence, but the experimental evidence suggest that subjects do not generally do this to the optimal degree. These two factors, the limited evidence available due to the finite sensitivity of the operator, and his innate

conservativism, mean that his subjective estimate of the defect probability will lag behind the actual probability. The ability of the inspector to adjust his response bias to the new probability can be regarded as a separate attribute to his skill at estimating this probability. This question has been investigated by Sims (1972) in a laboratory simulation of on-inspection task, with equivocal results. Signal detection theory, with its separation of sensitivity and bias parameters, and its specification of an optimal criterion, is clearly an ideal model with which to investigate this problem. The question will be considered again when the detailed programme of experimentation is set out.

2.10 Summary

The theoretical foundations of SDT have been reviewed in detail. The necessary conditions for applying the theory have been set out, and the available inspection studies in which SDT has been used have been considered from the point of view of their adherence to these conditions. Very few studies were seen to have tested the assumptions underlying SDT prior to employing the sensitivity and bias parameters beta and d'. Those studies that have employed SDT more rigorously have suggested that the theory is applicable in this area, particularly in the context of on-going inspection tasks. Further research is seen as necessary in establishing this validity in a wider variety of inspection situations, and also as a tool in investigating a number of important practical problems. In particular, the ability of the inspector to adjust his criterion appropriately, if the incidence of defects changes, is seen as being highly amenable to a SDT approach.

CHAPTER 3 A REVIEW OF THE LITERATURE OF INDUSTRIAL INSPECTION AND RELATED THEORETICAL AREAS

3.0 INTRODUCTION

Producing a comprehensive classification scheme for inspection literature presents a number of difficulties. Inspection is an activity carried out in a very wide range of industries and one which utilizes many differing skills. Some of the taxonomic difficulties present in this area will be discussed subsequently. The literature review which follows will be divided into two broad sections. In order to establish the context for the research objectives of this study, the more important areas in the inspection literature will first be surveyed. Although the emphasis of this review will be on studies which can be directly applied to practical situations, it will also be necessary to consider some of the theoretical areas which underlie the applied studies.

In the second part of the review, the topics which have been selected as part of the experimental investigations will be treated in greater depth. From a detailed consideration of these areas, together with insights gained from the SDT literature reviewed in Chapter 2, the broad outlines of the experimental investigations will emerge.

3.0.1 Some taxonomic considerations

It is not proposed in this study to develop a formal task taxonomy for inspection tasks, although there is certainly a need for such an endeavour. In applied research particularly, as the body of data on human performance grows, it is increasingly necessary to be able to generalize research findings from laboratory studies to operational settings and from one operational setting to another. (Levine et al., 1971). An extensive research programme concerning the problems of behaviour taxonomies has been in progress at the American Institute for

Research, e.g. Fleishman et al. (1970), Miller (1971). Several different lines of approach have been considered in this area. One of the earlier attempts by Miller (1962) was motivated by the desire to obtain data for design decisions in man-machine systems. The main characteristics of Miller's scheme resemble those of the informal model for the inspection process presented in Chapter 1. The behavioural task structure suggested by Miller proposes scan, identification, decision making and effector phases which can be readily equated with the acquisition, decision, identification and action phases proposed in Chapter 1.

Unfortunately, both schemes share another characteristic - they do not provide a very satisfactory means of organizing the available research and applied literature in a particular area. One of the difficulties is that the model is in terms of separate psychological processes, whereas in any real situation the importance of the various hypothesized stages in the inspection procedure is determined very largely by the characteristics of the task. This problem is of course common to any taxonomy. Another problem that occurs when attempting to classify inspection studies according to the scheme set out in Chapter 1, is that many of the important variables are global in nature and could affect performance through a number of the stages postulated. Examples of such variables are individual differences and social factors. It seems then, that although the earlier descriptive scheme is useful in specifying the sequential stages in the inspection task, and the variables which need to be considered at each stage, a different approach is necessary from the standpoint of structuring the literature.

The scheme eventually decided upon follows the task characteristic approach proposed by Farina and Wheaton (1971). Four major sets of variables are seen to determine inspection performance: the

characteristics of the inspection tasks themselves, the physical environment in which the tasks are performed, the organizational and social structure of which the inspection function is a part, and individual, operator centred variables. Of course, although these categories and the variables they comprise are described as independent factors, performance in a real inspection task will be a complicated interaction of these variables. In general, inspection studies have tended to concentrate on single or at the most two interacting variables, although there have been some exceptions, e.g. McFarling (1974), who examined the interactive effects of noise, sex and pacing variables.

Many theoretical areas impinge on the inspection situation, but the emphasis in the first part of the review will be on studies that are either applied, or attempt to simulate at least one aspect of real life inspection tasks. However there will be a preliminary discussion of the major research areas that have implications for a large number of inspection problems. The inspection model described in Chapter 1 will be utilized during the review where appropriate.

3.1 General inspection literature survey

3.1.1 Some relevant theoretical areas

Three theoretical areas of psychology have considerable relevance for inspection tasks. These are decision theory, vigilance and visual search. Decision theory has already been discussed in detail from the orientation of signal detection theory in Chapter 2. Vigilance is an important area because many inspection tasks are largely perceptual in nature and involve prolonged periods of attention with a low probability of signal occurrence. Visual search considerations provide

insights into many of the task characteristics which will be considered in this review. For this reason we will begin with a brief consideration of the latter two theoretical areas.

3.1.1.1 Vigilance and its relevance to inspection

Vigilance has been an important area of psychological research since Mackworth (1950) was able to demonstrate in the laboratory some of the performance decrements occurring during prolonged watchkeeping which had first been observed in radar operators during the war. The apparent relevance of vigilance research to many applied problems gave rise to a voluminous literature. A review by Halcomb and Blackwell (1969) referenced 700 articles. No attempt will be made here to review this literature. Several comprehensive reviews are available, e.g. Davies and Tune (1970), Mackworth (1969 and 1970), Broadbent (1971). Mackworth identified a number of factors which contributed to the performance decline over time, such as a low signal rate, adverse environmental conditions, and unfamiliarity with the work. Later researchers added sleep deprivation, (Wilkinson, 1960) inappropriate signal expectancies (Colquhoun and Baddeley, 1964, 1967) and poor motivation of experimental subjects (Mackworth, 1970).

The similarity between vigilance tasks and inspection tasks is clear. Both involve prolonged attention by the subject for signals (or defects) which may occur infrequently, randomly in time and space and be difficult to readily discriminate.

Many of the findings in vigilance experiments parallel those found in inspection. For example, one of the most consistent results found

in vigilance is the importance of signal rate in determining efficiency. Experiments by Colquhoun and Baddeley (op.cit.) demonstrated the importance of signal rate in influencing the overall level of performance in vigilance tasks. Increases in signal probability produce increases in both the detection rate and the false alarm rate, (Baddeley and Colqhoun, 1969), a finding which could have been predicted from SDT considerations (Chapter 2). Closely analogous results are found in the inspection literature (see later section) and there does seem to be a close affinity between the classical vigilance task and many inspection situations. Many writers, e.g. Poulton (1973), make the assumption that most vigilance data can be readily applied to inspection tasks as long as the subjects have received sufficient practice.

However, other workers have had reservations about the applicability of much vigilance research to real life industrial problems. Kibler (1965) made the following comments when comparing the basic task dynamics of typical vigilance research with those of contemporary monitoring tasks.

1. The weak, brief duration signals typically employed in laboratory vigilance studies are rarely encountered in applied monitoring tasks.

2. The human monitor is typically required to keep watch over multiple information sources, and frequently more than one type of target or information class is the object of his vigil.

3. The signals are often complex and multidimensional rather than the simple, unidimensional events usually employed in laboratory studies.

4. In most monitoring tasks, determining the appropriate response to a

signal event entails a decision process much more complex than those required in vigilance studies.

Elliott (1960) suggests that the classical vigilance decrement has never been observed in any closely simulated radar task and that the social isolation usually found in vigilance studies is not typical of military situations. Smith and Luccaccini (1969) maintain that a vigilance decrement has never been demonstrated in an industrial situation. They suggest that this is due to the greater complexity of the industrial task and that the vigilance decrement can be explained by the lack of motivation of laboratory subjects who are insufficiently aroused to continuously perform what is an essentially meaningless task. Harris (1969) also cautioned in directly applying laboratory results to industrial inspection situations. Belt (1971) attempted to clarify this question by comparing performance by the same subjects on a laboratory vigilance task with an authentically simulated industrial inspection task. He found that the usual vigilance decrement occurred with the laboratory task but that a constant level of performance was maintained with the inspection experiment. The author suggested on the basis of subjects' comments that this was a result of the greater motivation present on the inspection task.

The general conclusion that emerges from these considerations is that we cannot blindly use all the results from vigilance experiments to predict performance in industrial situations, particularly when the laboratory studies have task characteristics unrepresentative of inspection tasks. On the other hand it would be foolish to ignore the considerable body of knowledge that has been gained on human performance in monitoring situations, particularly when results obtained from

vigilance studies are paralleled by data from tasks more representative of the inspection situation. As usual a process of discrimination is necessary when generalizing from research findings to the real world. It is necessary to examine the characteristics of any specific inspection task in detail in order to decide whether or not vigilance research is applicable. The more infrequent the defects, the less arousing the task conditions and the more prolonged the inspection period, the greater the likelihood of vigilance data being applicable, particularly if the defects are simple in nature.

3.1.1.2 Visual search considerations

This topic is surveyed in depth in Bloomfield (1970) and its applications to inspection described in Bloomfield (1975).

During visual search the eye makes a series of fixations and it is assumed that the probability of detection of a defect decreases as its distance from the fixation point increases. This leads to the concept of a visual lobe, which is a hypothetical area about the fixation point within which there is some arbitrary probability, e.g. 0.5, of detection. During search, the saccadic movements of the eyes give rise to a series of overlapping visual lobes which cover the area to be searched. Search is efficient if the area is covered completely with a minimum overlap of the visual lobes. The larger the visual lobe the more efficiently the area to be searched will be covered in the minimum time. The size of the visual lobe will be influenced by a number of factors. Individual differences in peripheral visual acuity, background luminance, length of exposure time, and discriminability of the target are all relevant variables. In addition to the size of the visual lobe, search efficiency will also be influenced by the fixation strategy adopted by

the observer. Both random and systematic sampling will cover the whole area to be searched and will ultimately detect any defect which is discriminable if it falls within the visual lobe. A systematic, regular strategy which optimally covers the entire search area with minimum overlap between the lobe areas will always be quicker on average. Bloomfield (1970) showed however, that even with well practised subjects, their search strategy was better fitted by a random rather than a systematic scanning model. In general the time taken to detect a target consists of two components, the search time itself and the time taken to respond when the target falls within the visual lobe i.e. there is no search involved. With readily discriminable targets this approximates to simple reaction time, but with near threshold targets more complex considerations using SDT or other models may become important (Pike, 1971).

One tends to assume that visual search considerations only become important in situations where large areas have to be scanned for defects. In reality, even if the total area to be searched is relatively small, if the targets, i.e. defects, are similarly small, the visual lobe is effectively reduced in size and so search is still necessary. Bloomfield (1975) considers three categories of visual inspection in which search is important. Where displays contain a number of small items some of which may be defective, the mean search time has been shown to be inversely proportional to the square of the discriminability of the defects. Discriminability can refer to differences in dimensions such as size, area or shape between good and defective items. This is known as a competition search situation. In multipart inspection a single complex object is shown to the inspector and he has to examine many features of the item which may be faulty. An example of this type of inspection is given in Harris (1966) in which ten items of equipment

were rated in terms of complexity, where this was largely in terms of the number of major parts each item contained. A high negative correlation was found between the number of defects found and the rated complexity. In the final type of inspection considered by Bloomfield, the inspection of sheet materials, a fault may be difficult to detect for several reasons. It may fail to emerge perceptually from its immediate background because of patterning effects, it may have a very low contrast difference with respect to the background, or it may simply be very small relative to the total area that has to be inspected.

The efficiency of visual search will be influenced by several factors in addition to those already discussed. Operator centred factors such as experience, eyesight and age will clearly be important as will task characteristics such as presentation rate (Perry, 1968) display size and shape (Baker et al., 1960) and the provision or otherwise of visual aids (Schoonard et al., 1973).

3.1.2 Task characteristics

3.1.2.1 Pacing and movement of the item being inspected

These variables have been considered in a number of studies because of their importance to industrial engineers in determining the maximum throughput of an inspection station possible without degrading the efficiency of defect detection. In terms of the inspection model of Chapter 1, the variables can be regarded as operating at the acquisition phase of the process.

Drury (1973) discusses two major aspects, the effect of the rate of movement of the item being inspected, and the effect of pacing per se.

These two aspects are not necessarily identical. It is well known, e.g. Blackwell (1959) that dynamic visual acuity is inferior to static visual acuity, and Sury (1964) has shown that pacing can degrade performance even if the subject is paced at the same rate as his unpaced performance.

With regard to the first variable, Williams and Borrow (1963) and Eriksen (1964) have shown that the rate of movement does not produce degradation in performance unless the angular velocity exceeds $7°$-$8°$ at the eye.

Drury (1973) considers a number of studies and attempts to deduce a general relationship between the time available per item being inspected and the probability of a correct decision being made. The two types of correct inspection decision concerned are P_1, the probability that a good item is accepted, and P_2, the probability that a faulty item is rejected. The studies considered covered a wide variety of inspection tasks, e.g. Fox (1964) concerning coin inspection, Perry (1968), glass bottle inspection, Sinclair (1971), food products, and unpublished studies on the inspection of sheet glass. In all of these studies the general effect is observed that as more time is allowed to inspect each item, the probability of rejecting a faulty item increases whilst the probability of accepting a good item decreases. Drury accounts for this finding by postulating that the inspection process consists of two phases, a visual search process until a potential defect is found or time runs out followed by a decision process. Thus for very short time intervals, corresponding to a high rate of pacing, relatively few of the potential faults are seen. This leads to a low value of P_2, the correct rejection probability, and also a high value of P_1, the correct acceptance probability, because the opportunity for making an error in

the other direction, i.e. false alarms, is lessened. At very long time intervals, or low pacing rates, search considerations become unimportant and the limiting values of P_1 and P_2 are largely a function of the SDT variables considered in Chapter 2, i.e. the intrinsic sensitivity of the inspector and his criterion.

A paper by Buck (1975) presents a highly detailed analysis of what he refers to as dynamic visual inspection, DVI. DVI occurs in a conveyor belt situation where the item being inspected moves past the inspection station at various speeds. Some of the factors affecting DVI can be identified as the direction of movement, the speed of the conveyor, the object interspacing distance and variability, and the lateral variability of the object on the belt. Buck discusses in detail the effect of various viewing constraints such as the 'viewing window' past which the items on the conveyor flow. DVI involves first visually tracking the moving inspection item. Crawford (1960) shows that up to an angular velocity of about 25^{o}-30^{o}/second the eyes can acquire a moving object in a single saccade. At greater speeds, additional eye movements are required and hence belt velocities of this order mean that the observer will spend more time in making visual corrections and therefore less time will be available for visual search within the object.

A number of studies, e.g. Ludvigh and Miller (1958) show that a complex set of factors affect dynamic visual acuity, e.g. the angular velocity of the item, and the exposure time available.

The belt velocity in DVI therefore plays a complex role both through its affects on visual acuity and the exposure time available for visual search (this variable will be considered in more detail subsequently).

Recent data from a number of studies (Rizzi et al., (1974), Nelson and Barany (1969), Smith and Barany (1971), Purswell et al., (1972) and Lion et al., (1975)) suggest that the following generalizations can be made about pacing and the more general area of dynamic visual inspection. DVI is improved with (a) increased exposure time for an object, (b) lower belt velocities, (c) greater interspacing between successive objects. Self paced inspection appears to be superior to externally paced work, (Williges and Streeter (1971), McFarling and Heimstra (1975)). Individual differences in peripheral visual acuity influence efficiency of DVI. The random or ordered arrangement of the items on the conveyor belt is another factor found to be important. The fact that the latter two variables are also important in static visual search, Bloomfield (1970) lends weight to the idea that DVI consists of detecting the oncoming item, visually acquiring it, searching it for faults and then making an accept reject decision, Buck (1975). Cochran et al., (1973) present a model for predicting the combined effects on DVI of change in visual angle, angular velocity, time to view, illumination and contrast.

3.1.2.2 Magnification, lighting and other aids to enhance defect discriminability

Techniques such as X-rays, ultrasonics, gamma rays and dye penetrants are all used in non-destructive testing in industry to render visible defects which could not be detected with the unaided senses. Often the resulting display or trace constitutes an inspection problem in itself, where the operator may have to continuously monitor the output for some subtle change which indicates a defect. These problems have been considered in Embrey (1975) (ultrasonic testing) and Lusted (1971) (X-ray photographs). The most common methods of enhancing defect

discriminability are lighting and magnification.

Lion (1964) compared the effects of fluorescent with tungsten lighting on a simulated inspection task involving grading ball bearings for size, and found that fluorescent lighting produced a significantly higher rate of work without any concomitant increase in errors. Lion et al., (1968) repeated the above comparison in a conveyor belt situation. The items used in one experiment were plastic discs with a link pattern drawn on their faces. 'Defective' items had a break in the link. The other experiment used plastic buttons as the test items, some of which had off-centre holes. Detection efficiency was higher for the link inspection task under the fluorescent lighting than with the tungsten lighting. However, there was no significant difference between lighting conditions with the button inspection task. The authors accounted for this latter result by suggesting that the link task was primarily a test of visual acuity whereas the button task involved more perceptual elements. It was proposed that the differential effect of the different light sources was due to the fact that the tungsten lamps, being point sources, were more readily obscured by the subject at the workbench, producing an effectively lower lighting level. This would readily account for the results in view of the relationship between visual acuity and ambient illumination. It seems strange that the authors did not verify this hypothesis by adjusting the different types of lighting to provide equal illumination with the subject in situ. The question of the very different spectral composition of the different light sources would seem to be another uncontrolled variable.

Further insights into this area are given by Sakguchi and Nagai (1973). In a comparison between various types of lighting in a Landolt ring

recognition task, it was found that eye fatigue, as defined by subjective reports, seemed to result from the narrow bandwidth of coloured lighting sources such as sodium lamps. This variable clearly needs to be considered when lighting for prolonged inspection periods is being specified.

Lighting considerations for inspection are far broader than merely specifying the optimum intensity and type of lighting to be used. Faulkner and Murphy (1973) illustrate a number of ingenious ways in which special purpose lighting can be used to enhance the discriminability of defects. They point out that the simple expedient of increasing lighting levels is not necessarily the most effective way of increasing task performance. They describe a number of lighting techniques including grazing illumination, polarized light, spotlighting, dark field illumination etc. which can be used in various situations. Case studies from the glass industry in which lighting is used to enhance defect discriminability are given in Gillies (1975).

Inspection using microscopes and other magnification aids has become increasingly important with the growth of the microminiature integrated circuit industry. Smith and Adams (1971) and Smith (1975) propose that over a wide range of conditions, optimum performance can be expected when the visual angle subtended by the defect, as a result of magnification, is between 9 and 12 minutes of arc. Froot and Dunkel (1975) point out, however, that a number of other parameters of the microscope system need to be taken into account, including resolving power, aperture and depth of field. They cite a case study where two groups of inspectors could not agree over the incidence of defects in a batch of products. It transpired that although they were using microscopes of identical magnification, they had differing resolving power for defects.

A detailed case study involving the use of magnification in the inspection of rubber seals is given in Astley and Fox (1975).

3.1.2.3 Complexity

The variable of complexity is an extremely difficult one to quantify. Firstly there is the question of defining an adequate index of complexity. In inspection studies complexity has usually been described in terms of the number of items which would be potentially defective on each unit inspected. For example Harris (1966) found that the index of complexity assigned to circuit modules by a panel of experienced judges, correlated highly with the number of major parts making up the item. It seems likely however, that variables such as the arrangement of the parts constituting the item are also important, either because of the Gestalt consideration suggested by Fox (op. cit.) or through the facilitation or otherwise of an efficient search strategy. At least two other aspects of complexity can be considered, the complexity of the defect itself and the complexity of the background in which it is embedded. The conspicuity of the defect could well be regarded as the degree to which it shares common attributes with the field within which it is embedded. One might expect this variable to be partly dependent on the number of these shared attributes which occur within the defect and its background. Possibly a quantitative estimate of detectability as a function of this concept of complexity could be derived.

The clearest effect of complexity on inspection efficiency comes from the Harris study (op. cit.). Inspector performance in terms of defects detected showed a strong negative correlation with the complexity of

the module. A DVI task used by Purswell et al. showed a similar relationship, where subjects had to remove grids containing geometrical patterns from a conveyor belt. McFarling and Heimstra (1975) used printed circuit boards containing varying amounts of circuitry as constituting differing complexity levels. Decision time was found to increase with increasing circuit complexity whilst defect detection performance declined. Another variable investigated was pacing, and it was found that as circuit complexity increased, there were larger increases in decision time for self paced subjects than for their machine paced counterparts.

It seems therefore, that increases in complexity can be regarded as generally reducing defect detection probability. This can partly be accounted for by visual search considerations, because in paced situations there will be many occasions when the inspector will not have had time to examine each potentially defective attribute of the item in the viewing period allowed. The persistence of the effect even in self paced situations suggests however that other, perceptual variables may also be important. Studies of image interpreters, e.g. Powers et al., (1973) have shown that in multiple defect situations, interpreters often have a far lower detection efficiency for subsequent defects after an initial one has been found.

3.1.2.4 Display organization

This variable is another which one might expect to be influenced both by visual search and perceptual considerations. Williges and Streeter (1971) found no significant differences in performance as measured by correct defect detections and false alarms between an ordered and a random display consisting of 600 transparent discs containing

occasional pin hole defects. This result is in conflict with visual search studies and with other inspection studies which have considered the same variable. The authors suggest that this may have been due to the close proximity of the inspection items in both the ordered and random arrangement. In fact many of the subjects after completion of the experiment stated that they had been unaware that the discs had been displayed in both a random and ordered fashion.

In a laboratory visual search situation Bloomfield (1970) found that an irregular display took longer to search than a regular one. One would therefore expect this variable to be of importance in paced situations, and in fact in the Williges and Streeter study a significant interaction was obtained between paced and unpaced presentation and the ordering or otherwise of the items. Detection performance was significantly better with the regular display under the paced condition than the irregular display.

The effects of display arrangement are not necessarily due to visual search considerations alone. An industrial study by Fox (1964) considered the effects of the random or regular arrangement of coins on a conveyor belt at the Royal Mint. The regular display proved considerably superior to the random arrangement and Fox proposed that the result could be explained in terms of Gestalt theory, i.e. the defective coins emerged more readily from the regular, 'good gestalt' background than from the random, and therefore difficult to perceive arrangement. An alternative explanation is simply that the search time was longer in the irregular display case. Since the inspection was paced, a longer search time would produce a lower rate of detection of defects independent of any effects on the intrinsic perceptibility of the defects. It

is not possible to decide which of these explanations is appropriate from the Fox paper. Indeed it is difficult to see how the two variables could be experimentally disentangled. Nevertheless, the possibility of effects at a perceptual level in addition to the peripheral consideration of visual search cannot be ruled out. Scott Blair and Coppen (1942) suggested that 'learnt gestalten' seemed to develop in certain skilled operators and Thomas (1962) quotes several industrial examples where inspectors seemed to detect defects on a wholistic, figure/ground basis, rather than by a process of search. Eye movement studies in situations of this type would presumably clarify the issue as to whether search was taking place.

Lion et al., (1975) compared performance on a single line conveyor belt system with a three line system. The test items consisted of plastic discs which contained either a broken link pattern (defects) or a complete pattern. Detection performance was superior on the three belt system. However, in order to provide the same amount of material per unit time to the inspector, the single belt was run at 18cm per sec., i.e. three times the speed of the three line belt. Since this exceeds the limits proposed by Williams and Borow (1963) (i.e. 2cm per sec.) at which performance decrement occurs, it seems likely that the obtained results were due to this variable rather than the organization of the display.

3.1.2.5 Signal rate

Signal rate is an important variable which has received considerable attention in both vigilance and inspection research. Early work in vigilance tasks, Jenkins (1959), Kappauf and Powe (1959), suggested that lower signal rates produced reduced detections by causing an increased vigilance

decrement. Colquhoun and Baddeley (1964, 1967) showed however that this was at least partly an artefact due to the subjects having an inappropriately high expectancy of the signal rate. It was further shown by Colquhoun (1961) that it was the conditional probability of a signal given an event occurred, that determined detection efficiency rather than the actual frequency of signals in time. In general it has been shown that an increase in signal probability produces an increase in both the detection rate and the false alarm rate, e.g. Baddeley and Colquhoun (1969), as would be predicted from the SDT considerations discussed in Chapter 2.

Rather few inspection studies have explicitly varied defect rate or probability as an explicit experimental variable. Fox and Haslegrave's study (1969) had the virtue that it was conducted in an industrial environment. They attempted to verify Colquhoun's finding of the importance of signal probability as a determinant of defect detection efficiency as opposed to stimulus frequency. They investigated both a static and a conveyor paced situation where screws were inspected for a variety of faults. No significant differences in correct detection probability were found in the paced condition, but the static condition showed the expected straight line relationship between detection probability and defect probability. It is not clear why the effect was not observed in the paced condition, but it seems likely that other variables such as the conveyor speed (52 feet/minute) swamped the probability effect.

It is clear however that data from experiments of this type have to be interpreted with care. A well known study by Harris (1968) employed four different defect rates on a scanning type inspection task using

four groups of inspectors. Using as measure of inspection accuracy the proportion of defective items that were detected, Harris stated that the accuracy of inspection declined with decreasing defect probabilities. Baker and Schuck (1975), however, took the Harris data, and using the equal variance SDT model, calculated that d' was not significantly different for any of the probability conditions. They therefore stated that Harris was in error and that inspector accuracy did not change as a result of signal probability. This is a good illustration of the confusion that ensues when differing measures of inspector performance are used. In reality both author s were correct within the descriptive terminology they were employing. If anything, Harris's conclusions were to be preferred to those of Baker and Schuck because the latter authors did not bother to check the appropriateness of the equal variance assumptions implicit in their use of the SDT model. The results are of course explicable in terms of the inspectors employing a criterion appropriate to the ongoing defect probability.

The use of artificial signals in an inspection task to increase the apparent defect rate and thereby enhance the detection rate (and false alarm rate) has been proposed by Wilkinson (1964). Although his 'inspection' task is actually a laboratory vigilance task, there seems to be no reason why the technique should not be applicable in a real life situation as long as the artificial 'defects' could be readily separated from the real ones and the situation was such that false alarms were not 'expensive'.

The only dissonant study on the general conclusion that an increased defect rate enhances detections is one by Sosnowy (1967). This study, which was a simulated inspection of ball bearings, only showed the

usual relationship at a high rate of pacing, i.e. 240 items per minute. Unfortunately the original study could not be obtained by the reviewer (it is reported in Badalamente and Ayoub, (1969)) and hence the index of inspector accuracy used and the precise experimental conditions could not be ascertained.

One study exists in which the variable discussed in Chapter 2, a within session change in the defect rate is considered (Sims (1972)). The study used printed circuit board inspection as a simulated industrial task, and showed that although inspectors generally accurately perceived the quality level of the incoming product, there was considerable inter-subject variability in their ability to adjust to changes in defect rate.

The remaining studies in which defect rate is an important but not explicit variable have been discussed in the SDT review of Chapter 2.

3.1.2.6 Number of inspectors

Schlegel et al., (1973) compared the performance of a single versus a dual inspector system in a Landolt ring, conveyor based inspection task. The task used an inspection period of 45 minutes in order to investigate the effects of the two systems on a possible vigilance decrement. In terms of probability of detecting a defect, the two inspector system was significantly superior to the single inspector. The false alarm probability showed no significant difference. These results were in line with what would be predicted from the simple statistical combination of the two inspectors efficiences. The two inspector system had the additional bonus that there was significantly less vigilance decrement in performance. The reasons for this are not obvious, since the

inspectors were separated by screens and could not monitor one anothers performance. It seems possible that the element of competition in the two inspector case provided an increased level of arousal and hence reduced any vigilance decrement. This study has obvious implications for real life inspection systems. It should be noted, however, that the stimuli used were considerably easier than those normally found. A slightly different arrangement was used by Lion (1975). In an experiment described in detail earlier (Section 3.1.2.5), the performance of inspectors on a three line arrangement of items on a conveyor belt was compared with the performance of two inspectors seated on opposite sides of a six line conveyor. Performance in the two inspector arrangement was significantly better than using a single conveyor both in terms of correct detections and fewer false alarms. The authors attributed the result to the stimulus of doing a job in unison, the feeling of competition and the reduction in boredom due to the opportunity to talk. Since the length of each session was only 12 minutes, the first two factors seem more likely to be important than the last.

A study by Morrissette et al., (1975) showed a similar improvement in performance in team monitoring using a laboratory task. The evidence suggested that performance was improved by social facilitation, i.e. the monitors performed better when working together than separated.

The evidence strongly suggests therefore that inspection efficiency will be improved by operator redundancy. In practice, such redundancy does occur in industrial situations when particularly critical products are being inspected.

In using more than one inspector, a decision has to be made whether the

economic advantages accruing from the higher detection efficiency expected exceeds the cost of additional inspectors.

3.1.2.7 Repeated inspection

Batches of product are sometimes inspected repeatedly in order to raise the probability of defect detection. This may be done by the same inspectors or an independent inspection may be utilized. This latter arrangement is clearly preferable since re-inspection by the same inspector means that the same defects are likely to be missed particularly if there are systematic errors in inspection strategy. Belbin (1957) gives an example of the repeated inspection of ball bearings, although he does not state whether or not it was independent. After the first inspection 63% of the defects were found, and a further 16% were found after the second pass through the system. Harris and Chaney (1969) p.78 describe the repeated inspection by ten independent inspectors of an electronic module. Performance in detecting critical defects continued to improve (at a slightly increasing rate) with the addition of independent inspections up to a total of six, after which performance levelled off. Performance in detecting non-critical defects, on the other hand, continued to improve as additional inspections were added. These results tend to confirm that different inspectors do not necessarily find the same defects. Eilon (1961) presents an operational research type model which specifies situations, depending on the cost of inspection, when it is economically worthwhile to recycle products through an inspection station. The question of the efficiency of independent repeated inspections versus a team inspection approach is a complicated one. One way of looking at the situation is to assume that an interacting inspection team combines the advantages of social

facilitation noted in the last section with the theoretically higher probability of detection due to the possible 'blind spots' of the individual inspectors being eliminated by the overlap of abilities. On the other hand it is possible that the team consensus as to what constitutes an acceptable product may be incorrect. The effect of this could mean that the findings of more accurate inspectors within the groups could be rejected, or not reported, due to group pressures. Further work is clearly needed in this area.

3.1.3 Environmental factors

Traditionally, these factors include heat, lighting, noise and workplace design. Lighting has already been considered in an earlier section. There have been no studies in which heat has been considered as an explicit variable. The evidence from vigilance tasks (J.T. Mackworth (1969), p.167) suggests that whereas cold may slow reactions and interfere with detections particularly at the beginning of a session, heat tends to increase missed signals at the end of a prolonged vigil. These results may be extended to inspection tasks with the caution suggested earlier.

McFarling (1974) considered performance under noise conditions in a simulated printed circuit board inspection task which investigated the effects of a number of interacting variables. Inspector performance in a quiet condition was significantly better, measured by defect detection probability, than under a 90dB white noise condition. False alarm scores were not significantly different in each case. This latter result is unexpected since in vigilance tasks the decrement in detection performance under noise conditions is generally a result of an increase in beta (Broadbent 1970). The result obtained suggests a

lower sensitivity in noise. The author was unable to account for the effects but J. Mackworth (1969) suggests that similar results in vigilance experiments are due to distraction effects. A very similar study by Ehlers (1972) also considered the effects of noise on the inspection of printed circuit boards. Three white noise levels, 50, 70 and 90 dB were employed, and defect detection performance was significantly worse in the 90 dB condition. A deterioration in performance over the 70 minute inspection task occurred under the 50 dB noise condition but not under the 70 or 90 dB condition. False alarms were not analysed because only 10 in the 19,000 good circuits inspected were called bad. These results suggest that in common with other monitoring tasks noise levels of 90 dB and above should be avoided in inspection situations, although a moderate level of noise serves to reduce vigilance effects, presumably via its arousing qualities.

Workplace design is clearly a variable which needs to be considered in the design of inspection stations, where a worker may be carrying out a visually demanding job for long periods. Astley and Fox (1975) present a case study which shows how anthropometric considerations are taken into account in this situation.

3.1.4 Organizational factors

Organizational factors are important determinants of the effectiveness of any real inspection system. If there is disharmony within the inspection team, or conflict between production and quality control, then even the most perfectly designed inspection system will either fail to function effectively, or its findings will go unheeded. We will first consider the more general socially orientated factors that

influence the effectiveness of inspection and then some of the specific organizational factors which directly influence inspection performance.

3.1.4.1 Managerial and social aspects

McKenzie and Pugh (1957) consider the effects of the relationship between the production and inspection departments in industry. Where lack of communication exists, production departments will generally be critical of attempts by the quality control section to assess their work. Social pressures by their workmates will often persuade inspectors to modify their judgements even when there has been no objective change in quality. The authors comment on the very high degree of individual variation in inspectors and even inconsistency within their own judgements. The general point made is that it is this inconsistency that leads to the deterioration of relationships between inspection and production departments. Recommendations are made to regularly calibrate inspectors with reference standards.

Belbin (1957) discusses the issue of the differences between the customer's quality standards and those adopted by inspection departments. He presents a real life example which shows that many of the complaints from customers of a particular firm were due to defects for which the inspector had not been told to look for. Similarly many of the faults for which inspectors did reject items were of no consequence to the customer. The effect of the continuing complaints from the customer was to make the inspector reject more and more products for the wrong reasons (presumably via a change in criterion). Belbin points out that the quality standards required may fluctuate due to variations in demand, and suggests a scheme for defining quality levels so that

inspection criteria can be readily modified. The effects of social pressures on inspectors is illustrated by an example from a hosiery factory where inspectors had to feed back repairable defects to their workmates for mending. The defects could be classified as rejects or mendable, the latter producing a much higher rate of pay for the menders. A combination of social pressure from the operatives on the inspectors, and a lack of clearly defined quality standards meant that virtually all of the defects were classified as being mendable.

McKenzie (1958) regards inspection accuracy as being determined by basic individual abilities, environment and formal organization, and interpersonal and social relations. One of the most important organizational factors is the provision of reference standards for the inspector. Thomas (1962) emphasizes the importance of clear, unambiguous examples of both rejects and perfect product being provided because of the tendency of perceptual judgements to drift with time. Equally important is the provision of inspection instruction. Raphael (1942) describes how some viewers inspecting fabric were rejecting 53 per cent of the product whilst others in the same group were rejecting only 13 per cent. It transpired that the specification allowed a tolerance of 3 millimetres but some of the viewers had not been informed of this. The question of the drift of perceptual standards and the possibility of ameliorating this by a weekly 'calibration meeting' is discussed in McKenzie (op. cit.). The same paper cites further examples of the effects of social factors on inspection standards. A group of two operators and an inspector worked as an isolated group apart from occasional visits by a supervisor. There was obviously a tendency for the groups to identify itself as a cohesive whole and hence the inspector could not be expected to make decisions that might disrupt

the group, unless they were based on unambiguous evidence. Mitchell (1935) provides evidence of poor social relationships between operator and inspector biasing the latter's inspection standards, as does Roethlisberger and Dickson (1939).

McKenzie (op. cit.) points out that inspection sections tend to have a tightly knit social structure, partly because of their small numbers and partly because of their control function vis a vis production, which tends to lead to strained relations between the two functions. A survey conducted amongst patrol inspectors suggested that they felt that they were viewed unfavourably by production operatives.

Thomas and Seaborne (1961) criticize many experimental studies of inspection because they remove the individual task from its socio-technical context and examine it purely in terms of psychophysical performance. They point out that the laboratory study lacks much of the concomitant information which serves as a frame of reference in the industrial task. Often the industrial inspector ulitizes sources of information, such as a knowledge of the supplier, which enables him to use an appropriate criterion, in the SDT sense. Of course if such information is unreliable, the inspector's accuracy may be reduced. There is very little opportunity for inspectors to develop consistent standards in a situation where the range of quality of the input items varies widely and where there is little feedback as to the quality of the final product required. Again, in SDT terms, such feedback is known to be necessary to develop a stable criterion. In many real inspection systems the inspector is required to continuously modify his standards to take into account factors such as market conditions, level of output and demand. Because of the emphasis on inspection

studies in critical areas where quality specifications are clearly defined, this variable quality aspect of much 'bread and butter' inspection has tended to be neglected. In an analysis of a particular inspection task, Thomas and Seaborne (op.cit.), showed that the inspector's function could be regarded as satisfying the sometimes conflicting needs of the sales organization, the production department and the raw material purchasing department. The inspector utilized his knowledge of the manufacturing process to inform production operatives of deterioration in the process. His inspection of raw materials provided information influencing his judgement of the final product. Finally he was frequently approached by the sales manager, and on the basis of information on the state of the market would raise or lower his standards with regard to certain faults.

This analysis suggests that in real life situations the sources of information utilized by the inspector in arriving at his accept/reject decision are far more complicated than in the laboratory situation. He receives feedback from sales and from any check inspection that may be carried out. He receives feedforward information about the state of the raw materials and from his knowledge of the manufacturing process, and finally he utilizes the sensory data present in the actual item being inspected.

This situation can readily be analysed in SDT terms. We can regard the training and any reference standards that are provided as influencing primarily the sensitivity of an inspector, his ability to recognize the cues which indicate good or bad product. The instructions which apply to an inspection task at a specific time will provide the base data by which the inspector sets his criterion level, i.e. decides on what standards of acceptability shall apply at that time. This information

is modified by feedback from sales as to the importance of particular defects. This can be regarded as changing the inspector's implicit pay-off matrix. Feedforward from production modifies the criterion by affecting the subjective probability of a defect occurring. We can see that in real inspection systems, the ability of the inspector to modify his criterion on the basis of additional information, is, as pointed out in Chapter 2, an important quality.

3.1.4.2 Motivational variables

As Weiner (1975) points out, it is surprising that there has been little scientific investigation of the effect of motivational factors on inspection, particularly in view of the popularity of this approach in industry. Many propaganda-style exercises such as the 'zero defects' programme (Swain 1972) have been tried and varying degrees of success reported. Unfortunately the quality control journals which report such research do not employ indices of performance which are sufficiently precise to make unambiguous conclusions possible.

The use of financial incentives is the obvious way to influence motivation, and has no doubt been employed in many companies for this purpose. There is, however, little hard evidence as to its efficacy or otherwise. Ergonomists working in inspection have tended to reject financial incentives in this area. However as Weiner (op. cit.) points out these assumptions may not be true for other forms of incentive such as knowledge of results. Mitten (1957) describes how female roller bearing inspectors were allowed to go home as soon as they had achieved their inspection quota. Although this incentive seemed to be effective in this case, social disharmony could result in inspection systems such as

that investigated by Embrey (1970) in which there were a wide range of ages and abilities.

Vigilance studies have not shown financial incentives to be very effective in maintaining performance, possibly because the financial rewards offered are unrealistically small (Wiener 1969). In many SDT studies attempts have been made to influence performance by manipulating the payoff matrix. As shown in Chapter 2, this has not proved very successful, largely because of the difficulty of relating subjective utilities to external rewards. Perhaps this is again a case of insufficient quantities of money being offered. It is possible that financial rewards may influence performance through indirect means such as modifying an inspector's visual strategy. Bloomfield (1970) describes experiments in which he produced extremely large increases in visual search performance by offering monetary rewards.

The use of Knowledge of Results (KR) in training for perceptual skills will be considered in detail in a subsequent section. It has been suggested by some workers that KR exerts a motivational effect quite distinct from its informational content. For example in vigilance tasks, where there is very little informational content, KR is known to enhance performance. Even though KR may be difficult to provide in a real inspection task, the evidence suggests that the benefits of KR extend to sessions where it is withdrawn, Wiener (1963), Annett (1966). Drury and Addison's (1973) industrial inspection study clearly demonstrated the enhancement of performance due to KR, but it was still difficult to say whether this was due to its informational content, motivational effect, or a combination of both. Despite Annett's (1969) position that motivation is an unnecessary construct in explaining the

effects of KR, the evidence overall seems to suggest that at least some motivational effect must be present whenever information feedback is given.

3.1.5 Individual factors

In virtually all studies of inspector proficiency the largest contribution to the variance of the results is individual differences between the inspectors. In view of this, it is surprising that so little work has been done on identifying the source of this variability. Usually individual differences are regarded as a 'nuisance variable' which experimenters seek to eliminate from their designs. A detailed consideration of the nature of the individual differences which affect inspection performance would seem to be a most effective way of enhancing performance in inspection systems where this task and environmental variables have already been optimized. The major ways in which individual differences affect the performance of an inspection system are through selection and training. These will be considered in detail in subsequent sections.

3.1.5.1 Selection

As implied by the inspection model in Chapter 2, two major groups of variables will affect a person's ability to perform inspection, peripheral factors such as eyesight which affect the acquisition of the necessary sense data, and cognitive factors which determine how the sense data will be interpreted. Although differences in cognitive skills can be reduced by training, it would not be surprising if there were intrinsic differences in certain cognitive abilities necessary for

inspection which selection procedures could identify. As we shall see, up to now, research on selection methods appropriate to inspection has proceeded in a very ad hoc manner. Very little attempt has been made to analyse the non-intellectual skills which may be necessary for inspection. Nearly all of the studies which exist on selection for inspection have attempted to correlate some performance index with a more or less arbitrary group of standard industrial tests.

In the sections which follow, we will first consider the individual variables that more directly affect the target acquisition phase and then consider the work on the intellectual and cognitive factors that affect inspection performance.

3.1.5.1.1 Visual abilities

It is clear that the visual skills of an inspector are an important determinant of his overall efficiency. Depending on the nature of the task, static or dynamic visual acuity will be important. Ayers (1942) and Tanalski (1956) both found strong relationships between various static visual measures and indices of inspection performance. Nelson and Barany (1969) have described a dynamic visual acuity selection test suitable for conveyor based inspection. Standard visual tests have been developed for static visual skills required for various industrial tasks including inspection. The best known of these, the Orthorater, is described in Trimby (1959). In situations where colour is an important cue in identifying defects one of the standard colour blindness tests needs to be incorporated in the routine testing schedule.

Virsu (1972) presents a sophisticated review of the visual factors affecting the inspection of radiographs. The use of the Modulation

Transfer Function (MTF) (Cornsweet 1970) approach to measuring visual performance is proposed as being far more efficient than traditional measures of visual acuity. He proposes that in the case of radiographs, the translation of monochrome photographs into coloured slides using MTF techniques would optimize the visual performance of the interpreter.

3.1.5.1.2 Age

The effects of ageing are to produce a gradual loss in sensory capabilities such as visual acuity. This is however compensated for by the perceptual skills acquired through long experience. Few studies are available in which age and inspection accuracy have been considered. Sheehan and Drury (1971) and Drury and Sheehan (1969) report a gradual decline in d' with age of about 0.2 units per ten years of age. Since their results were based on only five inspectors however, they may not be readily generalizable. Jamieson (1966) showed increasing performance with age in electronics inspectors, Jacobsen (1953) found accuracy increased up to age 34 and then declined to age 55. Evans (1951) found no effect at all.

The evidence from vigilance tasks shows only slight effects of ageing, unless the stimulus presentation rate is high, e.g. Thompson et al., (1963). This is consistent with the general finding that short term memory capacity declines with age. This will clearly have a deleterions effect on rapid conveyor paced inspection, where the older inspector may have difficulty in retaining the information from the display in his short term memory before the next item has to be inspected.

In general it seems likely that older persons will make good inspectors as long as they are employed in tasks in which their perceptual skills

are utilized but which do not place heavy demands on their sensory or information processing capabilities.

3.1.5.1.3 Personality variables

Virtually no work has been done on the use of personality tests on selecting for inspectors. Colquhoun (1959, 1960) used the Heron personality inventory to investigate time of day effects on detection performance in extreme personality groups but the task was a laboratory experiment. In vigilance tasks of this type, the Eysenck Personality Inventory (EPI) has been extensively employed to investigate the importance of the extraversion-intraversion personality variable in influencing performance. Although it is often stated that the intrinsically more highly self-aroused (according to Eysenck's theory) intraverts do better at vigilance tasks, the results are somewhat more equivocal (Mackworth 1969). A number of other studies using personality variables in vigilance experiments, have found very few significant correlations with performance. In spite of the lack of success in utilizing these variables in vigilance tasks, there does seem to be a case for investigating the use of some of the more easily administered tests, such as the EPI, in an inspection context, particularly if prolonged monitoring is involved.

3.1.5.1.4 Sex

Although women are more often found in the inspection departments of manufacturing industry than in most other sections, there is no solid evidence that there are substantial differences between the sexes in ability for inspection work. Only one inspection study, McFarland (1972),

seems to have included sex as a major variable, and the only difference found was in the greater variability of response times for women. Of the vigilance studies that have considered this variable, six found no significant differences at all (Waag et al., (1973), R. Smith et al., (1966), Gale et al., (1972), Kappauf et al., (1955), Kirk and Hecht (1963), and McCann (1969)), in two men performed better (Neal and Pearson (1966), Heimstra et al., (1967)) and in three others there was no significant main effect of sex but significant interactions with other variables (Bakan and Manley (1963), Krkovic and Sverko (1967), Whittenberg and Ross (1953)). There is clearly insufficient evidence for inspectors to be selected purely on the basis of sex.

3.1.5.1.5 Selection tests

A number of testing procedures have been applied to the selection of inspectors with generally disappointing results. Many early studies suffered from the drawback of correlating supervisors ratings, rather than objective measures of performance, against test scores. Wiener (op. cit.) points out that such ratings are probably based on the supervisor's perceptions of earnestness and co-operation and that the correlations between these variables and actual inspection performance are unknown.

Link (1920) obtained a correlation of 0.50 between output rate and tests of card sorting number cancelling and number group checking, with munition inspectors. Wyatt and Langdon (1932) were unable to obtain significant correlations using four standard industrial tests and eight inspection tasks. Sartain (1945) obtained a high multiple correlation, $R = 0.79$, between standard industrial tests and supervisor ratings.

Only low correlations were obtained using similar techniques, by Schuman (1945) and Tiffin and Rogers (1941).

One test specifically designed for inspection work exists, the Harris Inspection Test (HIT), which is a short pencil and paper instrument. Harris (1964) found significant correlations between the HIT and three out of four electronic inspection tasks. When the test was administered as part of a battery to 26 machined parts inspectors, however, no significant correlation was obtained between test scores and job sample measures of performance (Harris and Chaney 1966). By combining together the two most valid measures in the battery, the number comparison section of the Minnesota Clerical, and the Industrial Mathematics test a multiple correlation of R = 0.75 was obtained. No further validations of the HIT have appeared in the literature.

It is clear that there has been no systematic attempt to isolate the underlying individual factors which are important in inspection and to incorporate these variables in selection procedures. It is not surprising, therefore, that attempts to correlate performance with arbitrarily selected standard tests have been unsuccessful. The fact that the HIT, although successful in the application for which it was originally designed, did not predict performance in another type of inspection, suggests that it did not measure any general underlying ability.

We can conclude from the survey of the selection procedures currently in use in the inspection area, that a more fundamental approach is needed. Such an approach, based on the cognitive skills underlying inspection, will be set out in the theoretical section of this review (part 3.2.1).

3.1.6 Training for inspection

Training for inspection is a neglected area in industry. Many inspector training schemes described in the quality control literature (e.g. Browne 1965) have as their aim the imparting of the background information judged necessary to perform inspection. However there is virtually no emphasis at all on training for the perceptual skills necessary to detect defects and to recognize acceptable products. When training schemes which purport to accomplish this latter aim are described, the emphasis is often on the large savings that the scheme is alleged to have produced rather than the details of the technique and the method employed for evaluating it.

Learning effects in laboratory simulations of inspection tasks have been noted without specific training being given, e.g. Smith and Adams (1971) and Lion et al., (1968). In an industrial setting Chaney and Teel (1967, 1969) have employed a variety of techniques in training machined parts inspectors and photomask inspectors. Their most commonly employed technique, known as job sample instruction, essentially involves giving inspectors knowledge of results (KR) after inspecting test items containing typical defects.

KR (Knowledge of Results) has been employed in training for perceptual skills in a number of applied studies. Martinek (1965) reports a study in training photo-interpreters. Photo-interpretation, or image interpretation as it is alternatively known, consists of identifying military targets, which may be camouflaged, on aerial reconnaisance photographs. This task bears a close resemblance to an industrial inspection task. Martinek found that providing the photointerpreters with an error key

which analysed the commonly occurring errors of previous interpretations, produced significantly fewer errors of commission than a key in which the characteristic features of the various target types were set out. The error key can be regarded as a form of group KR, although it is incomplete. As there was no significant difference in the number of targets detected, this suggests that the result was due to a change in sensitivity rather than response bias. Another image interpretation study, Cockrell and Sadacca (1971), showed that KR was more effective in a team context when team members first scanned a photograph independently and then discussed the results together immediately afterwards. The greatest gains in proficiency were made by the least able members of the team and although detection performance was improved, significantly, there were greater gains in reducing the number of misidentifications and false alarms.

Another image identification study, Powers et al., (1973) attempted to improve performance by modifying the interpreter's scanning strategies rather than by KR techniques. Structured search practice increased the number of target detections at the expense of a greater number of false alarms. A 'speed reading' training technique, designed to reduce fixation times and expand the visual field, succeeded in halving the search time without changing the accuracy of the interpreter. Training using the 'error key' approach discussed earlier significantly reduced false alarms. Brock et al., (1974) provided a complete training programme for non-destructive testing radiograph examiners. The programme consisted of tape/slide presentations which gave examples of various types of defect, and then a KR phase where radiographs containing defects were presented and the student was given full KR after he had attempted to identify them. A self-administered test then determined if

the student should go on to the next module or re-take previous modules. There was a gradual increase in difficulty of the radiographs both within and between modules. The programme was highly successful in that it reduced the time to train to the required criterion from an average of 80 hours to 10.9 hours. Wallis (1963) describes the use of KR in the training of a complex perceptual task associated with a weapons system. Although no details were given of the classified system, substantial reductions in training time were found.

In view of the importance of training as a relatively low cost means of increasing inspection accuracy, it is surprising that no studies appear to have been performed to specifically investigate training techniques for inspection. It is felt that training is one of the areas which requires experimental investigation using tasks which are more representative of industrial inspection situations. In section (3.2.2) some of the theoretical considerations which will provide guidelines for research in this area will be considered in depth.

3.1.7 Conclusions regarding the general literature of inspection

It seems clear that inspection performance in a given situation will be the result of a complex interaction between some of the factors considered in the review up to this point. Although there is a paucity of research in a number of areas, two in particular will be considered in further detail in the theoretical review which will constitute the next part of this chapter. The first of these will be the implications for selection of some of the cognitive variables which may account for individual differences in inspection skill. The remainder of the

review will consider in detail some of the work that has been performed in the area of training for perceptual skills.

3.2 Theoretical literature survey

3.2.1 Cognitive variables in selection

An analysis of the cognitive skills necessary for the performance of inspection tasks provides useful guidelines as to possible new approaches to the problem of selection for these tasks.

In Chapter 2 the decision making aspects of inspection were emphasized in the context of SDT. It was implicitly suggested that decision making skills were learnt rather than innate. Although there may well be innate differences in the ability of individuals to utilize evidence concerning the existence of a defect to the full, it is not clear how one would devise selection procedures to identify such individuals, apart from simulation exercises.

Other aspects of the cognitive skills utilized in inspection seem to hold more promise. In an earlier section, when considering the visual search aspects of inspection, it was pointed out that in very many inspection tasks, the critical defect in an item is difficult to detect because it does not emerge perceptually from the background in which it is embedded. In fact this is by far the commonest situation. The defect virtually always occurs in noise. The noise may be an external random perturbation which degrades the detectability of a defect by obscuring or modifying the critical cues such as contours, angles or shade gradients which define the defect. Alternatively we can consider

a higher level 'cognitive noise' which results because the background within which the defect is embedded shares so many attributes of the defect that it is difficult to separate them perceptually. On this basis, if there are inherent individual differences in ability to separate a wanted configuration from the background in which it is embedded, this should provide a basis on which to select inspectors.

In fact a very considerable body of research exists on the ability of subjects to 'disembed' stimuli from their backgrounds. This dimension of individual differences is concerned with the field dependence - independence continuum. Field dependent individuals show great difficulty in breaking up an organized visual field in order to keep a part of it separate from that field. Field independent subjects on the other hand are readily able to extract a wanted configuration from the confusing field in which it may be embedded. These contrasting styles of functioning are found to be extremely stable with time and to affect performance in a wide variety of tasks. The work in this area is associated with Witkin and his collaborators (Witkin (1950), Witkin et al., (1962), Witkin et al., (1954)). Witkin developed an effective pencil and paper test for discriminating between field dependant and independant individuals, the Embedded Figure Test (EFT). This test has been validated in a wide variety of situations (see Witkin et al., (1971) for a review) and has demonstrated, for example, consistent sex related differences on the field independence dimension.

In view of the relevance of this aspect of individual functioning to many applied perceptual tasks it is surprising that so few studues have considered this variable. There are in fact very few studies applicable to the inspection area. Thornton et al., (1968) used an

identification task in which subjects had to locate and identify a series of small buildings in an aerial photograph. There was a highly significant correlation (0.72) between scores on the EFT and target detection performance in terms of number of targets detected during the time limits for the test.

Seale (1972) however, found no significant relationship between EFT scores in an aerial target acquisition simulation using aircrew as subjects. However, examination of the data suggested that the subjects were already a highly selected field independent group (airline pilots were used as subjects) and hence the test was likely to be too insensitive to differentiate between individuals. This may not detract from its general value as a selection instrument however.

There seems to be a good case for further work to be carried out using the field independence variable.

Other cognitive factors also deserve attention as potential dimensions upon which to base selection procedures. Field dependence is related to, but distinct from the dimension of distractability. It was originally hypothesized that field independence could be interpreted as the ability to resist distraction rather than to overcome the effects of the embedding context (Witkin et al., 1962). Subsequent work by Karp (1963) using factor analytic techniques suggested that distractability was a separate factor from field independence, but that the two factors were moderately correlated. Karp (op. cit.) differentiated between distraction and embeddedness as follows. In the distraction situations, the figural properties of critical items remain intact. In the embeddedness situation, the critical item or its parts are organized into new,

competing gestalts which serve to break up the original figure. A distracting context may be thought of as obscuring a critical item without changing the nature of the item, whereas an embedding context serves to obscure a critical item because it changes the nature of the item. Both types of field occur in inspection situations. Karp (1962) produced a number of tests designed to measure the characteristic of distractability. These will be described in detail in a later chapter.

Sack and Rice (1974) consider attention to have a directional aspect which can be analysed into at least three processes: degree of selectivity, resistance to distraction and shifting. Gardner and Moriarity (1968) discuss a factor termed 'field articulation', an index of a subjects ability to attend selectively to cues. Although the Witkin approach does not interpret field independence as an attentional phenomenon there is clearly a considerable overlap between the two concepts. Sack and Rice (op. cit.) consider distraction to be an involuntary change in an established attentional focus. Whichever interpretation of distraction is 'correct' there is no doubt as to the importance of this variable in inspection tasks. In many situations, particularly where the inspected item may only appear for a short time, as in paced tasks, if the inspector is too readily distracted by extraneous stimuli, he may miss defects. Smith and Barany (1970) present evidence that such non-observing does in fact take place in inspection tasks: the 'shifting' variable considered by Sack and Rice (op. cit.) is related to the ability to change one's attentional focus at will. It is clearly important to be able to readily change an attentional focus, as for example when one must lay aside one task and attend to another. The inspection of discrete items requires a readiness to shift attention from item to item as they are presented. This

variable also seems worth considering as for a possible means of selecting inspectors.

In summary it is apparent that the cognitive variables considered constituting an extremely promising area of research and one which will be pursued further in subsequent sections of this study.

3.2.2 Theoretical approaches to perceptual learning

Virtually all of the applied studies reviewed in section 3.1.6 have employed some form of KR as the basic training paradigm. This is partly because of the success of the KR approach in motor learning and the corresponding assumption that it is the most efficacious approach for perceptual skill training. There is still however some debate whether there may be other, equally efficient methods of training for these skills.

Most of the work that has been done has not used tasks which resemble the typical inspection situation. However, it is useful to consider in detail the approaches of two particular workers who have made substantial contributions to this field. These are Wiener, who has advocated the use of KR in these tasks, and Annett who has investigated cuing techniques in some depth.

It is useful at this point to define in more detail the two approaches. Knowledge of results (KR) has been defined as 'knowledge which an individual or group receives relating to the outcome of a response or group of responses' (Annett 1961). Forms of KR encountered in the detection context include immediate feedback as to whether a response was

a correct detection or false alarm or the provision of all or part of this information in the form of summaries.

KR is the classical learning paradigm and is said to exert its effects by reinforcing an S-R link, by reinforcing observing responses and by maintaining alertness via its motivational effect. The Law of Effect suggests that KR, contingent on a learners response, serves to reinforce the association between stimulus and response. The subject must make a response before the information can be obtained, and a 'corrected guessing' technique is used for training.

Annett, influenced by Gibson's (1953) suggestion that perceptual learning is a distinct type of learning requiring its own formulation, and not necessarily analogous to motor learning, has made a study in depth of the technique of cuing. Cuing has been defined (Annett 1959) as the provision of stimulus information before or during a response such that the response is made more effective or more likely to occur than without such information. Annett suggests that an overt response is only necessary as a means of acquiring information otherwise impossible to obtain, e.g. the 'feel' of a control. In identification tasks the stimulus and its name can be presented together and the learner does not necessarily have to make a response to the first in order to receive the second. Annett asserts that learning in perceptual tasks can take place via a simple association principle rather than a reinforcement paradigm. In his view the repeated pairing of the stimulus pattern to be learnt and its name, leads to the building up of a template against which a stimulus to be identified is compared. This theoretical position was part of Annett's general orientation that the facilitation learning by the provision of KR was not through its reinforcement or motivating

properties, but through the information conveyed to the learner (Annett 1969). Annett has presented an impressive array of evidence to support his position, although, as we shall see, his later work was more equivocal as to the superiority of cuing as compared with KR in promoting perceptual learning.

3.2.2.1 Cuing in perceptual skills training

Annett (1966) quotes an experiment in which subjects had to estimate the number of dots present in a tachistoscopically exposed field. Both cuing and KR training was superior to a simple practice control group. In a Landolt ring experiment in the same study, subjects were cued by placing the targets and non targets on different coloured backgrounds. The other conditions were a control and an 'easy' training condition in which larger gaps in the Landolt rings were used. The training effect was significantly greater for the cuing method. This suggests that simply making the task easier does not promote learning. Cuing makes the task easier without changing the critical dimensions of the stimulus. The final experiment in this series utilized a simulated sonar task in which 1000 Hz tone bursts were superimposed on a 50dB background of white noise. Four training methods were compared:

a) Cuing - a warning light was turned on half a second before each signal.

b) KR - as above, but the warning occurred after the signal, usually before the subject had responded.

c) Summary KR - a summary of hits, misses and false alarms was provided at five minute intervals.

d) Easy material - the background was reduced by 5dB so that most signals were readily detectable.

The cuing group was significantly superior to the others in enhancing performance. Annett concluded that these results were consistent with the view that perceptual learning takes place when a stimulus is unambiguously paired with its name or designation. He proposes that perceptual learning can be regarded as a simple example of paired associate learning. The failure of KR methods in some cases of training for auditory detection (Campbell (1964), Swets et al., 1964) is cited as further support for his case.

To further elucidate the relationship between cuing and KR methods of training, Annett and Clarkson (1964) conducted experiments using the same experimental set up as in the last study described. Five training groups were employed:

a) 100% cuing. A yellow light flashed half a second before each signal.

b) Retrospective cuing (non-contingent KR). A blue light flashed 2 seconds after each signal.

c) KR (contingent). Correct responses were followed immediately by a green light, incorrect responses by a red light.

d) Partial feedback cuing. The first signal was cued. Subsequently, if a signal was missed, the next signal was cued, if detected the next was uncued. False positives had no effect.

e) Partial feedback cuing and contingent KR - conditions for groups (c) and (d) combined.

According to the hypothesis that the most effective training procedure would provide the subject with as many authentic examples of the signal as possible and its distribution over the training period, it was expected that conditions (a) to (c) would produce the same rank order of effectiveness. Condition (d) was added in case cuing was effective,

but that the learner became dependent on it. It was intended as a form of conditional cue removal. Condition (e) was added when no training effect was found with condition (d). It was felt that this was because (d) gave no information on the signal distribution in time and hence partial KR was provided to enable the subject to obtain some knowledge of this through his responses.

Groups (a), (b) and (c) showed learning effects in the predicted order. Group (a) has maximum exposure to the signal plus full information about its frequency and distribution in time. Group (b) has fewer signal samples available but full information on signal frequency and distribution. In group (c) there are about the same number of signal samples available as in (b), but less distribution information. Group (e) with a proportion of authentic signals plus the opportunity to gain incomplete distributional information through contingent KR, is better than (b) or (c) but inferior to (a) which is again consistent with the hypothesis being proposed. The failure of group (d) to produce any training effect at all could be accounted for by the low proportion of signals cued and hence the reduced opportunity to accumulate the necessary experience. The cue withdrawal was initiated before any learning could be established. A notable feature of the results was that the improvements in detections found with KR were accompanied by an increase in false alarms, which was not the case with the cuing training. This is consistent with the SDT interpretation that cuing improves sensitivity but KR promotes a criterion change. In terms of the degree of vigilance decrement found, the cuing group showed a greater resistance to the decrement than the KR group. This was interpreted in terms of the expectancy theory of vigilance (Baker 1963) such that the subjects with the most accurate knowledge of the signal

distribution would show the smallest decrement.

Later experiments (Annett and Paterson 1966) using a similar experimental task provided further insights into the differing roles of cuing and KR in perceptual training. It was found that giving subjects information on the distribution of signals by flashing a warning light when a signal would have appeared, without actually presenting the signal, produced performance increases comparable to those obtained through cuing. This suggests that what is learned during cuing is not the nature of the signal itself, but some appreciation of its distribution. It should be noted however that in this task the signal, a 1800 Hz tone, had very few characteristics that one could learn. The results also suggested that the apparently lax criterion found in free response situations was at least partly due to an attempt by the subject to gain more information on the signal characteristics. With a fixed interval condition and complete KR there is much less difference in the style of performance between KR and cuing, although KR still induces a slightly lower criterion. In the third phase of Annett's study (Annett and Paterson 1967) subjects were trained in three attributes of sonar operation, i.e. pitch discrimination, intensity discrimination and duration discrimination. Cuing, KR, and a combination cuing/KR conditions were employed. All three methods were effective in training pitch and intensity discrimination but none for duration discrimination. There were no significant differences between the various training conditions. It was hypothesized that this was because a technique specifically designed to eliminate the effects of change in response bias was employed (2 alternative forced choice) and that the only difference between the various techniques is in terms of the change in response strategy produced. The most recent of Annett's work, Annett (1971), treats

cuing and KR as being essentially equivalent. This is in line with his general orientation that perceptual learning takes place purely by the pairing together in time of the stumulus and its name. If the experimental conditions are such that the subject does not have to make extra responses in order to obtain samples of the stimulus during training, then it does not make very much difference if the information is presented before or after the stimulus.

Annett's work has been discussed in some detail because of the comprehensiveness of his experimentation in this area. The issues arising from this work and possible further research that it might generate are discussed subsequently.

3.2.2.2 Knowledge of results in perceptual training

Wiener and his associates have produced a number of studies investigating the use of KR and other techniques mainly on visual vigilance type tasks. Wiener's early experiments (1963) agreed with Annett's findings that the provision of KR contingent on the subjects' responses (i.e. correct detections and false alarms) increased detections at the expense of a higher false alarm rate. Improvements due to full KR persisted after the KR was removed. A later study (Wiener 1968a) showed that detections continued to improve after 5 sessions of KR training, but that the training effect did not persist in a follow-up five weeks after the training. There was a significant increase in false alarms for the KR group over the first two sessions but not for the rest of the training, suggesting that a genuine increase in sensitivity was occurring. In these experiments complete KR was being given and hence there was no necessity for the subject to produce a high rate of responding to gain information.

An earlier paper (Wiener 1967) had shown that a group trained with KR on a visual meter monitoring task showed a significant improvement, compared with a control group, when transferred to a different type of visual monitoring task. There was no change in false alarms. This result tends to disconfirm the Annett hypothesis that perceptual learning consists of building up a 'template' of signal characteristics, since this would suggest that training would not readily transfer to tasks with different signal types. The results can, however, be accounted for quite readily by Annett's other proposal, that learning the signal distribution is almost as important as learning its characteristics. In Wiener's experiment the signal distribution (in terms of intersignal interval) was virtually the same for both tasks. Two other explanations were put forward by Wiener. One is that KR increases the subject's motivation to perform at a high level, and that this carries over to the transfer session, and the other that KR enables the subject to gain a general skill at maintaining vigilance in some unspecified way. In view of the fact that a vigilance decrement occurred in the transfer session, the latter two explanations seem less likely than Annett's proposal. In a direct test comparing cuing, KR, KR+cuing and a control, significant training effects were obtained with the KR, and KR+cuing groups but not with the cuing only group (Wiener and Attwood 1968). There was no significant change in false alarms on transfer with the groups receiving KR, but the cuing group made significantly fewer.

In SDT terms the results could be accounted for by the more cautious criterion induced by cuing, indicated by the lower false alarm rate in the transfer session. It is known (Broadbent and Gregory 1963) that during a vigilance task an increasing stringency of criterion is observed. This, combined with the already high criterion induced by

the cuing, could account for the observed absence of training effect in this experiment. An excessively high criterion would of course depress both false alarms and correct detections. The criterion effect would presumably exert a greater effect on detections than any learning associated with cuing, because of the simple nature of the signal in this experiment. A similar lack of efficacy of cuing and superiority of KR was reported for an auditory task in Annett and Paterson (1966). It was suggested in this case that the protracted nature of the training may have lead the subjects to become bored with the passive cuing training. However, the results can equally well be accounted for by the SDT hypothesis proposed earlier.

3.2.3 Some conclusions on KR versus cuing

We have seen from the analysis of Annett's and Wiener's work that their basic orientation to the question of training for perceptual skills is different, in spite of the fact that more recently Annett has been prepared to allow that cuing and KR may be largely equivalent. Annett's basic assumption is that perceptual learning is based on a simple association principle, and that learning will take place via a simple contiguity in time of a stimulus and its name. It follows therefore that both KR and cuing are important from the standpoint of the information they provide as to the characteristics of the stimulus and its distribution in time. Wiener's position is that the Law of Effect is the appropriate training paradigm and that it is primarily via the motivational effect of KR that the subject learns to maintain a higher level of arousal in the transfer session. Any knowledge of the signal characteristics that the subject learns through cuing or KR is by way of a bonus. Wiener also suggests that there is a danger that subjects may

become 'cue-dependent' and be unable to transfer any training gained to subsequent sessions.

The differing standpoints and perhaps the results of the workers in this field can be seen to be at least partly due to the differing types of task that they are interested in. Annett has not been concerned with vigilance phenomena as such, and although the signals he has employed have generally been fairly simple, the long term aim seems to have been towards the accurate identification of complex, near threshold stimuli, rather than reducing the vigilance decrement. The tasks employed by Wiener have, on the other hand, been largely visual, above threshold and extremely simple. The meter monitoring task was in fact chosen because it was known to produce a vigilance decrement. It is not surprising therefore that Wiener has been more interested in the motivational effects of KR. Although he has not stated as such, Wiener's orientation can be regarded as being concerned with enhancing signal detection through arousal mechanisms mediated via motivational variables. In this way, performance can be regarded as being improved both through an improved signal detection ability and a greater resistance to vigilance decrement. Annett would probably see the main gaol of training as improved signal detection ability, and any resistance to vigilance decrement as a bonus to be gained through a more accurate apprehension of the signal distribution. Lau (1966) and Wiener and Attwood (op. cit.) suggest that cuing may be more effective as the perceptual complexity of the task increases, whilst Annett and Paterson (1966) propose that KR may be judiciously mixed with cuing to provide a more interesting and therefore effective training regime for subjects.

In spite of the Wiener and Attwood experiment which showed no particular advantage for mixed KR and cuing training, several studies e.g. Weiz and McElroy (1964), Swets et al., (1962) and Swets et al., (1964) have shown this form of training to be advantageous.

It is clear that in spite of the considerable amount of research that has been expended on the question of KR versus cuing in perceptual training, the issue is still very much open. A number of omissions in the research to date can be identified. The most important of these is the absence of signals representative of real life tasks. In the context of this thesis, it would be of interest to apply some of the findings discussed in this section to the complex stimuli encountered in inspection tasks. Another important development would be to attempt to apply SDT in a more rigid way to the issues discussed up to now. Many of the changes in performance would be clarified by the application of SDT methods to isolate changes in sensitivity from those of response bias. This work would provide a useful link with the SDT approach to inspection developed in Chapters 1 and 2. A particular area of interest is the question of the development of a knowledge of the distribution of defects using KR or cuing techniques. The ability of an inspector to be sensitive to a change in defect distribution is important, since as discussed in Chapter 2, he can only employ an optimal criterion if his subjective estimate of the defect density is in accord with reality. It would be interesting to investigate if any of the techniques discussed up to now would develop any general ability to recognize a change in defect density. Presumably any technique which enhanced sensitivity would provide a larger sample of defects from which the inspector could more easily infer the defect distribution in time. The use of SDT would seem to be the most effective way to investigate this problem. These

possibilities would seem to be potentially fruitful areas for further research.

3.2.4 Other factors in perceptual training

Although the approaches discussed up to now represent the most consistent and prolonged studies of training for perceptual skills, a number of other models of perceptual learning exist and a variety of other variables can be seen as important in this area.

Wallis (1963) presented an interesting model of the perceptual learning process. Learning to identify complex patterns is seen as a blend of analytic and synthetic processes. Initially the trainee analyses the pattern to be detected into cues and features such as lines and angles in a visual stimulus. Eventually a process of synthesis takes place and perception occurs as an wholistic process, a Gestalt that is detected in its entirety. In his description of a training technique, Wallis stresses the demonstration of the relevant cues embedded in the complex whole by techniques such as drawing attention to one cue at a time and using training materials in which they are readily visible. As training proceeds, this analytic approach is gradually modified until an overall synthesis takes place. The main method of guidance utilized in this process is KR, and emphasis is placed on the importance of using realistic materials in training. It is pointed out that augmented feedback i.e. KR and cues, must be withdrawn at an appropriate time so that trainees do not come to depend on it for successful performance.

The problem of trainees becoming dependent on cues and being unable to successfully transfer from training has been pointed out earlier. A

similar difficulty could occur in KR situations under certain circumstances. Abrams and Cook (1971) propose that identification skills involve the development of internal references. If KR is provided continuously, the learner will utilize it to sustain performance at the expense of learning. Their experiments indicate that fading KR throughout the training session enhances the retention of identification skills. It is suggested that the removal of KR by fading creates the need for learning, and the continued, but reducing provision of KR the necessary information. Another finding which confirms the ideas of Wallis is that learning is enhanced by a gradual increase in the stimulus complexity during the training programme. Caution may be necessary in applying these results to inspection training however, since the stimuli were complex auditory signals.

Another issue arising from Wallis' analysis of perceptual learning is the relative efficiency of analytic compared with synthetic training techniques, otherwise known as part and whole methods. Annett (1971) reviewed a number of studies considering this variable. He conducted experiments comparing a large number of different analytic and synthetic methods. The overall result was that simple whole methods are as effective as any of the more complex part methods which attempt to draw attention to identifying features of complex stimuli.

3.3 Directions for research

The specific research proposals which emerge from the literature reviews of this and the preceding chapter can be seen to have three main themes. The first of these is the utilization of SDT in a sophisticated form in the analysis of real life and realistically simulated inspection tasks.

Secondly it is proposed to use SDT as a tool in the investigation of training techniques for the perceptual skills important in inspection. Finally the important but neglected area of selection will be considered from the standpoint of the cognitive skills necessary to perform inspection.

Prior to these more theoretically orientated areas, a description and analysis of the real life inspection systems which will form study vehicles for this thesis will be given and the results of on-line experimentation in an industrial context described.

CHAPTER 4 CASE STUDY I : THE INSPECTION OF BUBBLE CHAMBER PHOTOGRAPHS

4.0 INTRODUCTION

In this chapter a case study will be presented in which many of the theoretical topics discussed in the review chapter will be examined from the point of view of their applicability in a real inspection situation. The results of this case study and that considered in the next chapter will provide a useful orientation for the theoretical experimental work considered later.

Although the analysis of bubble chamber photographs may seem to be a somewhat specialized area, it will be demonstrated in this chapter that this task is directly comparable to inspection tasks found in industry.

The inspection system considered, which is in the Physics Department, University of Birmingham, has been described in a previous report (Embrey 1970). The work set out in this thesis is, however, previously unpublished.

4.1 General considerations

High energy nuclear physics research is carried out in many centres throughout the world. One of the most important research activities that these groups perform is the investigation of the structure of matter using large and very expensive particle accelerators such as the machine at Cern in Geneva. Beams of high energy particles produced by such machines are fired into devices known as bubble chambers, which consist of large containers filled with liquefied gas. The particles ionize the gas to give distinctive configurations of tracks, some of which may have been produced by new particles created as a result of

collisions between the incident particle beam and the gas atoms. It is the detection and analysis of these patterns which constitutes the bulk of the research effort in this field. The volume of data is such that it is not possible for experienced physicists to examine all data produced. Each time a beam of particles is fired into the bubble chamber, automatic cameras photograph the resulting tracks, giving rise to perhaps a million photographs from a particular experiment, each of which needs to be examined.

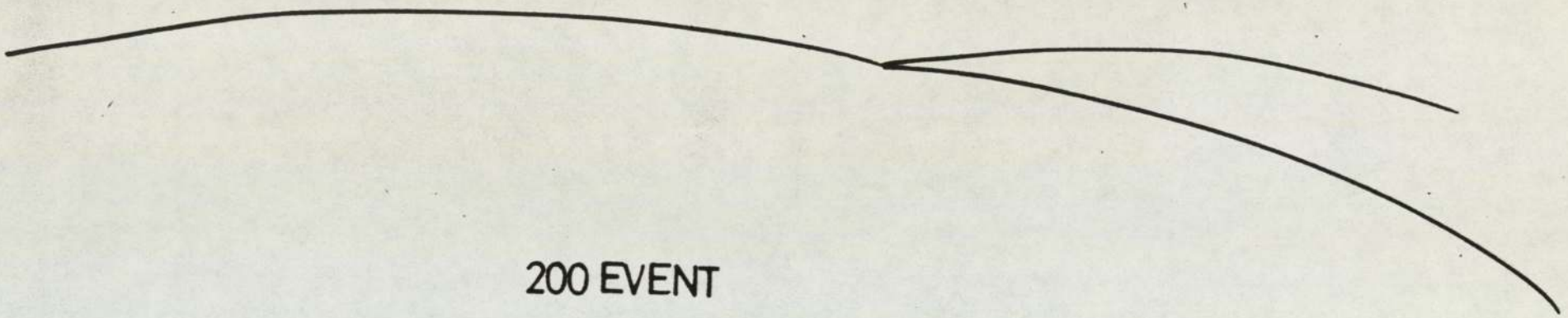
In order to cope with this inspection problem, data analysis groups have been set up in which the films are scanned and particular configurations of tracks, known as events, are detected and categorized.

The importance of the film examiner (or 'scanner') is not to be underrated. High energy nuclear physics research consumes a significant proportion of the funds available for scientific research in this country. In spite of the considerable expenditure on hardware it is salutory to note that the inspection efficiency of the unaided human operator is a vital link in the chain of analysis. For this reason the film scanning task constitutes a useful and valid area of study in its own right in addition to its implications for industrial inspection.

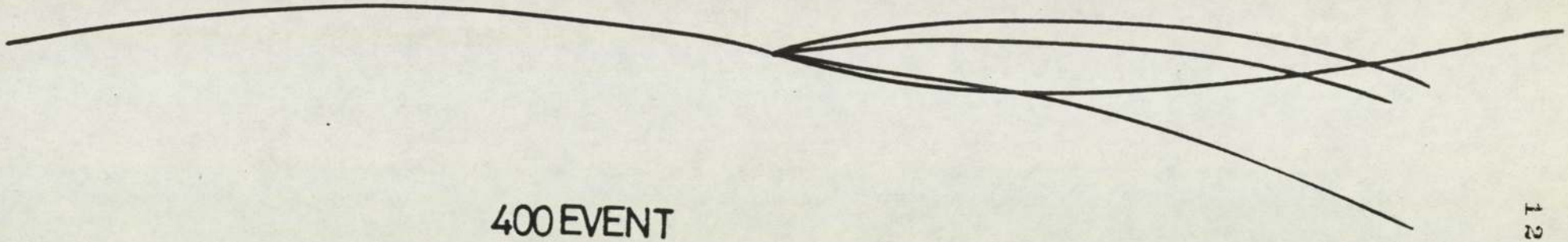
4.2 The scanning task

The items to be examined consist of photographs of bubble chamber events taken from three different angles, giving rise to three separate rolls of film, referred to as views 1, 2 and 3. Each roll consists of 750 frames. The appearance of a typical film frame can be seen from the enlarged photographs shown in Figure 4.2. The series of parallel curved

FIGURE 4.I DIAGRAMMATIC REPRESENTATION OF BUBBLE CHAMBER EVENTS.



200 EVENT



400 EVENT

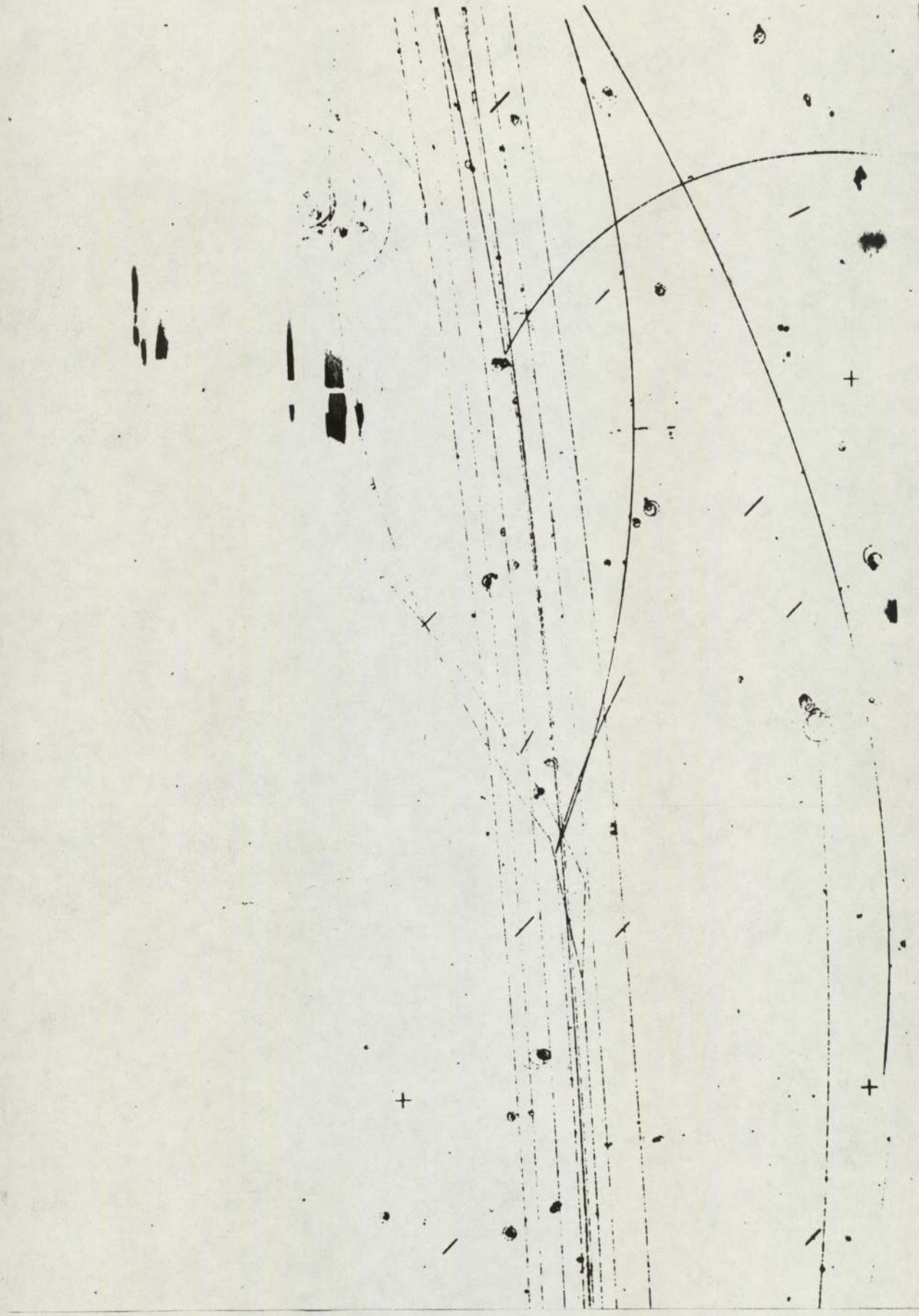


FIGURE 4.2 BUBBLE CHAMBER EVENT.

tracks entering at the bottom of the frame are known as beam tracks. As can be seen from the figures, most of the beam tracks pass through the bubble chamber without interacting. In the centre of the frame a typical 'event' can be seen where a number of prongs emanate from a 'production vertex'. The configuration of tracks which make up an event, known as its 'topology' can be described by means of a three number code. For the purpose of this study, it will be necessary to consider only two classes of event topology, those in which two or four prongs emanate from the production vertex, known as 200's and 400's respectively.

It is configurations of tracks similar to those illustrated which are of special interest to physicists. The task of the scanner consists of examining each film frame for meaningful patterns, in accordance with pre-assigned criteria. In many cases the scanners will be looking for a variety of different types of event on each frame although in some cases only one configuration is searched for. The film frame often contains a large number of unwanted patterns as can be seen in Figure 4.2. These tend to obscure the wanted configurations and can be regarded as visual 'noise'. The similarity with industrial inspection will be apparent.

The 'Shiva' scanning machines used to examine the film are illustrated in Figure 4.3. They consist of two scanning tables mounted side by side. Three rolls of film are loaded into a film transport mechanism at the side of the machine and the film images are then projected via an over-head mirror system on to the scanning table surface. The optical system contained in the scanning machine produces a magnification of 30 times to give a projected image of the same size as the scanning table i.e. 2.5 by 0.9 metres. The movement of the three views is controlled by the three switches to the left of the table surface. These enable the

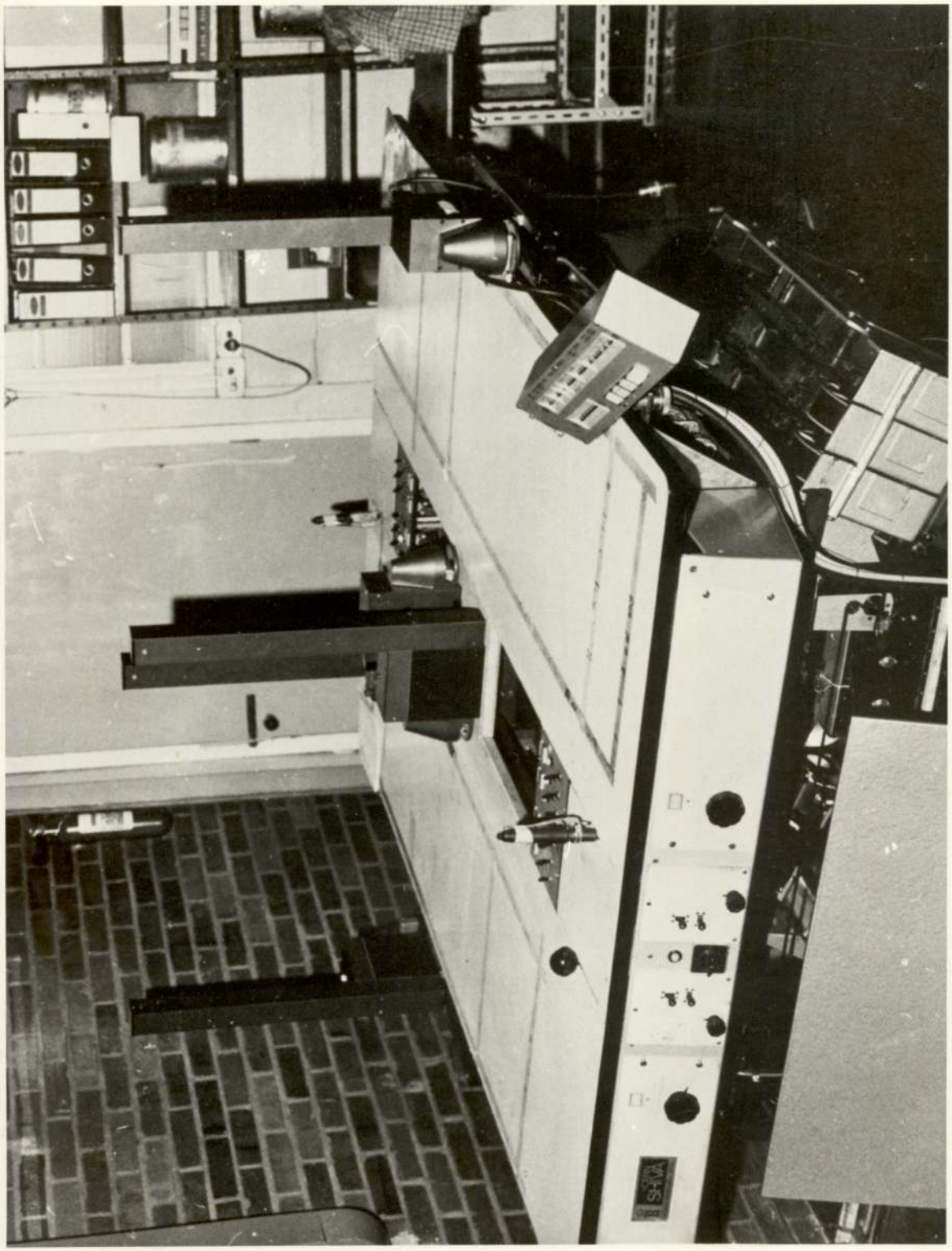


FIGURE 4.3 SCANNING TABLES

operator to advance any of the film views independently and to superimpose views if necessary. The whole optical system can be moved by the handle which can be seen to the left of the operator's position in Figure 4.2. This enables any part of the projected image to be moved close to the scanner for more detailed examination. Most scanners use this system extensively to scan the various parts of the display in which a suspected event lies.

4.3 Detailed task description

The scanner is required to examine every frame of the film specified, using at least two film views, and to find and record all the events on the film that satisfy the criteria detailed in the scanning instructions.

There is no detailed procedure laid down for carrying out the actual operations of scanning but the following is typical. The scanner advances view 1 of a particular film frame so that it is in a standard position on the scanning table. This is done by operating the view 1 film advance switch until a fiducial mark on the film coincides with a mark drawn on the scanning table. The operator then scans the projected image for wanted events, using the film transport handle to move parts of the field closer to the end of the table as necessary. Having scanned view 1, the scanner then changes to view 2, advances the film to the same frame as view 1 and then checks his findings from the first scan. This procedure is necessary because the appearance of an event can differ substantially from view to view, since each view is a two dimensional representation of a three dimensional space. Complex events are usually resolved unambiguously by use of the third view.

Having decided upon an interpretation, the scanner then feeds the information into a keyboard attached to the scanning table, from which it is transferred to on-line computer storage. Finally the scanner switches back to view 1, advances to the next frame and then begins the cycle of operations again. Each film is scanned twice, and lists compiled of any differences between the two scans. Finally the film is scanned a third time by a 'fine scanner', a more highly trained scanner who utilizes the comparison list and his own judgement to resolve ambiguities between the two previous scans. The 'fine scanner' is the final arbiter who decides which events are recorded.

4.4 Analysis of scanning as an inspection task

In the analysis of the data analysis group referred to earlier (Embrey op.cit.), a number of the variables important in inspection systems, e.g. selection, training and environmental aspects were discussed. The detailed consideration of these and other variables in the earlier review chapters enables a more complete analysis of the system to be produced.

4.5 Theoretical areas relevant to film scanning

4.5.1 Signal detection theory

SDT is clearly applicable to the scanning task in that it involves distinguishing the signal, in the form of the wanted event, from the 'noise' of the non-signal patterns in which it is embedded. It is of interest to consider if the equal or unequal variance model is likely to be appropriate in this task. If a population of experienced

scanners is being considered, they might be expected to know the characteristics of both signal and noise equally well and therefore the equal variance model would be more likely to apply than in laboratory based studies (Taylor op.cit.).

4.5.1.1 Factors affecting the criterion

Another consequence of a well practised subject population is that the scanners would be likely to be relatively homogeneous with respect to the criterion they adopt. This would be partly on the basis of experience per se, in that over a long period of time, there is considerable opportunity for feedback, informal or otherwise, to stabilize the criterion. Additionally the implied payoffs for correctly detecting an event or making a false alarm do not change substantially over time. The fact that all the events discovered by a scanner are subsequently checked by a fine scanner encourages the use of a lax, low criterion. On the other hand the low average probability of an event occurring (about 0.2 for the events considered in this study) would tend to raise the criterion. The way in which the criterion is affected by the combination of a priori signal probabilities and payoffs of the various types of division possible, was set out in chapter 2, page **27** . The probabilities of the various types of event that are scanned for is approximately constant over time, which also tends to give rise to a relatively constant overall degree of bias. Predicting the numerical value of beta is difficult because we do not have any quantitative estimates of the payoffs involved. However, by using the value of beta obtained experimentally, it should be possible to determine the subjective utilities employed by the scanners, assuming that the equal variance assumption model is appropriate. Alternatively, by assuming

some values for the costs and values of false alarms and correct detections, it would be possible to establish if the scanners were employing the theoretically optimum criterion. These questions will be considered during the experimental analysis.

4.5.1.2 Sensitivity considerations

In view of the relatively homogeneous criterion predicted in the previous section, performance variability would be primarily due to differences in sensitivity, either because of differences in sensory skills such as visual acuity, or intrinsic differences in the detectability of different event topologies. A fairly high overall value of d' would be expected, since most events are readily detectable by experienced scanners, given sufficient time to examine the film frame in detail.

There is a distribution of event discriminability across films which is a function of the contrast level, the number of beam tracks and the presence of extraneous tracks which are confusable with events. Within a particular film, different events of the same topology will vary in discriminability depending on, for example, whether the production vertex is obscured by an overlapping track, or the acuteness of the angles between the event 'prongs' and the beam track. Event prongs which have a very narrow angle are easily confused with the beam track. The number of 'awkward' events on a film is relatively small, but they give rise to protracted response latencies, whilst the scanner makes a very careful examination of the frame in order to resolve the ambiguity.

4.5.2 Visual search

It would be expected that much of the visual search data would apply to this task, as there is clearly a considerable search involved to cover the large area of the scanning table. However, the search is certainly not free search over the whole area, since the nature of the task constrains the scanner's attention to specific areas of the projected image. Events can only be produced from a beam track and these are concentrated in the centre of the frame. Typically the scanner will look along the curved, parallel beam tracks to see if their symmetry is broken by the oblique prongs of an event. A more general search would be carried out after the main event had been discovered, to ascertain if there were any other configurations associated with the main event.

In view of the inhomogeneous nature of the scan required, it seems unlikely that search time could be readily predicted from the techniques described by Bloomfield (1970). Because the task is self-paced, search time considerations are of less importance than in a conveyor belt situation. However, the overall length of time to completely scan a film will obviously be influenced by search variables, and so the overall throughput of the system is partly a function of this variable.

4.5.3 Vigilance aspects

It is difficult to predict a priori whether or not vigilance effects will significantly affect performance of the task under consideration. In terms of Kibler's comments cited in the last chapter (Kibler 1965), the signals encountered in scanning are more complex and (usually) of greater frequency than those found in laboratory vigilance experiments.

Also the fact that usually many categories of event are scanned for simultaneously, indicates that a more complex decision process is necessary than in vigilance experiments. Against these factors must be set the fact that many scanners subjectively find the task very boring. Certainly the author's own experience of scanning has tended to confirm this view. In theory scanning is carried out for periods of up to two hours continuously, although in practice the fact that all the scanning is carried out in one room means that there are many informal breaks when work stops for conversations with other scanners. In general it is felt by the scanners that the complexity of the task is such that it would be difficult, if not impossible, to perform whilst carrying out a conversation.

In view of these conflicting factors, it was felt to be of interest to investigate this variable experimentally.

4.6 Task characteristics

4.6.1 Pacing

The task is self-paced, which, as suggested by the review in the last chapter, should produce detection performance superior to a paced situation.

4.6.2 Enhancement of signal discriminability

Although the optical system produces a magnified image of the events, the actual inspection procedure is carried out on these magnified images, and hence no additional magnification is employed. The luminance of the scanning table surface (4 lumens) was within acceptable

limits. The only technique employed to enhance the discriminability of the defects was the informal one described in section 4.5.2 in which scanners look along the beam tracks in order to detect the oblique tracks of events.

4.6.3 Complexity

In considering how the complexity of the films being inspected affects event detection probability it is important, as pointed out in the last chapter, to consider the dimensions along which complexity is to be measured. In scanning, the complexity of the film can be regarded as the number of non-signal configurations likely to be present on a given film. Although there are occasional films in which there are a large number of beam tracks, which produce a high incidence of unwanted configurations, steps are usually taken to ensure that the beam tracks are widely spaced and are relatively few in number.

Another aspect of complexity that needs to be considered is that of the events themselves. Whether a four pronged event is more discriminable than say, a two pronged event, remains an empirical question. On commonsense grounds, one would expect differences, since the four-pronged event (known as a 400 in scanning terminology), has more of the event characteristics (oblique tracks) than a two pronged 200. In fact many scanning errors consist of misidentifications of events rather than missing them completely. In most cases it is necessary to ensure that both views are used to identify an event.

4.6.4 Signal rate

Because scanning is a self-paced inspection task, the signal rate in time will be determined by the average time that a scanner takes to scan a frame, and the overall probability that an event is present on a particular frame. Work by Colquhoun (1961), suggests that it is the probability of an event occurring given that a frame is presented that would determine detection efficiency, rather than the total number of frames scanned. Therefore one would not expect 'fast' scanners to detect a greater proportion of the events than 'slow' scanners, if one is considering only the event incidence in time. In general the probability of an event remains approximately constant for a given event type. We would not therefore, expect to encounter problems associated with the inspector being unable to modify his response strategy to take into account a sudden change in event probability, as was discussed in the last chapter. However, where a scanner was required to inspect for a different event type, with a different probability of occurrence than he had been used to, performance could be sub-optimal because of an inappropriate criterion.

The commonly occurring types of event have a probability of about 0.2. Usually, the scanner will be scanning for a range of events with probabilities varying from 0.2 to approximately 10^{-6}. These latter very rare events have an extremely high 'payoff' associated with them in terms of their interest to physicists, so that the criterion associated with them would not be as strict as might be expected from consideration of their probability alone. In the actual scanning task then, the inspector would be utilizing a range of criteria for the different event types.

4.6.5 Number of inspectors

Although there is only one inspector to each scanning table, the presence of the other inspector at the opposite end of the scanning machine, using the other scanning table, should enhance performance through the social facilitation noted in the studies reviewed in the last chapter.

4.6.6 Repeated inspection

All films that pass through the data analysis group receive two independant inspections from different scanners. Subsequently, comparison lists are compiled of differences between the two scans, and the film is then scanned for a third time by an experienced 'fine scanner', who utilizes the lists to resolve the differences between scans, as described earlier. Therefore the repeated inspection studies described in the last chapter are relevant here.

4.6.7 Environmental conditions

The environmental conditions are described fully in Embrey (op.cit.). The main problems are the poor ventilation and heating in the scanning room, and the bursts of high intensity noise from measuring machines in the same room.

4.6.8 Organizational and social factors

The organizational details of the system have been described in Embrey (op.cit). Considering social factors within the group, there tends to be friction between the old established workers, and the students who are

employed during vacations. The latter often quickly learn to perform the task satisfactorily but are easily bored.

4.6.9 Selection and training

Selection is on the basis of an interview with the supervisor of the data analysis group and the departmental personnel officer. Training is informal, and consists of an introductory lecture on the organization of the group and some of the basic elements of scanning. Subsequent training is largely 'on the job', the trainee being assigned to sit with an experienced inspector during the scanning procedure. In view of the complex nature of the stimuli, the perceptual skills required for scanning usually take a long time to acquire.

4.7 Conclusions regarding scanning from an ergonomics standpoint

In view of the fact that the task is self-paced and that repeated inspection is employed one would expect the system to be highly efficient in its operational objective of detecting particle interactions on film. Selection and training appear to have received little attention however, and the badly ventilated and noisy working conditions are likely to affect performance. There seems to be a possibility that there will be performance decrement with time, but this needs to be verified experimentally. The discrimination required to detect events requires both adequate visual acuity to acquire the information and the possession of high level perceptual skills to distinguish the wanted from irrelevant patterns. The scanner can employ learnt strategies, e.g. looking along the beam tracks, to enhance the detectability of the events. Although search is employed to locate an event, it is not clear how conventional search theory might be employed to predict the

time taken to achieve this. Signal detection theory would seem to be applicable to determine the degree to which the scanners employ an optimal decision criterion and the factors which affect their sensitivity for defects.

4.8 Experimental objectives

4.8.1 Specific practical goals

The analysis of the scanning task in the last section suggests several possible research objectives of interest. However, the experimental work eventually carried out had to reflect the specific needs of the organization concerned.

In the current investigation, two factors were of particular interest to management. These were whether different types of event differed in detectability and whether the noisy conditions in the scanning room degraded performance in the inspection task. When the possibility of performance deterioration as a result of prolonged periods of scanning was discussed, it was felt that this variable should also be investigated.

The interest in establishing whether there were intrinsic differences in detectability for the different event types, stemmed from certain statistical considerations connected with high energy physics. The calculation of various physical parameters was based on the probability of occurrence of different event types as inferred from the number of events detected by the scanners. These calculations assumed that all types of event were equally easy to detect. There is little data available on the effect of defect complexity as such on detectability.

Most of the studies using complexity as a variable have considered the overall complexity of the items being inspected, rather than defect complexity itself. As suggested in section 4.6.3, we would expect the number of prongs emanating from the production vertex of an event to determine its detectability, since it is primarily these characteristics which differentiate an event from the most commonly occurring form of background noise, the beam tracks. This hypothesis needs to be tested experimentally however.

The proposal to investigate the effect of noise levels on the detectability of defects was prompted by the fact that the management were trying to decide, on cost effectiveness considerations, whether or not to scrap the older machines that were responsible for most of the noise problems. They had not previously considered the possibility that the noise levels being generated might be affecting scanning performance, and hence were interested in obtaining objective evidence for this effect.

If definite evidence of performance deterioration with prolonged periods of scanning was obtained, management were prepared to consider the possibility of rescheduling the rest pauses.

4.8.2 Theoretical goals

This study provided a useful vehicle to investigate the practical utility of the theoretical orientation discussed extensively in Chapter 2: Signal Detection Theory. It was of considerable interest to evaluate the extent to which theories of this type could be translated from laboratory situations to real world tasks. As the earlier review suggested, many of the applications of SDT to date seemed to have not

adequately checked whether the underlying assumptions of the equal variance model apply.

If SDT in some form could be applied to the experimental data, then this would facilitate the separation of bias and sensitivity effects in the detection results and allow the comparison of the criterion used by the scanners with the theoretical optimum.

The effects of noise on detection situations employing patterns as complex as those found in this study had not previously been investigated, and it was of interest both theoretically and practically to see if those effects were different from those found in simpler detection situations. It was also of interest to investigate whether the nature of the noise had any differential effect on detection performance. Most studies had employed continuous white noise as the stressor, whereas the noise in the scanning room consisted of intermittent bursts as the machines were operated.

The investigation of time related decrements on the scanning task would provide additional evidence for the generalizability or otherwise of vigilance research to real world tasks.

Finally the possibility of interactions between the variables of noise pattern complexity and time on task was thought likely to provide further insights of theoretical interest.

4.8.3 Summary of research objectives

A. <u>Practical</u>

1. To investigate whether there were intrinsic differences in

detectability between events of differing complexity.

2. To evaluate the effects of various types of auditory noise on detection performance.
3. To determine whether detection of events was affected by prolonged scanning, i.e. time on task.

B. Theoretical

1. To investigate the applicability of SDT as a usable model in a real inspection situation.
2. To consider the effects of auditory noise on various parameters of detection performance. Different types of noise were to be considered.
3. To verify or otherwise the applicability of vigilance data to the scanning task.
4. To consider the interactions of the major variables present in the study.

4.9 Experimental philosophy

As the objectives of this study were to collect data in as realistic a situation as possible, it was necessary to conduct the experiment using real films and employing the scanning machines customarily used by the inspectors. The most authentic results would have been obtained by introducing a test film into the everyday work of the scanners and subsequently scoring it for accuracy. There were a number of disadvantages to this procedure. The unofficial breaks taken and other occurrences difficult to allow for, lead to the decision to conduct the experiment using a greater degree of control.

4.10 Experimental work

4.10.1 Hardware considerations

As far as the scanning task itself was concerned, two aspects needed modification. The first of these was the tendency of some of the scanners to advance the film so that only part of it was visible on the scanning table. They would then examine the tracks by moving a separate handle which moved the optical system independently of the film. In order to obtain consistent estimates of the total time to scan a frame, it was necessary to ensure that each frame was presented at a standard position at the commencement of each scan. Ideally the frame needed to be positioned on the scanning table such that the whole of it was visible. A second problem concerned the use of the various views available. Some scanners tended to scan using one view only, and when they encountered an event which was difficult to resolve on a single view, they would wind on the second and sometimes even the third view to obtain the additional information present on the corresponding frames. Sometimes these views would be very far behind the current frame and a considerable time might elapse until they were wound to the appropriate position. This procedure would clearly adversely affect any estimates of scanning time.

Both of these problems were resolved by modifying the film advance mechanism of the scanning machines. As described in detail in Embrey (op.cit.) an electronic film advance mechanism was designed such that by depressing a button, all three views of the film advanced simultaneously. When the frame image was at the correct position on the scanning table, a photocell sensor stopped the film advance mechanism.

During the actual experiment, subjects were asked to depress one of two buttons after they had scanned a film frame, depending on whether they felt that there was an event of a specified type on the frame. Depressing one of the buttons caused an oscillator tone to be recorded on tape, and simultaneously initiated the film advance sequence. The tapes were subsequently analysed using SETAR (Welford 1952) to give an output indicating the nature of the response (i.e. event present or absent) and the elapsed time since the previous response, i.e. the response latency in a self-paced task.

4.10.2 Experimental design - general

The basic conditions to be investigated were auditory noise levels and types, differing complexities of events, and time on task. Two of the noise conditions utilized continuous white noise, one level being a masking noise condition of 65dB and the other corresponding to the noisiest conditions in the scanning room of 85dB. The third noise condition consisted of an actual recording of the highly variable noise environment in the scanning room, the average intensity of which was 85dB. The object of this condition was to investigate whether the nature of the noise, apart from its intensity, had any effect on performance. The two most commonly occurring types of event, the four pronged and two pronged topologies, were chosen as the two levels of pattern complexity, which were known to be approximately equiprobable on the film to be used. A time interval of thirty minutes was chosen for the task duration, which was divided into three periods of 10 minutes for the purpose of the performance decrement analysis. Subjects started scanning at random points on the film, subject to the proviso that there were sufficient frames available for the fastest scanner to work for 30 minutes. It was not possible to control

completely for time of day effects because different subjects belonged to shifts which started at different times. Attempts were made, however, to ensure that the various sessions took place at approximately the same time within each shift for each subject. The subjects employed were seven experienced scanners, all with at least one year's experience. They all had normal or corrected vision. There were six males and one female (subject 7).

4.10.3 Experimental design - statistical

A 3 x 2 x 3 x 7 complete factorial repeated measures design was employed, the factors being noise conditions, event complexities, time intervals and subjects respectively. This design is discussed in Kirk (1968) p.237, Myers (1966) and other standard texts. Poulton (1969) has criticized such designs on the grounds of possible carry-over effects between treatments. In the opinion of the present author however, these criticisms are really applicable primarily in laboratory studies where learning effects between trials are almost inevitable with the time available for practice in typical experiments. In industrial experimentation, where very highly practised subjects are available, as in the present study, it is felt that such effects will be minimal, and hence the repeated measures design is considered appropriate. The desire to test a number of variables and the relatively limited number of trained subjects available made the subjects X treatments design a natural choice.

All the factors in the experiment were assumed fixed. In the case of the experimental treatments, the particular levels of interest of the variables considered were all included in the experiment. The justification

for using subjects as a fixed rather than a random effect was that the conclusions drawn from this study were intended to be specific to the group under study. Additionally, scanners are a specialized group and cannot be regarded as being randomly sampled from the population. The analyses of variance were performed by a program from the IBM Scientific Subroutine package, modified extensively to produce the particular design used. Where appropriate, arcsine or log transformations were used to reduce the heterogeneity of variance of the raw data or where obvious skewness of the distribution existed.

4.10.4 Analysis of data

The basic data from the experiment consisted of yes or no (event present or absent) decisions for each frame together with the time interval from the presentation of the frame to the response by the scanner. This was obtained from SETAR as described in Embrey (op.cit.). A computer program (DATAL, Appendix D) converted this data to give the measures set out below:

1. Performance measures based on SDT theories

(a) Parametric

d'

beta

(b) Nonparametric

(I) Sensitivity

Pollack Norman index

Latency sensitivity index (Navon 1975)

(II) Bias

Hodos-Grier index

ZFA

2. Inspection performance indices

   Indices A1 and A4 (McCornack 1961)

   Correct detection probability

   False alarm probability

3. Response latency measures

   Correct rejection (i.e. correct decision that frame did not contain an event) latency

   Correct detection latency

   False alarm latency

   Missed defects latency

The inspection measures A1 and A4 are defined as below:

$$A1 = \frac{NCR + NCD}{NFA + NCR + NCD + NOM} \times 100$$

$$A4 = \frac{NCD - NFA}{NCD + NOM - NFA} \times 100$$

where NCD = no. of correct detections of signals

NFA = no. of false alarms

NOM = no. of missed signals

NCR = no. frames correctly rejected as not containing signals

The program also performed arcsine and log transforms on some of the data prior to the analysis of variance. Most of the measures set out above have been discussed in Chapters 1 and 2. Any new measures considered will be discussed in the text.

4.10.5 Treatment of zero cells in SDT analyses

I am indebted to Dr Raj Parasuraman for suggestions on this topic. In this study, as in most detection experiments, on several occasions subjects made no false alarm responses during the experimental period

being analysed. This creates difficulties for the calculation of SDT parameters because without a pair of correct detection and false alarm probabilities beta and d' cannot be calculated. When these probabilities are obtained from detection experiments, the quantity we are actually measuring is the relative frequency of a particular response category over a number of n trials, which tends to the actual probability as n becomes large. If a given time period does not contain any false alarms, for example, this does not automatically mean that commission error probability during this period is zero. It simply indicates that the probability is too small to be estimated by the use of relative frequency techniques.

Several methods are available for the estimation of false alarm probability P(S/n) during a period in which n non-signal trials occur and no false alarm responses are made.

The technique that has usually been adopted, Jerison, Pickett and Stenson (1965), Wallack and Adams (1969), is to assume that half a commission error has occurred. This is equivalent to assuming that:

$$P(S/n) = \frac{1}{2n} \qquad (1)$$

Another technique is to take a weighted average of the probabilities associated with each noise trial in the observation period, i.e.:

$$P(S/n) = \frac{1}{2} + \frac{1}{2}^2 + \frac{1}{2}^3 \ldots + \frac{1}{2}^{n-1} \; / \; 2n$$

$$= \frac{1 - 1/2^n}{2n} \qquad (2)$$

A final possibility considers the likelihood of no commission errors occurring within a period.

Probability of a commission error not occurring during a single trial = 1 - P(S/n).

Assuming a fixed probability, the probability of no commission errors occurring in n trials = $(1 - P(S/n))^n$.

The likelihood of this can be tested by setting $(1 - P(S/n))^n \leqslant \frac{1}{2}$, from which $P(S/n) \geqslant 1 - \frac{1}{2^{\frac{1}{n}}}$, where $p(S/n) = 1 - \frac{1}{2^{\frac{1}{n}}}$ (3)

For large n (>50) equation (3) gives the best estimate of P(S/n), and was therefore used by the programs where zero false alarms occurred.

4.11 Results and discussion

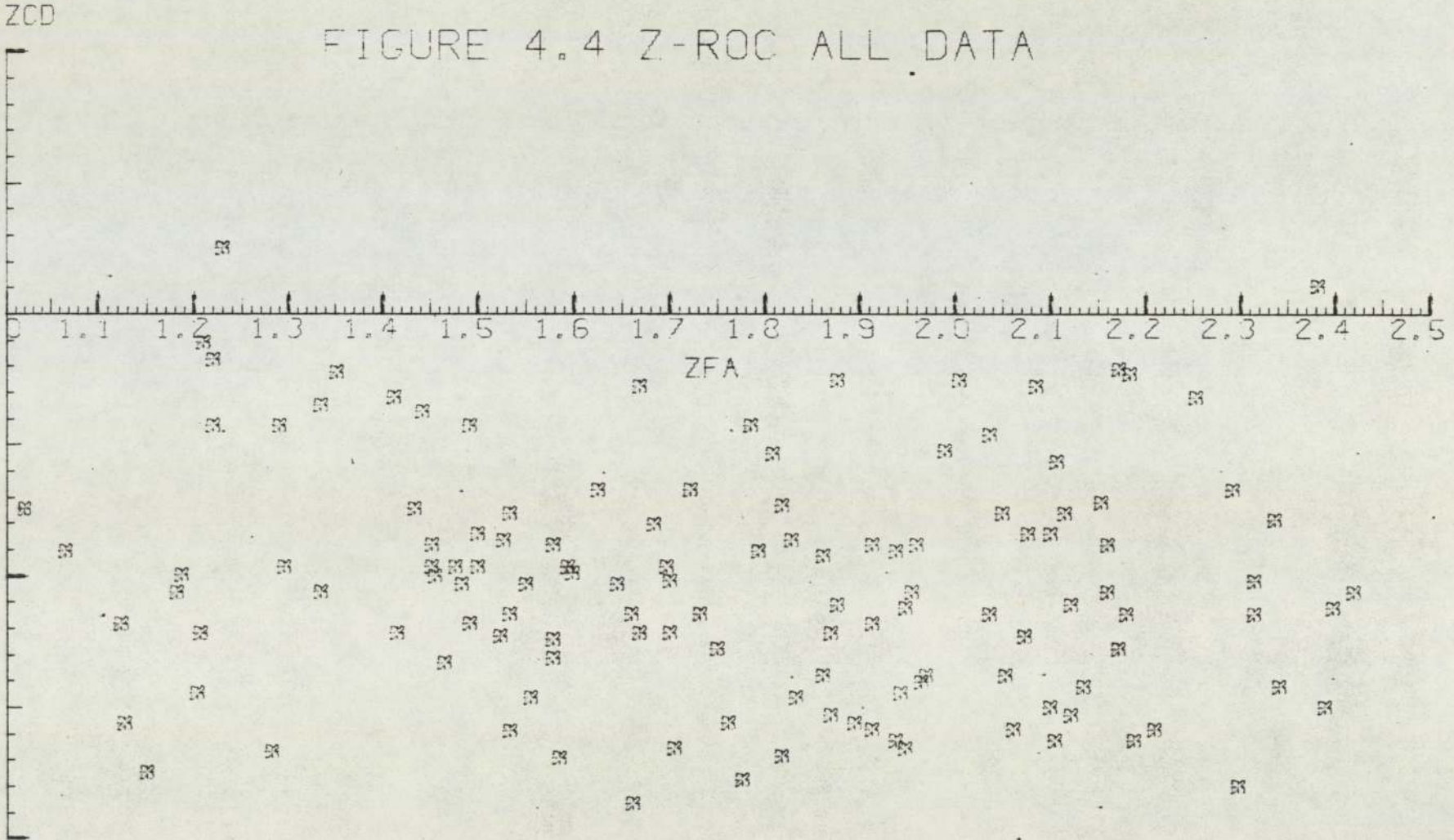
The summary data and means are given in Appendix B and the analyses of variance referred to in the text can be found in statistical Appendix A. The first group of results discussed will concentrate on the theoretical implications of the data prior to a consideration of their practical significance.

4.11.1 SDT considerations

4.11.1.1 The applicability of SDT to the data

One of the major goals of this study was to investigate the applicability of SDT in the task under consideration. The first test which can be applied is to plot the Z transforms of the false alarm and correct detection probabilities against one another to give the normalized or Z-ROC curve. Figure 4.4 shows this for all 126 data points. The data clearly do not fall on a straight line as predicted by SDT.

FIGURE 4.4 Z-ROC ALL DATA



A nonsignificant correlation confirmed this. On the other hand this result is not too surprising in that we are superimposing ZFA and ZCD values obtained across a wide variety of experimental conditions and subjects. Analysis of variance 1 for ZCD shows significant differences between subjects ($p < .001$) and time intervals ($p < 0.05$) whereas for ZFA there is a significant time interval x subject interaction. In order to obtain meaningful ROC curves it is therefore necessary to consider the ROC curves within time intervals within subjects. It was mentioned in Chapter 2 that because of the error inherent in the estimation of both false alarm and correct detection probabilities, the ROC curve should be obtained using maximum likelihood rather than least squares techniques, which assume an errorless independent variable. Considering the data within time intervals, within subjects, allows the utilization of the Grey-Morgan maximum likelihood (ML) fitting program, Grey and Morgan (1972). This program normally only accepts data from rating experiments in which a series of ascending confidence ratings have been made. Each pair of false alarm and correct detection frequencies in the data subset under consideration provides one point on the ROC curve, but if the raw frequencies were entered into the ML program without regard to order, the program would fail in attempting to fit a single straight line. A preliminary program (SIGROC, Appendix D) therefore first calculated a series of ZFA values from the false alarm frequencies. Since ZFA is monotonic with the magnitude of the sensory evidence regardless of the variances of the noise and signal + noise distribution, (Chapter 2), sorting the ZFA values into ascending order together with the associated FA and CD frequencies provided the program with the appropriate inputs. Part of the output from the program is given in Appendix A. The straight line Z-ROC fits the data in every case, as tested by chi-square. The mean ratio of the signal + noise to the noise variance is 1.362 which differs

significantly from 1 ($t = 14.47$, $p < 0.001$). Although the Grey-Morgan program, strictly speaking, was written with the rating experiment as its underlying model, its use in the present context is justified on the grounds that it is being employed purely as a device for fitting a straight line to the data using maximum likelihood techniques.

The use of the program also enables us to correct the beta values found by assuming the equal variance model, to the actual betas employed by the subjects in the unequal-variance situation. The Grey-Morgan program produced estimates of the variance ratio (the reciprocal of the slope of the fitted Z-ROC curve) for each block of time data within subjects that it was applied to. The unequal variance betas are obtained simply by multiplying the equal variance betas by the appropriate Z-ROC slope (McNicol (1972) p. 92). This was done for each of the 21 blocks of data for which the Z-ROC had been fitted.

Hence the data appears to be describable by the unequal variance SDT model. This is in accord with most laboratory studies using SDT with visual tasks, but does not agree with the earlier suggestion that the well practised subjects used in this study might be expected to know the signal characteristics as well as the non-signal attributes, and hence have equal variance internal distributions. It seems possible that because of the very low incidence of false alarms found in this study, the other cause of apparently unequal variance distributions, the greater sampling error involved in the calculation of the signal distribution variance, may be an important factor. The present finding differs from the only other published industrial study using SDT, Drury (1973), which found the equal variance model to fit the data. The main difference between the two studies was that Drury's subjects

were provided with rapid feedback during the latter part of his study, whereas no direct feedback for the scanners in this experiment was provided. However, even before feedback was provided in Drury's study, a slope of 1 for the ROC curve was obtained. Probably the main reason for the different findings in each study is in the nature of the signals in each case. The variability of glass faults is relatively small compared with the extremely wide range of configurations that are found on bubble chamber film, and this would tend to increase the variance of the signal distribution in the latter case.

4.11.1.2 Other tests of the SDT model

A number of methods are available for testing whether certain other assumptions of the SDT model hold in this situation. Ingleby (1974) points out that although the likelihood ratio criterion, beta, is the theoretically ideal measure of how much weight to attach to a sensory datum, it has never empirically been established that human subjects actually do set their criteria in terms of beta rather than, for example, the sensory evidence x itself.

It can be shown (McNicol (1972) p.64) that:

$$\log \text{beta} = d'x - d'^2/2 \qquad (1)$$

in the equal variance case, (x = ZFA, since this is monotonic with the sensory evidence used by the observer). In the unequal variance case, the expression becomes:

$$\log \text{beta} = x^2 [(\sigma_s^2 - 1)/2\sigma_s^2] + d'x - (d'^2/2) - \log \sigma_s \qquad (2)$$

If the observer is actually positioning his criterion on the basis of beta, the first implication of 1 and 2 is that there should be a linear relationship between x , i.e. ZFA, and log beta in the equal variance case, and a parabolic relationship in the unequal variance

situation. In fact, with the value of the variance ratio found in this experiment (1.3) it would be difficult to distinguish between a linear and parabolic regression with the range of probabilities which occurred. We can therefore take a good linear fit of the relationship between ZFA and log beta as reasonable evidence that a likelihood ratio criterion is being used.

The graph of log beta v. ZFA is plotted in Figure 4.10 and it can be seen that a good fit is obtained, ($r = 0.74$ $p < 0.001$).

Confirmation that the unequal variance model applies can be obtained by considering the change in ZFA for a given change in log beta at different values of d'. The parabolic relationship between log beta and ZFA of equation (2) suggests that ZFA should change less at high values of d' than at low values, with changes in log beta. We can test this by obtaining the regression equations for ZFA v. log beta for each subject. Since there are significant differences between d' for subjects (analysis of variance, Appendix A p.) a plot of the slope of the regression lines (a measure of the rate of change of ZFA v. log beta) against $^{1}/d'$ should be linear. As Figure 4.6 shows, there is a high degree of relationship with $r = 0.757$ ($p < 0.01$).

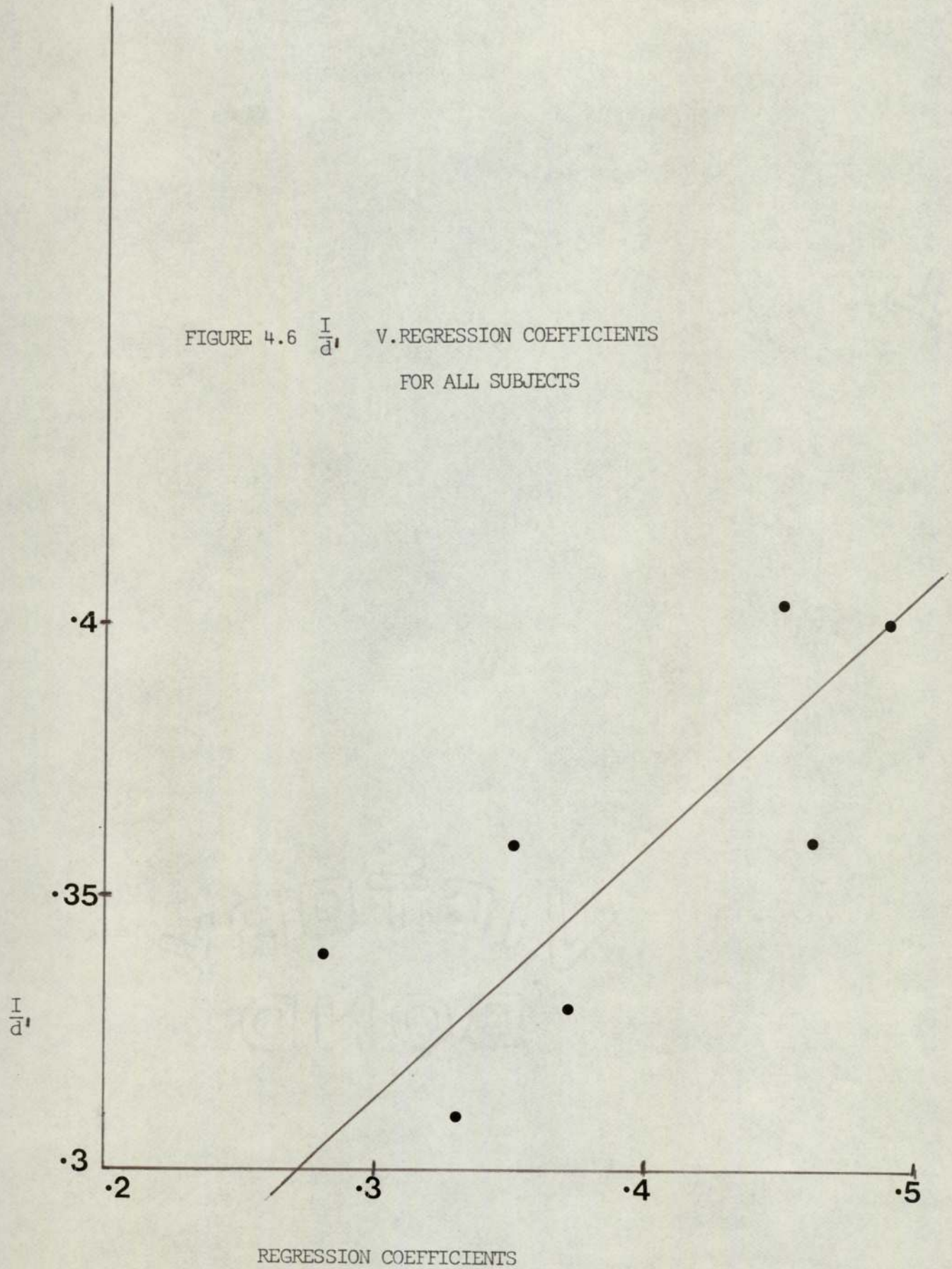
The preceding tests strongly suggest that the SDT model provides a good description of the data. The evidence is not entirely unequivocal however. The equal variance model suggests that beta should be related to the a priori probability P by the equation:

$$((1 - P) / P) \times (\text{relative cost factor}) = \text{beta} \quad (3)$$

(see Chapter 2 p. 27).

A significant correlation was only obtained between beta and $(1 - P)/P$

FIGURE 4.6 $\frac{I}{d'}$ V.REGRESSION COEFFICIENTS FOR ALL SUBJECTS



for one out of seven subjects. Although the fact that the unequal variance model does not produce a linear correlation, since the variance ratio is not very large, one might have expected a larger number of significant correlations. The result obtained is probably due to the fact that the a priori probability of the events does not vary sufficiently to produce a modification of the subjects' criteria.

One aspect of SDT predictions that has not yet been considered is the effect of the payoff matrix on the positioning of the criterion. It is clear that in the scanning task we do not have a symmetrical payoff matrix, in that missed signals are more expensive than false alarms, because the latter errors are likely to be picked up at the fine scanning stage. As mentioned in Chapter 2, the payoff matrix is given by the expression:

$$\frac{\text{value of correct rejections - cost of false alarms}}{\text{value of correct detections - cost of missed signal}} \qquad (3)$$

In order to provide an approximate check of the effects of these utilities, a survey was carried out amongst the scanners, and the mean utilities for the various decision alternative was found to be as follows:

value of correct rejections (of non-signal frames) = 4

cost of false alarms = 1

value of correct detections =10

cost of missed signal = 5

Substituting in equation (3), assuming on a priori signal probability of 0.2, the theoretical optimum value of beta obtained is 2.2. By comparison the overall mean beta obtained for the experiment is 2.71. The closeness of these figures supports the view that the inspectors are using a likelihood ratio criterion.

We can conclude, therefore, that the evidence suggests that the SDT model is an appropriate one in the inspection situation under investigation.

4.11.1.3 Relationships between performance measures

The fact that the unequal variance model is appropriate to the data makes it useful to consider the applicability of the non-parametric measures of performance in investigating the effects of the various experimental variables. It is also of interest to look at the relationship of some of the inspection performance indices such as A1 and A4 to the SDT measures.

As discussed in Chapters 1 and 2, the unique advantage of the SDT parameters is that they allow the separation of sensitivity and bias effects. They also have the advantage of resting on a solid theoretical foundation, and of providing a predictive capability. None of the non-parametric measures available for yes-no data are able to offer these advantages and we are therefore justified in judging their usefulness by the extent to which they correlate with beta and d'. In the present experiment it was possible to correct the betas obtained under the assumption of equal variances by multiplying each block of values by the slope obtained from the ROC curve for that particular block, as described earlier. This provided a baseline of 'actual' betas against which to compare the non-parametric measures. Unfortunately such a blanket procedure is not available to correct the d' values. The appropriate version of d' in the unequal variance case is Δ m or d'_e as described in Chapter 2. These can only be obtained from the ROC curve; d' values obtained under the equal variance

assumptions cannot readily be rescaled. In the comparisons that follow it was decided to use the equal variance d' as a baseline. Since the variance ratio is not large, the comparisons will retain some degree of validity.

Plots of the Pollack-Norman and latency sensitivity indices against d' are given in Figures 4.7 and 4.8. These indicate a high degree of correlation in the first case and a smaller but still significant r in the latter. These are confirmed by the product moment correlations in Table 4.1 below:

comparison	correlation	sig.
d' v. Pollack - Norman index	0.923	$p < 0.001$
d' v. Navon latency index	-0.184	$p < 0.05$

Table 4.1 Comparison of sensitivity indices

The relationship between d' and the Pollack-Norman index seems to be slightly curvilinear in nature, although assuming a linear relationship would lead to only slight errors.

The scatterplots of the corrected values of log beta against the Hodos-Grier bias index and ZFA are given in Figures 4.9 and 4.10 and Hodos-Grier v. ZFA in Figure 4.11. The corresponding correlations are given in Table 4.2

comparison	correlation	significance
log beta v. Hodos-Grier index	0.893	$p < 0.001$
log beta v. ZFA	0.743	$p < 0.001$
Hodos-Grier index v. ZFA	0.746	$p < 0.001$

Table 4.2 Comparison of bias indices

FIGURE 4.7 D' V. P-N INDEX

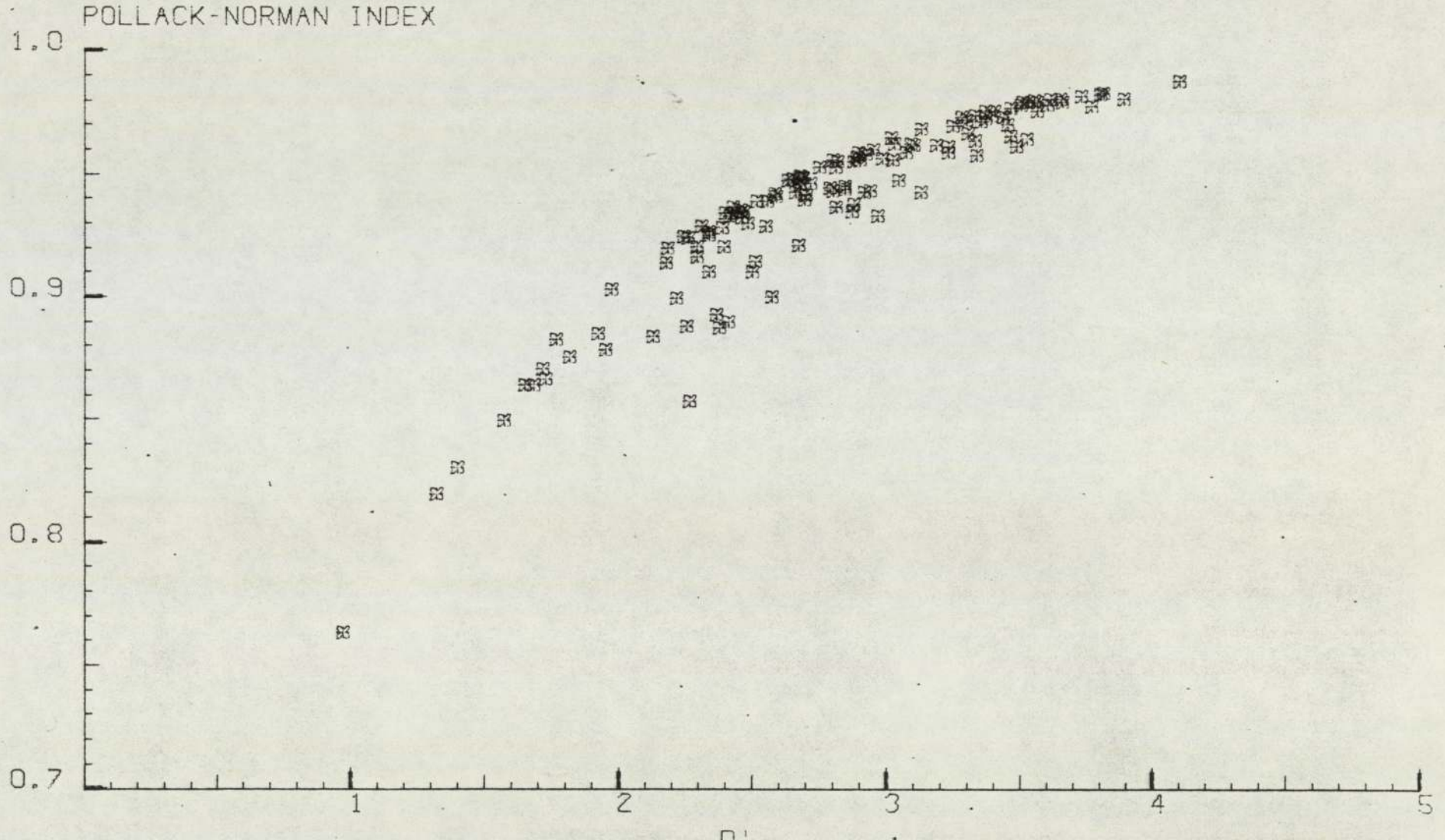


FIGURE 4.8 D' V. NAVON INDEX

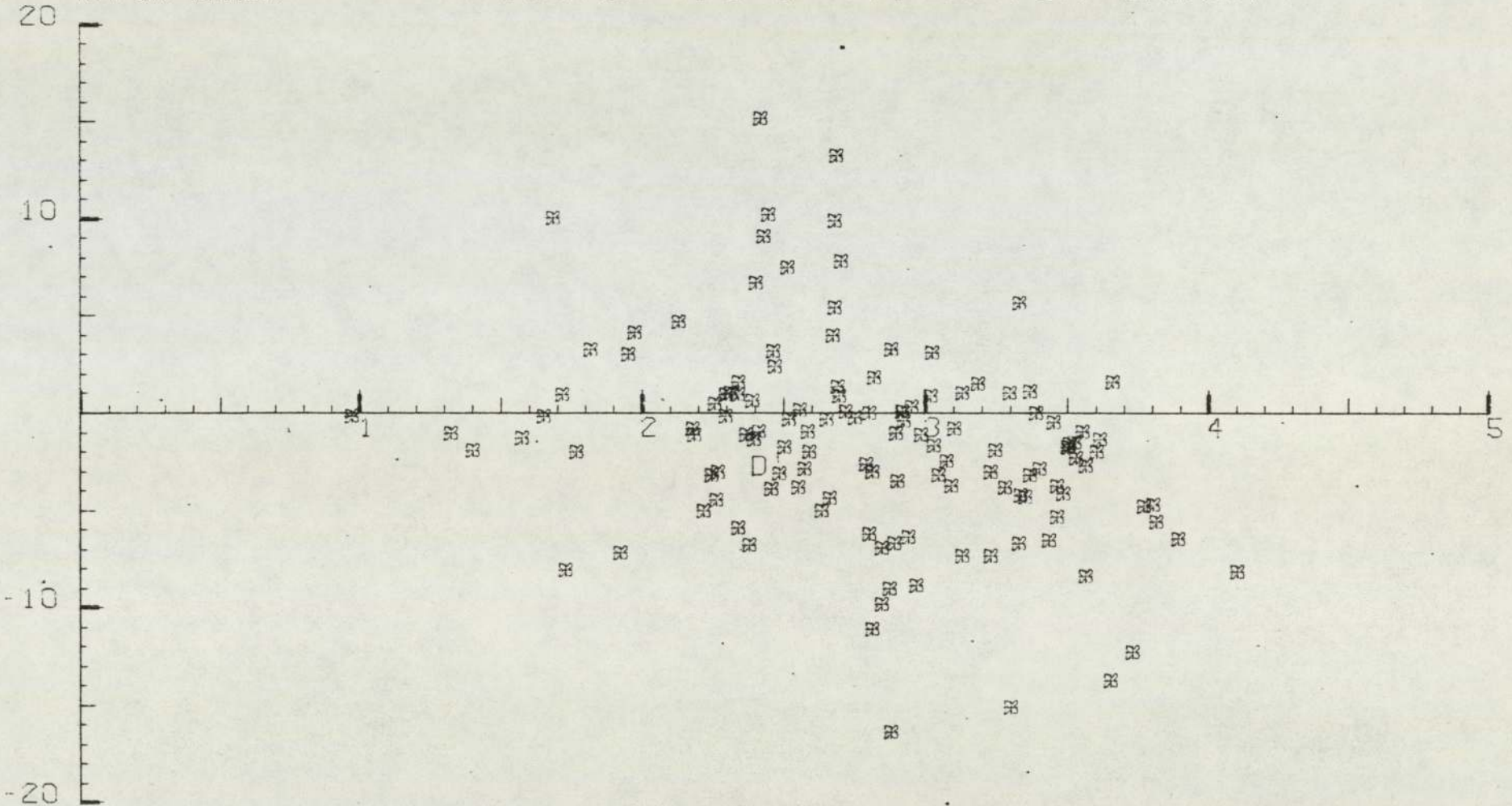


FIGURE 4.9 LOG BETA V. H-G INDEX

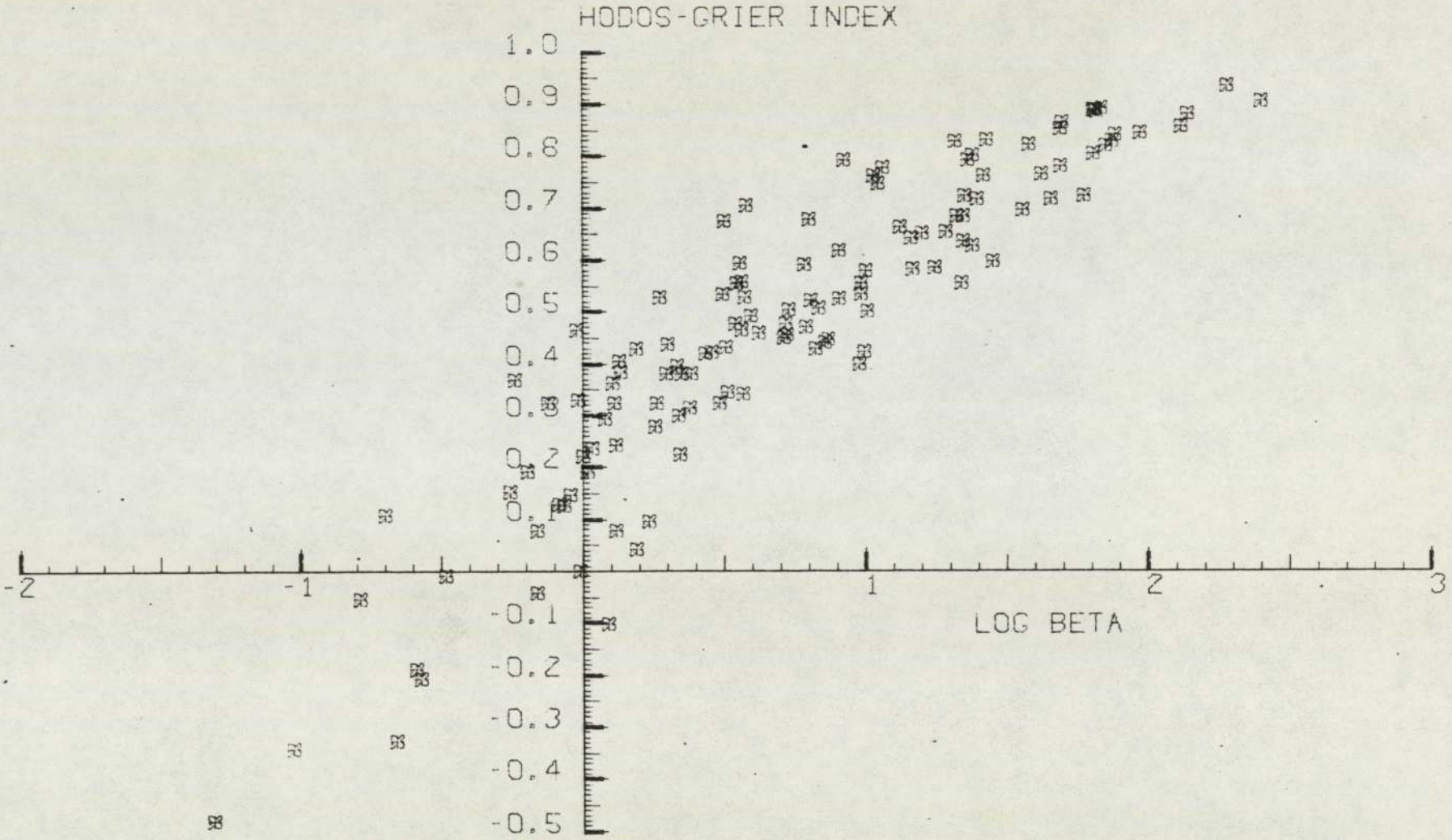


FIGURE 4.10 LOG BETA V. ZFA

FIGURE 4.11 H-G INDEX V. ZFA

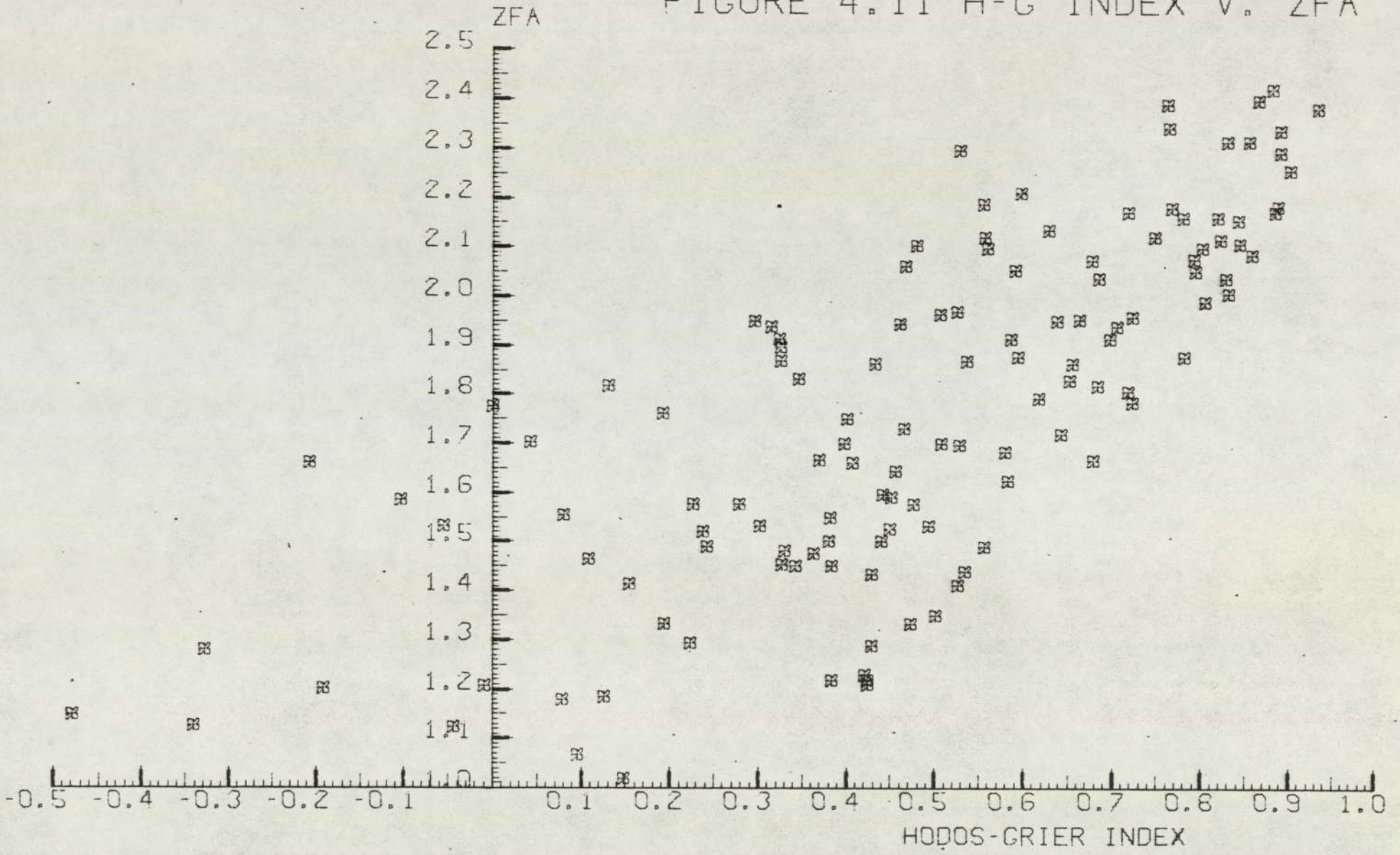
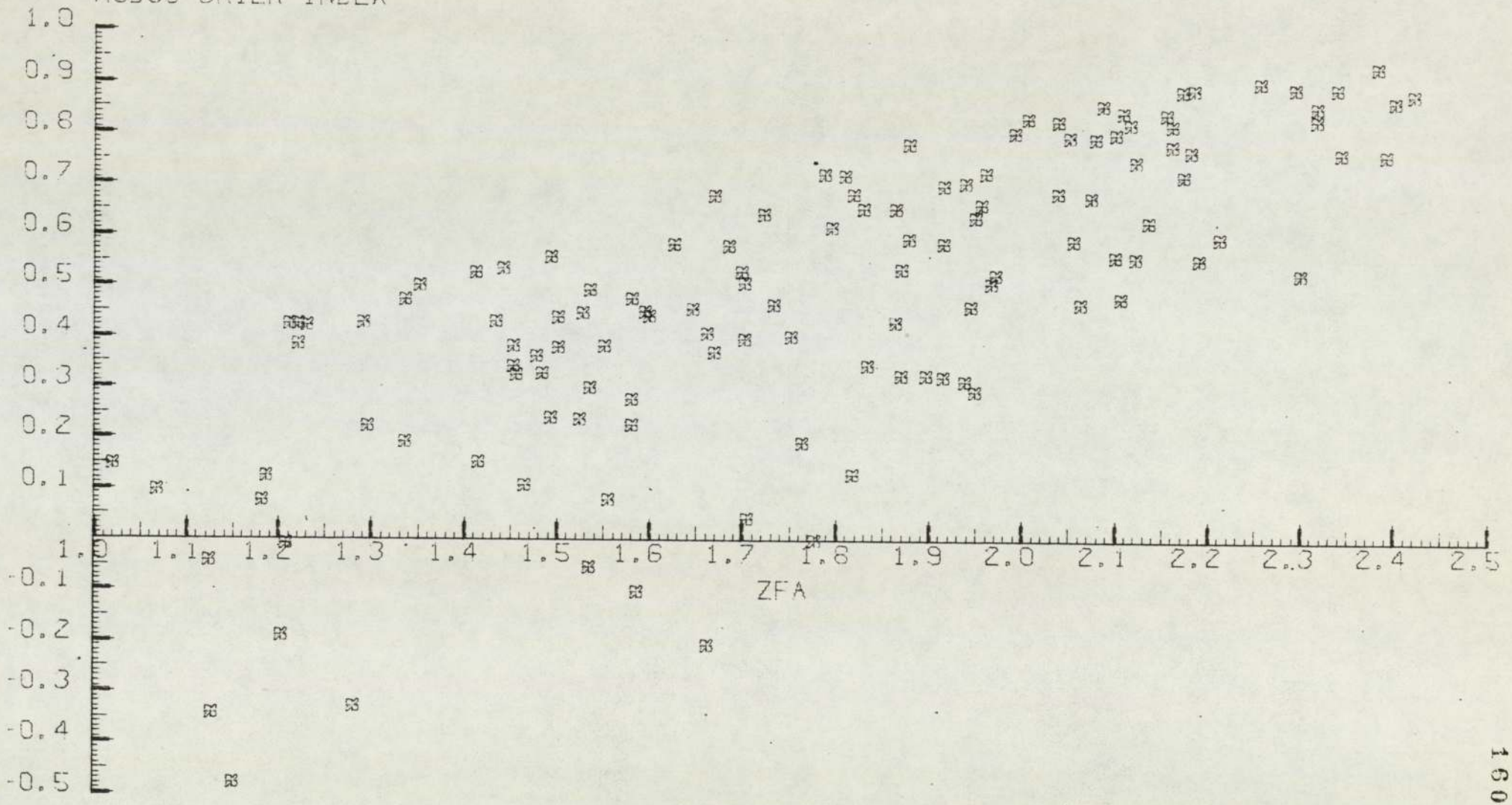


FIGURE 4.11 ZFA V. HODOS-GRIER INDEX



A1 INDEX V. A4 INDEX

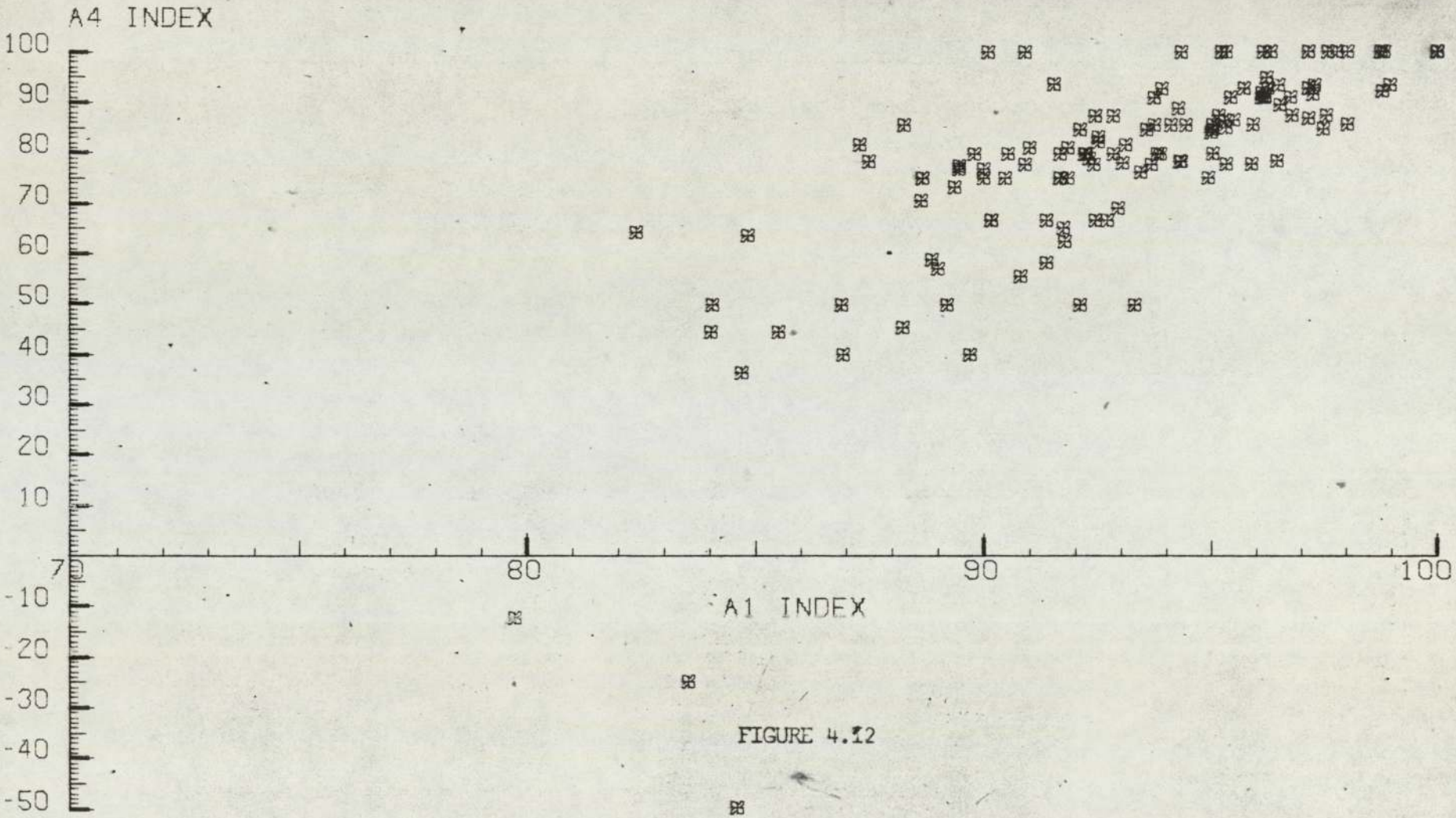


FIGURE 4.12

FIGURE 4.13 A1 INDEX V. D'

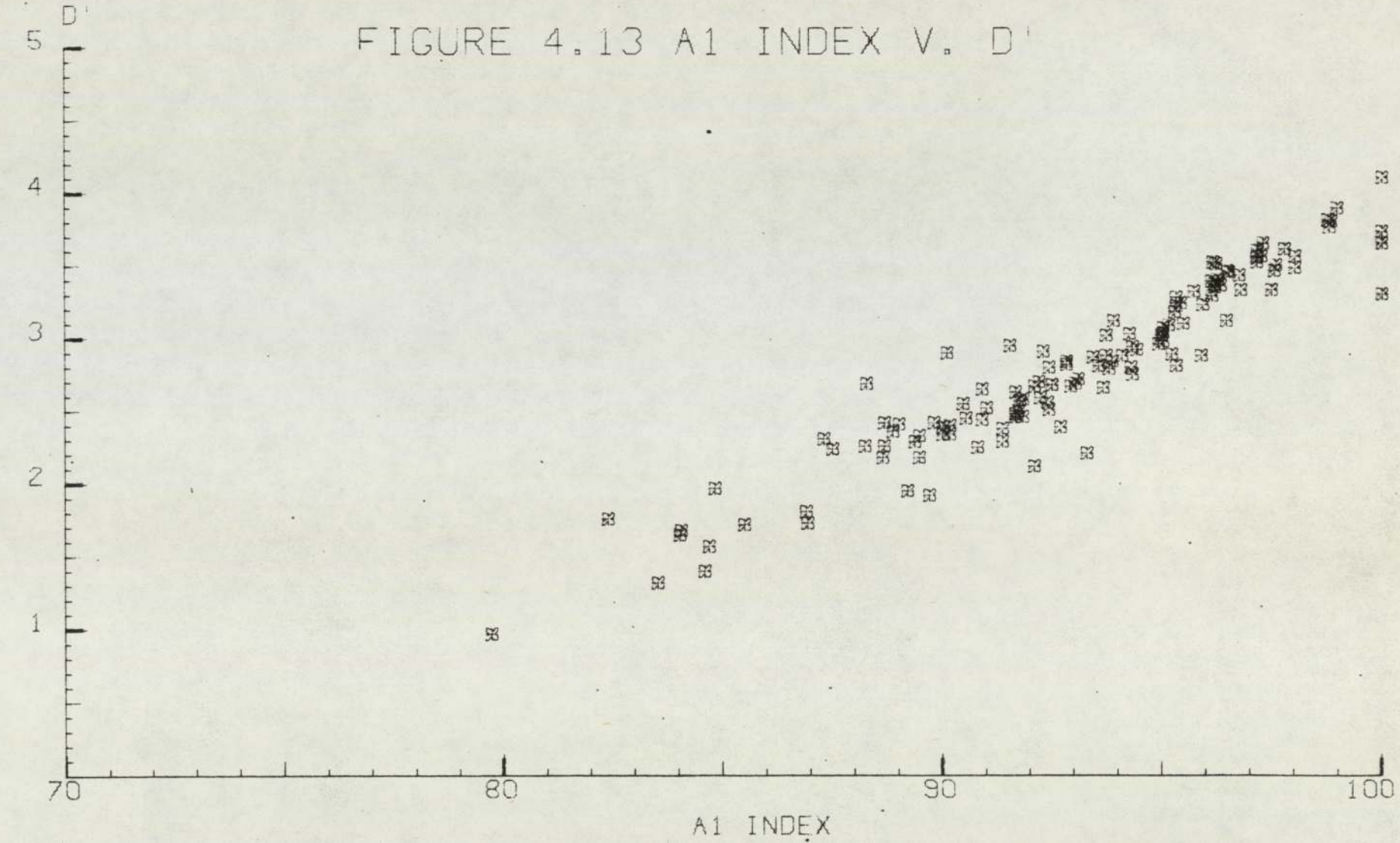
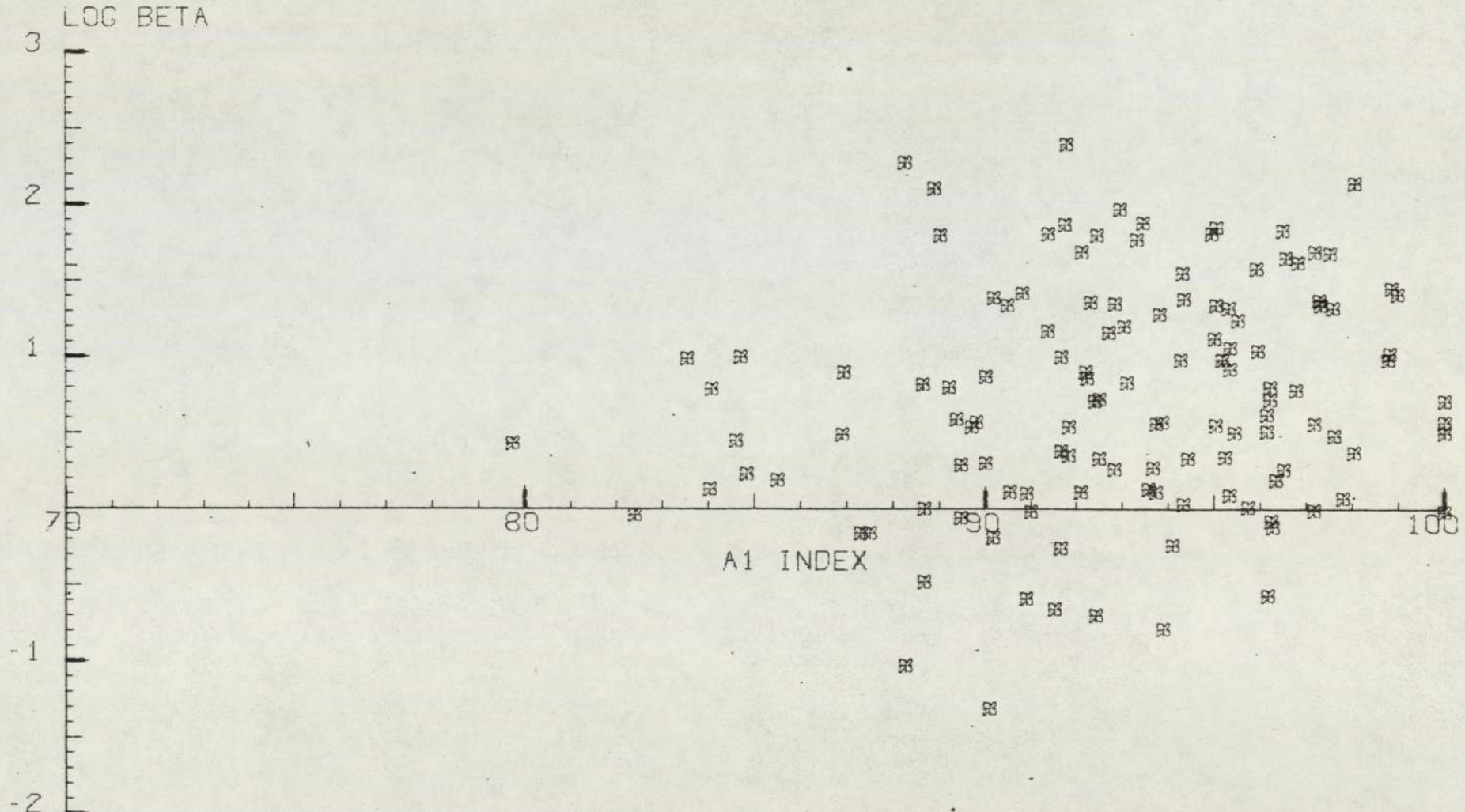


FIGURE 4.14 A1 V. LOG BETA



We can conclude that the non-parametric indices are closely related to the corresponding SDT parameters. This is not surprising in the case of ZFA, because of its monoticity with the evidence variable used by the observer as discussed earlier. The close empirical correspondence between d', log beta and the two corresponding geometrical indices is encouraging in view of the difficulty in establishing an analytical relationship. The fact that the latency sensitivity index also shows a significant correlation with d' suggests that the theoretical response latency model adopted by Navon in deriving the index was correct.

The next performance indices of interest are A1 and A4. Their scatterplot is given in Figure 4.12. The two measures are clearly highly correlated, which is confirmed by an r of 0.742 ($p < 0.001$). Plots of A1 against d', Figure 4.13 and log beta, Figure 4.14 also indicate a high and a more moderate degree of correlation, $r = 0.947$ ($p < 0.001$) and $r = 0.225$ ($p < 0.05$) respectively. These plots demonstrate the essential disadvantage of measures such as A1: they are dependent on both sensitivity and bias changes, and this clearly reduces the possibility of ascribing a cause to a given change in the performance index.

4.11.2 Effects of experimental variables on performance

4.11.2.1 Overall performance measures

Considering the analysis of variance for both the correct detection probability and its arcsine transformation, we see that there are significant differences between subjects ($p < 0.05$) and between time intervals ($p < 0.01$). In fact the time effect is an increase in performance with time, a comparison of means showing that the performance

on the last two ten minute periods of the experiment is significantly better than during the first ten minutes (Tukey test, $p < 0.05$).

The arcsine transformation of the false alarm probability shows a significant time x subjects interaction ($p < 0.01$). Examination of the interaction suggests that there are increases in false alarms with time intervals for five out of the seven subjects. The correct detection and false alarm probabilities taken together suggest that a change in criterion is responsible for the effects, rather than a sensitivity change.

The analysis of variance for the A1 inspection index indicates significant differences between subjects ($p < 0.01$) but does not indicate any time effects. No significant effects are found for A4, in spite of its high correlation with A1.

4.11.2.2 SDT performance measures

The analysis for d' gives significant differences between subjects ($p < 0.01$) but any conclusions drawn need to consider the fact that the unequal variance assumptions do not apply in this case.

In the analysis of variance for log beta the corrected values of beta were used as described earlier. Significant differences were found between time intervals, ($p < 0.05$), with a significant time intervals x subjects interaction ($p < 0.01$). In view of the importance of this variable, a simple main effects analysis was conducted to clarify the nature of the time effects for the different subjects. This is given in Table 4.3 below, and the corresponding graph showing the changes in log beta over time intervals for each subject for

Source	Sum of squares	df	M.S.	F	Significance
T at S1	1.56	2	0.78	2.07	
T at S2	2.78	2	1.39	3.69	$p < 0.05$
T at S3	1.63	2	0.82	2.18	
T at S4	6.75	2	3.38	8.97	$p < 0.01$
T at S5	1.5	2	0.75	1.99	
T at S6	3.4	2	1.7	4.51	$p < 0.025$
T at S7	1.9	2	0.95	2.52	
ERROR		24	0.377		

Table 4.3 Simple main effects analysis: time intervals x subjects interaction for log beta

which the differences were significant is given in Figure 4.15.

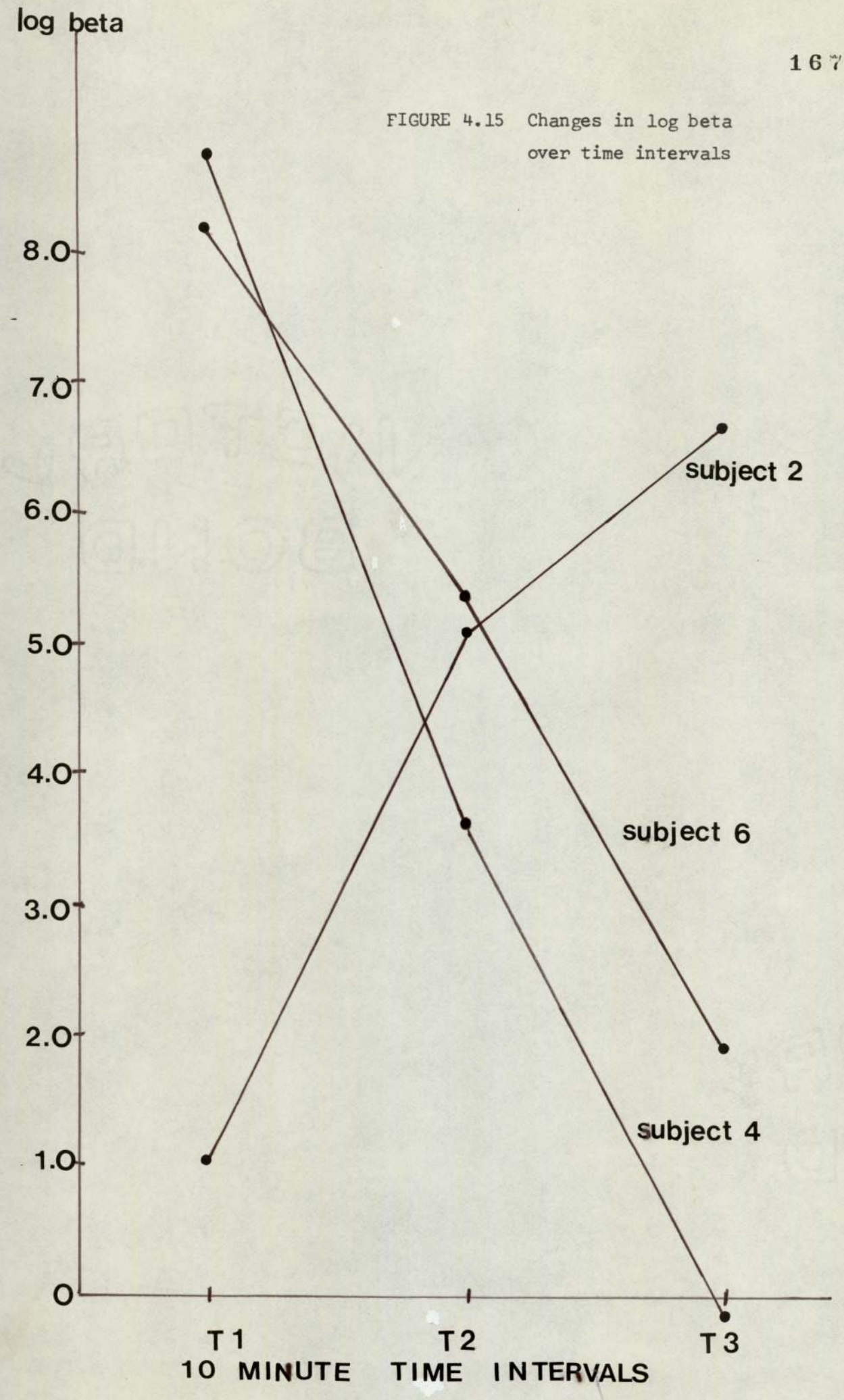
The simple main effects analysis indicates that log beta is significantly different between time intervals for subjects 2, 4 and 6. The significance of the difference between the means for these subjects is shown in Table 4.4 below (Tukey tests).

	Subject 2			Subject 4			Subject 6		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
T1		NS	$p < 0.05$		NS	$p < 0.01$		NS	$p < 0.05$
T2			NS			NS			NS
T3									

Table 4.4 Significance of the difference between log beta means (Tukey test)

With two of the subjects (4 and 6) there is a significant decrease in log beta between the first and last time periods, with subject 2 the trend is reversed. The overall trend, as indicated by the means, is for a decline in criterion with time on task.

FIGURE 4.15 Changes in log beta over time intervals



The Pollack-Norman sensitivity index shows significant differences between subjects on the analysis of variance ($p < 0.05$), paralleling the result for d'. The Navon latency sensitivity index analysis gives significant differences between subjects ($p < 0.01$) and two significant interactions, i.e. time x subjects ($p < 0.05$), and event types x time x subjects ($p < 0.05$). Most of these interactions occur because the quantities used to calculate the Navon index, the latencies for the response categories, are highly sensitive to the experimental variables. As a result, it is difficult to interpret them in this case.

The Hodos-Grier bias index shows a highly significant time effect ($p < 0.01$) and a significant noise condition x time x subject interaction. Consideration of the NxTxS summary table largely confirms the analysis for log beta, for times and subjects. The pattern of the results however, varies in an unsystematic way under the differing noise conditions.

ZFA shows a significant time intervals x subject effect ($p < 0.05$). This is analysed in Table 4.5 to give the simple main effects to compare with the corresponding analysis for log beta (Table 4.3).

It will be seen that the detailed analysis using ZFA does not exactly parallel that using log beta. Although subject 6 shows significant differences in bias across time intervals using both measures, subject 5 shows a significant difference using ZFA as the index and subjects 2 and 4 using log beta. It is apparent, therefore that different results are likely if the two indices are used in statistical analysis, even though the overall trend of the means is identical in each case.

Source	SS	df	S	F	sig.
T at S1	0.07	2	0.035	0.41	
T at S2	0.04	2	0.02	0.24	
T at S3	0.44	2	0.22	2.59	
T at S4	0.43	2	0.22	2.59	
T at S5	0.73	2	0.37	4.35	$p < 0.05$
T at S6	1.44	2	0.72	8.47	$p < 0.01$
T at S7	0.01	2	0.005	0.06	
ERROR		24	0.085		

Table 4.5 Simple main effects analysis for time x subjects interaction for ZFA

4.11.2.3 Latency measures

For the purposes of analysis, the latencies of the four categories of response, i.e. correct detections, correct rejections (of frames containing no events), false alarms and missed signals will be considered separately in addition to the total response time. In most cases, the latencies were subjected to a log transformation before the analysis of variance.

For reasons which will be discussed in the next section, a detailed analysis of some of the more complex interactions will not be given, particularly where these are associated with subject differences.

Considering the correct detection latencies, subjects show highly significant differences ($p < 0.001$) with a significant time by subject interaction ($p < 0.01$). Simple main effects analyses indicate that whether correct detection latency increases or decreases with time on the task depends on the subject, only subjects 2 and 4 showing significant changes.

The correct rejection latency shows a number of significant main effects and interactions. As usual there are significant subject differences, and the main effects of event type and time are also significant at $p < 0.05$ and $p < 0.001$ respectively. Conclusions drawn from these main effects need to be modified by the presence of significant noise x event type ($p < 0.05$) and time x subject ($p < 0.001$) interactions. Simple main effects analysis of the latter interactions shows that all subjects show significant differences between time intervals but that the change is not always in the same direction. As suggested by the time means, however, the predominant change is to a decrease in latency with time on task. Analysis of the noise x event type interaction shows that there are significantly longer correct rejection responses for the 200 as compared with 400 events, but only under the high white noise conditions.

Analysis of the log of the missed signals latency indicates, in addition to the usual subject differences, a significant noise type x subject interaction ($p < 0.05$) a significant noise x time x subject interaction ($p < 0.05$) and a significant event type x time x subject interaction. The means for noise types suggest a longer response latency under the high white noise condition compared with the other types. The event type means imply an overall longer response latency for the 200 events, but the significant interaction times and subjects suggests that this may not be constant over conditions.

Considering the false alarm latencies, the only significant effect apart from subjects is a noise x event type x subject interaction, ($p < 0.05$).

The final latency measure considered, the total latency, shows significant time and subjects main effects, ($p < 0.05$) and $p < 0.001$ respectively,

and significant event type x time ($p < 0.05$) and time x subjects interactions ($p < 0.01$). The means for event types indicate longer latencies for the 200 events and for the earlier time intervals compared with the later.

4.12 Conclusions from experimental study

4.12.1 The application of SDT

The evidence presented in the preceding sections suggests strongly that the unequal variance SDT model satisfactorily accounts for the data obtained. The original proposal that the equal variance model would apply is not supported by the evidence. In view of the fact that the subjects knew the characteristics of signal and noise equally well, as a result of extensive experience, alternative explanations for the good fit of the unequal variance model need to be sought. The greater variance of the signal distribution can be accounted for by the considerable intrinsic variability of the signal in this task and by the sampling error inherent in calculating the false alarm probabilities with the low incidence of false alarms observed in this experiment. These explanations are more tenable than the suggestion that the absence of feedback during the experiment prevented the scanners from learning the characteristics of the signal.

The fairly close correspondence between the optimum criterion for this task and the mean beta actually obtained, adds further weight to the use of the SDT model. It should be pointed out however, that there is a wide range of values of beta about the optimum criterion which will give most of the maximum value possible for the inspector's decisions (Swets and Green (1966) p.93.

The absence of overall significant differences in beta between subjects is in accord with the hypothesis proposed originally, that the scanners would be homogeneous with respect to their criteria, as a result of long experience on the task.

4.12.2 Comparisons of performance measures

Comparisons of commonly used indices of inspection performance A1 and A4, showed that they were correlated with both bias and sensitivity measures, thus limiting their usefulness in terms of suggesting causes for suboptimal performance.

Investigation of the relationships between the parametric and non-parametric indices of bias and sensitivity suggested that the non-parametric indices considered seemed to correlate highly with the corresponding beta or d'. At a detailed level of analysis however, they could not be expected to provide exactly the same results.

4.12.3 Effects of experimental variables

The first questions of interest are those relating to the effects of the experimental treatments on SDT parameters. The changes in correct detections and false alarms with time suggested a decreasing response criterion. This was confirmed by the analysis of variance for log beta, even though a significant decline was only obtained with two subjects, and an anomalous significant increase was obtained with a third. However, the overall pattern of the results was consistent with a reduction in beta with time. The change in beta to a more lax criterion is difficult to account for. In vigilance tasks, the opposite effect, the increasing stringency of beta with time on

task has been accounted for by the inhibition of neural responses, a greater stimulus intensity being required as time goes on, for a signal to achieve a given criterion (Mackworth (1969)). Another explanation advanced is in terms of an inappropriately low signal expectancy at the beginning of the session, leading to few signals being detected, which in turn leads to a lowering of the criterion (Mackworth, (1970)). In the present experiment, as has been emphasized previously, there was no reason to believe that the subjects began with an inappropriate expectancy. If this was not the case, we could account for the results by saying that the subject began the experiment expecting fewer signals than he actually encountered, and that he subsequently lowered his criterion as a result of detecting a high incidence of signals. The neural habituation theory also seems inappropriate, since it cannot be modified to predict a decline in the criterion. In fact the results seem to be more readily accounted for by a suggestion of Welford (1968), that an increase in arousal would move both signal + noise and noise distributions to the right, without modifying the position of the criterion. This would produce an apparent decrease in the criterion. On the other hand the task cannot be regarded as particularly arousing, and there seems no obvious reason why arousal should increase with time.

A more likely explanation would seem to be in terms of an initial modification of the subjects' utilities for the various decision classes, due to their perception of the experiment as a 'high risk' situation. The evidence suggests that initially the scanners approach the experiment with a criterion which was different than that employed in their day to day work. Because they were 'on test' it seems likely that they were initially utilizing particularly stringent criteria as to what constituted a wanted event. This would result in many 'borderline' events being rejected, even though in the real task they would

probably be accepted on the basis that the fine scanner would look at them again and make the final decision. In fact they could be regarded as putting themselves in the position of the fine scanner, who makes the final decisions about which events should be accepted or rejected, and who utilizes more stringent criteria. This effect can be interpreted in terms of payoffs, since during normal scanning, the fact that an event will be looked at again obviously encourages a lax criterion, because the cost of a false alarm is very low. In the experiment, one would expect the 'raised criterion effect' to decline as the scanners become more used to the experiment. Support for this hypothesis comes from an examination of the actual beta values for each of the 10 minute time intervals on the task, i.e. 3.356, 2.587 and 2.197. The final magnitude of beta obtained is very close to the calculated optimum i.e. 2.2. The anomalous subject who increased his criterion may have perceived the utilities of the situation differently than the other subjects.

The implications of this analysis are than the observed decline in beta was a characteristic of the experimental situation rather than an effect which occurs in the day to day performance of the task. Although this may be regarded as a 'negative' finding, it does seem to be of considerable importance when considering industrial experimentation in general. Any experimental study which utilizes an off-line investigation of the type described for signal detection experimentation, needs to control for variables of this type, or to perform a SDT analysis to isolate the effect. The results are also illuminating examples of the way in which subjects are able to modify their criteria without a concomitant change in d'.

The absence of any significant differences between event types was

expected, since there is no a priori reason to expect event complexity to affect beta. Although high intensity white noise is known to affect beta in vigilance tasks (Davies and Tune (1970)), this is generally at a higher level than the 85dB used in the experiment, and therefore the lack of any effect in this experiment is not surprising.
A surprising aspect of the results for d' and the non-parametric sensitivity indices, is that d' should be apparently unaffected by the characteristics of the event. The expected intersubject differences in sensitivity were found but no significant differences between event types, despite the strong a priori reasons discussed in section 4.6.3 for expecting these to differ in detectability. The lack of any effect of noise on d' is again probably due to the relatively low levels (85dB) employed.

The reasons for the lack of an effect on d' of the different complexity of the events are related to the self paced nature of the task, as will be discussed in the next section.

4.12.4 Latency measures

As has been discussed earlier, any response latency observed is the sum of the time to make a structured search to find a configuration of tracks which is a potential event, and the decision time to assign the configuration to the category signal or noise. Visual search theory and the various signal detection latency theories therefore do not in themselves provide good descriptions of the data. Some interesting insights emerge from a consideration of the latency results, although they will not be analysed in detail because of these theoretical difficulties and because response time is a relatively unimportant variable in practical terms for this particular task.

Let us first consider the degree to which the latency results provide support for the application of the SDT model. If we assume that the search time is some random time increment which is added to the decision making time, we can examine the total response latency in the light of SDT concepts. A further assumption that needs to be made is that the latency associated with each decision is a function of the distance of the observation point (in decision space) from the criterion. The further the distance from the criterion the shorter the response latency. This can be interpreted as the further the observation is from the criterion, the more discriminable it is, and hence the less time is required to sample sufficient evidence to make a decision. The notion of a difficult decision requiring a greater decision time is intuitively reasonable.

Considering Yes responses, incorrect Yes responses (false alarms F.A.) will, on average, be distributed nearer the criterion than correct detections (CD), implying that CD latencies will be shorter than FA latencies, i.e.

$$L(CD) < L(FA) \quad - (1)$$

Similarly for No responses, correct rejection (CR) latencies will be shorter than missed signal (Omissions, OM) latencies, i.e.

$$L(CR) < L(OM) \quad - (2)$$

Considering both Yes and No responses, we can say that in a situation such as the present experiment, where the probability of a signal is less than that of noise, No will be the dominant response. No responses will, therefore, on average, be distributed further from the criterion and will hence have a lower mean response latency. Hence both No responses, i.e. correct rejection and missed signal latencies

will be shorter than both Yes responses (correct detections and false alarms). This implies:

$$L(CR) \text{ and } L(OM) < L(CD) \text{ and } L(FA) \quad - \quad (3)$$

Combining (1) (2) and (3), the latencies for the four categories of response should fall in the order:

$$CR < OM < CD < FA$$

The actual mean latencies are, in seconds:

$$CR = 7.38,\ OM = 10.24 \quad CD = 10.42,\ FA = 12.31$$

which is in the predicted rank order.

A possible explanation for the lack of an effect of event type on indices of discriminability can be found in the response latencies for the two types of event. The latency for the 200 event is consistently longer than for the 400 for all categories of response. In an unpaced situation it seems likely that the inspector is able to overcome the effects of a low signal to noise ratio by utilizing extra time to sample more attributes of the stimulus. Presumably the d' values which would be obtained from a short, fixed interval experiment, would be lower than those found in the self-paced situation.

The results for the effects of the experimental variables on the latencies of the various response categories cannot be simply accounted for, because of the reasons discussed at the beginning of this section. The large number of subject interactions found are likely to be a function of the differing scanning strategies adopted by the subjects, which tends to obscure the effects of the main variables. There is some evidence that the high white noise condition produces longer response latencies for some categories of response and this could be interpreted as evidence of a distraction effect. However, the effect

is small, difficult to isolate from the effects of other variables, and is unlikely to have any practical significance.

In summary, the most important information gained from the latency data is that it appears to provide further support for the application of the SDT model. The data are also useful in suggesting how in a self-paced situation, the subject may utilize extra sampling time to make the correct decisions for difficult discriminations.

4.13 Summary and general conclusions

The inspection task performed in the Data Analysis Group, University of Birmingham, has been considered from the standpoint of some of the variables considered in the review chapters of this thesis. An experimental study was performed to answer a number of practical and theoretical questions. The first of these concerned the applicability of the SDT model to the inspection task under consideration. The interrelationships between various performance measures was also investigated. Finally the effects on performance of several variables, of particular interest to the management of the inspection system, were analysed.

The evidence strongly suggested that the unequal variance SDT model provided a good fit to the experimental data obtained. It was found that in general the non-parametric measures of bias and sensitivity correlated well with their SDT counterparts, but that this correlation was not sufficiently close to produce the same results from detailed statistical analysis. A consideration of the relationship between SDT performance measures and other commonly used inspection measures showed

that the latter failed to adequately distinguish between changes of bias and sensitivity.

Having established that the SDT model was appropriate, the effects of the various experimental variables on the measures of bias and sensitivity was considered. The only significant effects obtained with d' and the non-parametric sensitivity measure of Pollack and Norman was between subjects. The criterion measure, log beta, was corrected for the effects of the assymetrical underlying variances, before being used in the analysis. A significant effect of time on log beta was obtained, the criterion decreasing with time on task. Various explanations were considered for this, and it was concluded that it was likely to be due to the experimental situation, the scanner initially utilizing a high criterion, because of the perceived payoffs of the various types of decision in that situation. As the experiment progressed the scanner utilized a criterion which was more in accord with the utilities found in the everyday task, as suggested by a prior survey of the inspection staff. No other significant effects were obtained with log beta.

Although the latency data was of secondary interest in this study, its analysis in general terms provided further support for the SDT model. A number of complex effects of the experimental variables on the latencies of the various categories of response were obtained. Interpretation of these was restricted because of the difficulty of separating search time and decision time. Considerable between-subject effects were found, probably as a result of differences in scanning strategies. Systematic investigation of the visual search aspects of this task would necessitate a two - phase experiment, in which the decision times would be evaluated independently of the search component.

Although it would have been possible to have conducted further analyses on the latency data in the form of distribution fitting, it was not felt that this would contribute to modelling the situation, because of the problems discussed earlier. The data provided general evidence that the simpler pattern produced an overall longer response latency, although this effect was complicated by subject differences. This result gave an explanation for the lack of any significant differences between the event types as measured by the SDT indices. It was hypothesized that in a self-paced situation the inspector was able to employ additional sampling of a near threshold potential defect, in order to make a decision about it. Self-pacing in inspection could be regarded as a means of enhancing sensitivity from this standpoint.

The conclusions for management were as follows. There was no concrete evidence of performance decrements as a result of the effects of auditory noise. The indirect evidence that the lack of differences found between the two types of event was due to the self-paced nature of the task suggested that no attempt be made to introduce pacing. The fact that there was no evidence for performance decrements with time during the experiment could not necessarily be extended to longer time periods.

In conclusion, the study demonstrated the practical feasibility of the SDT approach to inspection, and the unique insights that can only be obtained by this technique.

CHAPTER 5 CASE STUDY II : THE INSPECTION OF FILM IN THE QUALITY CONTROL DEPARTMENT OF ILFORD LTD.

5.0 INTRODUCTION

The industrial site for this study was Ilford Ltd., at Brentwood in Essex, a large manufacturing unit concerned with all aspects of photographic film manufacture, including Xray, ciné and roll film.

This study was particularly useful in that it served to emphasize the difficulties that can arise when attempting to apply some of the theoretical concepts that were discussed in the review chapters, to situations where the definition of acceptable quality is highly subjective. Although this study and the Data Analysis Group study were superficially related, in that they both involved film, in fact the nature of the inspection task itself was very different. Although the difficulties encountered in this study meant that an in-depth theoretical analysis of the results was not possible, many of the practical problems encountered provided an impetus for the experimental work described in later chapters.

The first part of the study involved a semi-structured interview with four of the senior film examiners and with the Quality Control Manager, Mr R F Salmon. The interviews were tape recorded and their content provided the basis for the general description and the task description which follows. Two experiments were then performed in an attempt to establish the basic performance parameters of the system and to investigate the utility of SDT in tasks of this type. The difficulties involved in the analysis of these tasks will be discussed in detail subsequently.

In spite of these difficulties, a report was presented to the firm which produced a positive response, and the contacts established at

Ilford proved to be very useful during the laboratory phase of the research discussed in subsequent chapters.

5.1 General description

The manufacture of photographic and X-ray film involves the coating of a plastic base material with a photosensitive emulsion. The coating machines which perform this operation operate in total darkness, because of the light sensitive nature of the coating, and produce large rolls of film approximately 44 inches in width. These are subsequently sliced up into a size appropriate for the use to which the film is to be put. The film is sold to a very wide market, with the emphasis on the commercial and industrial sector rather than the domestic field. Consequently the potential users have high quality standards, and quantities of film are sent back to the manufacturers if they are found unsatisfactory.

Quality control at Ilford is the responsibility of the Technological Services Division. The quality control section performs two functions. One is concerned with the maintenance of the physical parameters of film quality, such as coating thickness, grain size and base thickness, which are monitored by routine laboratory procedures. This aspect of quality control will not be considered in this study. The other function involves the visual inspection of film, which will be the main area of interest.

5.1.1 Visual inspection

Visual inspection takes place at two points in the manufacturing

process. The first of these is known as production testing. Samples of coated film in the form of large sheets are taken at an early stage of production, developed, and inspected by being placed over an illuminated table. The function of production testing is to provide rapid feedback if a malfunction in the coating machinery is producing a defective product. This section is run on a twenty-four hour basis in order to do this.

The other type of inspection is known as viewing. In the case of photographic film this involves taking samples of between 200 and 2000 feet in length, with widths of 16mm. or 35mm. from selected parts of the original roll. These are then exposed to give an even neutral density, developed, and sprocketed to produce cine films known as test references, or simply references. The references are projected on to a large screen in a cinema-like projection room. As the film is projected, a clockwork mechanism unwinds a tape in front of the inspector (or 'examiner') at a rate proportional to the rate of film transport through the projector. When a defect is observed, the inspector makes a mark on the tape indicating the nature of the fault. Measurement of the distance of the mark from the beginning of the tape enables the position of the defect to be subsequently located on the reference.

After a number of test references have been viewed in this way, any film which the examiner wishes to examine more closely is put on a device known as a viewing box, which enables the film to be wound at high speed to any position indicated by the tape as containing a defect. The film frames in this locality are then examined by transmitted light from an illuminated panel set in the front of the box, using a hand magnifier if necessary. This procedure, known as

'boxing' enables a detailed examination to be made of any part of the test reference of interest. Samples of film may be cut from the reference and sent to the laboratory for further analysis.

The report that is produced after examination, is usually qualitative in nature and generally consists of a description of the predominant defect characteristics, e.g. 'slight incidence of black spots but acceptable'. The overall philosophy of testing is to examine the film under the conditions that it will eventually be used. Thus ciné film is projected on to a large screen and 16mm. film on to a screen more representative of those used by amateurs. A numerical count of the defect is only made in detail if some problem exists which is associated with the occurrence of a particular defect, or if a count is required for statistical purposes.

The examiner has some prior information as to the potential defects which may occur on a reference. Each reference film can contains a 'coating card' which indicates whether any obvious problems occurred during the coating operation. Since many of the defects arise during the coating procedure, this gives the examiner some prior information as to the nature of the defects which may be expected.

Viewing is regarded as a very important operation, and normally no film is released on to the market until adequate screening has taken place.

5.1.2 Nature of the defects

A large number of configurations occur on film which can be categorized as defects. There are about 15 commonly occurring discrete

defects, which include items such as 'fibres', due to a foreign body being on the film when it was coated, and 'insensitive spots' due to the coating not adhering to the film base at some point. There are also 'continuous' defects, for example there may be a continuous slight variation in the film density. Apart from the more common defects there is a very large category of defects which occur occasionally, and even some which occur once and are never seen again.

The defects vary in size although they are generally very small compared with the area in which they appear. The angular size of an average defect (a 'fibre'), is about 1.44 minutes, and it has a contrast relative to the screen of approximately 10 to 1. If a defect occurs on a single film frame, as is often the case, the presentation time is of course extremely short, approximately 0.04 of a second for films being projected at the usual rate. The rate of occurrence of defects is extremely variable and may change suddenly at any time.

5.1.3 The definition of acceptable quality

The definition of what constitutes an acceptable film is highly subjective in nature. The ultimate arbiters of quality are the experienced film examiners, and any references which may be regarded as borderline will usually be referred to them. Inherent in the criterion of acceptability in a particular case, is the type of product and the customer for whom it is destined. In the case of cine film for example, the occurrence of a relatively large number of defects will be very readily apparent to the user, who will view a large sample of the film in a relatively short time. By contrast, with microfilm for example, the potential viewing population will be

small and the presence of a small defect even on each frame will not significantly degrade the quality of the film for the purposes for which it is intended.

5.1.4 Physical environment

The viewing area consisted of a large cinema-like projection room containing rows of seats. Around the periphery of the room were six viewing boxes with their associated winding machinery. There is usually a small 16mm. projector in use which projects the lower grade film on to a small screen on the sidewall. The main projectors were large, standard 35mm. cinema projectors.

The size of the main screen was 9'6" x 5'5" and the viewing distance 11'6", thus giving maximum visual angles of 79° and 58° respectively. The brightness of the screen as measured by an SEI photometer was 3.55 foot lamberts. Most of the time the ambient noise intensity was 60dBA. When two of the winding machines were being operated, the noise intensity went up to 84 dBA. Although the room was windowless, temperature and ventilation appeared to be adequate. The examiner viewing the main screen was generally seated in one of two large, deeply sprung armchairs provided for the purpose.

5.1.5 Selection and training

No formal procedures existed for selecting inspectors. There were difficulties in obtaining staff generally, partly because of the rather restricted long term career prospects. Most of the younger staff had been taken on to work specifically within the quality control area, whilst the more established examiners had generally

moved into quality control from the manufacturing side of the firm. None of the employees spent the whole of their time examining film, but most of them had at least one session a week.

No formal training programme existed. Most trainees sat with an experienced examiner whilst he was inspecting film until he was satisfied with their general level of competence. Many of the most experienced staff regretted the absence of a training scheme, but felt that there was at present insufficient time for a senior member of the staff to spend long periods with a trainee. It was felt that it took a considerable time to learn to examine film satisfactorily.

The educational background of the examiners was variable. Most of the older workers did not possess any formal qualifications, but some of the recent employees had G.C.E's.

5.2.1 Signal acquisition factors

It is clear that the acquisition of the signal is considerably more difficult than in the task considered in Chapter 4. The very short presentation time of the signal (0.04 seconds) means that the screen area cannot be searched and hence the probability of detection will depend on the size of the visual lobe relative to the size of the total screen area. The visual lobe size in turn depends on physiological factors such as peripheral visual acuity, and perceptual variables such as the ability of the inspector to distinguish between signal and noise. This latter skill is particularly important in film examining because of the very high level of noise stimuli that appears on the projected film. This is a result of dust particles

and other extraneous configurations put on to the film during its development. The perceptual skills required to distinguish between defects and non-defect stimuli take a considerable time to develop, and this is part of the reason for the length of time required to train an examiner. In SDT terms, increasing d' requires a thorough knowledge of the signal and non-signal characteristics.

From the standpoint of possible vigilance effects, the task characteristics would suggest that these are a distinct possibility. The signals are brief and are usually simple rather than complex. They often occur at a very low frequency in time and space, and the task is normally performed in an unstimulating environment. The decision process attending each response is usually simple. These conditions are in accord with those proposed by Kibler (1965) and Mackworth (1970) as giving rise to vigilance decrements. The length of time taken to examine the longest reference is 20 minutes, which is the minimum period at which vigilance effects would become evident. At the end of the 20 minutes there is usually a break of two or three minutes whilst a new film is loaded into the projector, before the next examining session begins. In this way, an examiner may inspect for several hours if there is a particularly urgent batch of film to examine. It seems possible that the short recovery periods between films might not allow performance to return to its original level, in which case a progressive decline over a number of sessions would be observed. This possibility suggested an experimental investigation, which will be described in a later section.

5.2.2 Decision making factors

Decision making factors are of considerable importance in this task. They operate at a micro and a macro level, affecting the decision making applied to an individual defect, and the global decision making employed to decide whether the test reference as a whole is acceptable or otherwise.

If we adopt the SDT paradigm, we would expect decision making at both levels to be influenced by prior probabilities and payoffs. At a defect detection level, the obvious effect would be that due to the expected incidence of defects. If the incidence of defects increased, it is likely that they would not necessarily be detected immediately because of the inappropriate criterion of the examiner. They could be interpreted, for example, as being due to extra dust particles on the film. In fact, examples of this effect were mentioned to the author during the interviews. In one case a particular type of defect had gradually increased in severity over a number of weeks. This was not detected until complaints were received from customers. When the films were re-examined with this new 'set' the presence of the defects was obvious. Considering the payoff matrix, at the defect detection level, false alarms could be regarded as relatively 'inexpensive' since they could always be checked at the boxing stage. This would promote a lax criterion.

The decision as to whether a test reference is acceptable or not is a global one which will depend only partly on the evidence gathered during screening. In addition to this evidence, which is in the form of the number of defects present and the overall appearance of the film, the final decision will also be influenced by prior probabilities

and payoffs, which may be different from those employed during the decisions concerning individual defects. The final acceptability decision will clearly be made partly in terms of the prior probability of the reference actually being unacceptable. This is normally a function of the particular film type, a new product being likely to have a higher a priori probability of being defective than an established one. The payoff matrix is a function of the likely effects of the various categories of decision. The probability of declaring a film acceptable will be influenced by the consequences of releasing a batch of film which is actually defective on to a particular market. Similarly, declaring a film unacceptable carries with it the costs of a possible false alarm, which include long delays whilst the batch of film is thoroughly investigated. There appears to have been no formal attempts to verify if the quality standards operated by the senior examiners are actually in accord with those required by the customer. The general feeling was that if complaints remained at a reasonably low level then quality was acceptable.

Another question which appears to have been neglected is the question of the drift of criteria of acceptability over time. We have seen from studies analysed in Chapter 3 (e.g. Thomas 1962) that in the absence of regular recalibration sessions, or absolute standards to refer to, then it is quite common for criteria to change over time to such a degree that they are no longer in accord with the professed quality standards of the inspection system.

A model of the film examination process in decision theory terms is given in Figure 5.1. It can be seen from the model that the actual film examination procedure produces an accumulation of evidence as to the number of defects and the general level of acceptability of the

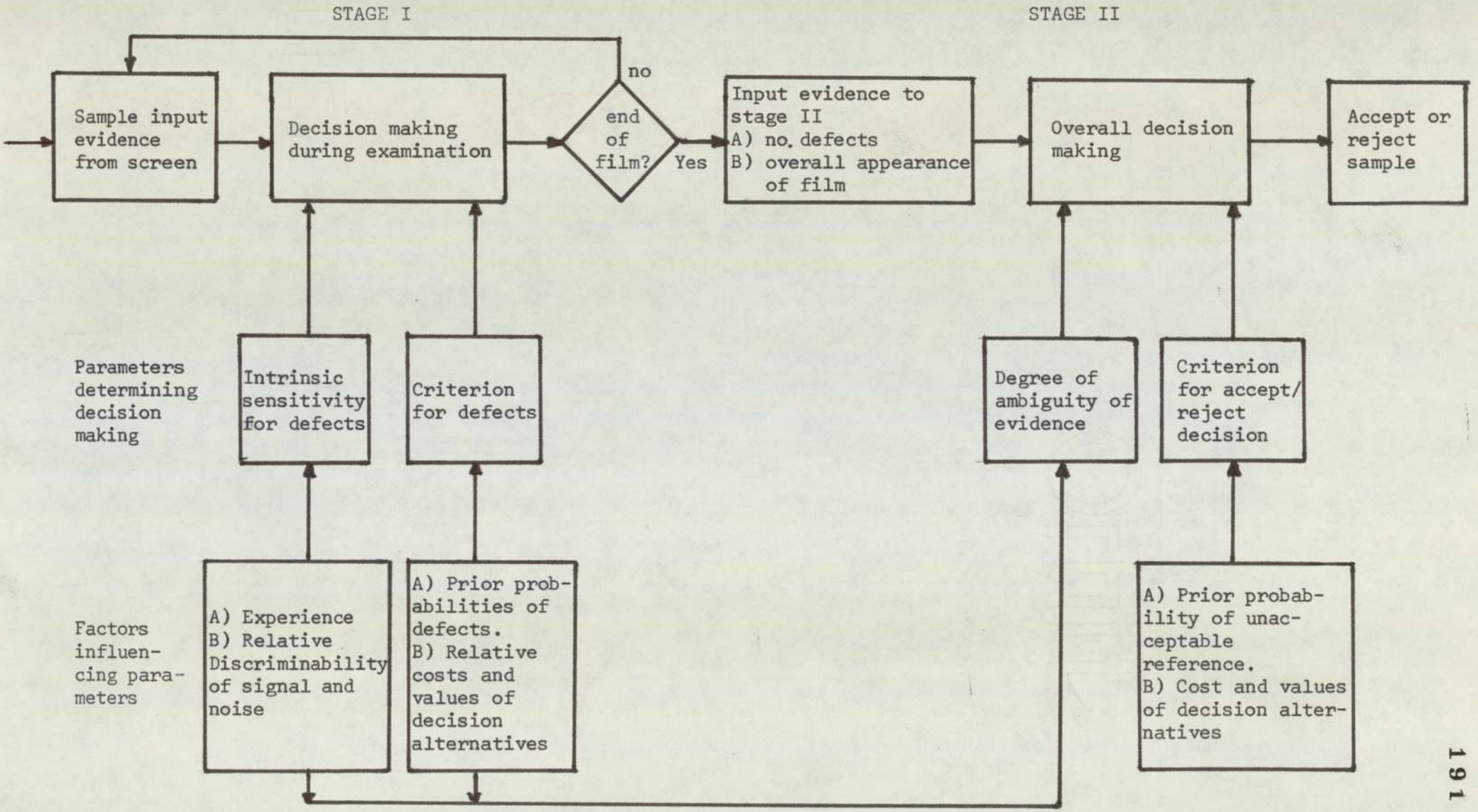


Figure 5.1 A decision making model of the film examination procedure

reference. The quality of this evidence will be influenced by the inspector's intrinsic sensitivity for defects and the criterial factors discussed earlier. If the inspector has no difficulty in distinguishing between signal and noise, (noise in this case could include other defect types if we are considering only one class of defects) then the evidence which is input to stage II will be highly reliable. During stage II, the examiner will have to decide, on the basis of the evidence from stage I, whether to accept or reject the sample. If we regard his sensitivity at this stage of the process as fixed, then his ability to decide between the two possible courses of action will be strongly influenced by the quality of evidence he receives. This is indicated in the model by the loop connecting the sensitivity and criterion factors of stage I to the degree of ambiguity of evidence box in stage II. In terms of SDT, the value of d' that the inspector is using at stage II is influenced primarily by the signal to noise ratio of the evidence. The payoff factors utilized at stage II discussed earlier, together with the prior probabilities of an acceptable or unacceptable sample, will determine the position of the criterion for the final decision.

If the examiner has a high d' at stage I, the evidence utilized at stage II will be truly representative of the actual state of the film, and hence the probability of an incorrect decision will be minimized.

5.2.3 Training

The SDT model proposed has various implications for training. The two stage nature of inspection suggests that any training programme will need to consider both phases. According to the SDT approach, training

should be aimed at two goals, enhancing the discriminatory powers of the inspector as much as possible, and modifying his response bias to the optimal position in a given situation.

Improving d' at stage I basically involves training the examiner to recognize the characteristics of defects as compared with spurious signals under the conditions that they appear on the screen, and also distinguishing between defect types. A consideration of the nature of the defects makes it obvious why training is so prolonged. The very short exposure time of the stimuli makes their initial detection difficult and hence any feedback from the trainer will only be effective on a relatively small proportion of the trials. Another difficulty concerns the lack of reference standards which would allow the trainee to gain insights into the appearance of the defects as they appear tachistoscopically on the screen. There is also an absence of a clearly defined nomenclature for the defect types. Some examiners gave examples of confusion that had arisen as a result of different inspectors using different names for the same defect. It is obvious that this situation is very confusing for a beginner. The training techniques which are appropriate for this situation have been reviewed in Chapter 3 and will be considered further in Chapter 7. The question of training an examiner to adopt an optimal criterion at stage I is a difficult one. An optimal criterion can be calculated from SDT from a knowledge of the a priori defect probability and the payoff matrix. The first difficulty involves the definition of the payoff matrix. Assigning numerical values to the various decision alternatives is particularly difficult in this case, where the results of the stage I decisions are used as input to the stage II decision making. Another difficulty concerns the a priori defect probability. As we have seen, it is pointless training an

inspector to rigidly adhere to a criterion appropriate for a particular defect incidence, if this is likely to change. The solution would seem to be to devise a training procedure which would allow the examiner to alter his criterion readily in the light of the prevailing defect incidence. Presumably this flexibility would also extend to changes of beta implied by changes in the payoff matrix. Although it may not be possible to assign precise values to the various payoffs, it should be possible to say in which direction the criterion should move in a particular case.

Considering the stage II decision making, it is clear that considerable experience is required to know what constitutes acceptable product for a particular market, particularly as the criterion is likely to be weighted by factors such as the urgency of a particular order and the possible cost penalties of not fulfilling it on time. Whether such complex utilities could be trained for explicitly seems doubtful. In any case, before a decision was made about a particularly important market, it is likely that confirmatory opinions would be sought from a number of experienced examiners, and further samples taken. In spite of this, there does not seem to be any reason why examples of acceptable quality film for various products and markets should not be utilized in training to give some indication of the required criteria.

It is not proposed to delineate a training programme at this stage. The review of perceptual training techniques in Chapter 3 would indicate that some form of KR or cuing technique would be an appropriate basis for training for sensitivity. The experimental studies in Chapters 6, 7 and 8 will provide further insights into this area.

5.2.4 Enhancement of the detectability of the defects

Another approach to improving the efficiency of the system, apart from training considerations, is to consider the ways in which defect detectability could be enhanced.

The most obvious way of doing this would be to slow down the rate of projection. This could only be done up to a point, since flicker would begin to be present at very low projection speeds. Nevertheless, it seemed reasonable to suppose that an effectively longer presentation time for each frame would effectively improve the detection probability. This proposal was put to management and met with an unenthusiastic response. The reason for this was that the proposal conflicted with the philosophy of viewing the films under the same conditions as they would be viewed by the customers. Even though more defects might be detected it was regarded as being more important to retain the fidelity of the inspection procedure to the conditions of eventual use. Although this attitude was certainly partly rooted in a desire not to disturb what was regarded as an effective system, there were other arguments that could be advanced to support it. The strongest of these was the fact that the senior examiners had established their criteria of what was acceptable for a particular market and product over a number of years. Any changes in the incidence of defects on the film, due to changed inspection techniques which were not related to true changes in quality, would require a complete recalibration of their standards in order to inspect to the same criteria as before. This emphasizes the fact that it is the integration of information over the entire examination of the film that is the important parameter of quality, not merely a count of defects.

Another suggestion made was to produce negative prints of the test references. This would mean that the defects would appear as light configurations on a dark background, rather than dark on light as at present. The proposed change would mean that afterimages of the defects would remain on the retina for a considerably longer period than the actual presentation time of the defect, thus considerably enhancing their detectability. Using fully dark adopted subjects, data from Alpern and Barr (1962), suggests that the after image would remain for an order of seconds. An additional bonus would be that spurious stimuli due to dust shadows would be absent.

Although this suggestion seems attractive, it was rejected on two grounds. The first of these was that, as discussed earlier, the test would be unrepresentative of its eventual use. Secondly there was the question of the time involved in the special processing of the film. There was constant pressure on the Quality Control Department from Sales to inspect the film references as quickly as possible so that the batch could be released for sale. It was felt that the extra processing time that would be required was not available.

5.2.5 Conclusions from ergonomics considerations

Film examining is a difficult discrimination task because of the short duration of the stimuli, the presence of a high level of visual noise and the highly subjective nature of the definition of acceptable quality. The present training technique would seem to be inadequate to produce trained examiners in a relatively short time, in the light of the difficulty of the task. A number of the techniques for training for perceptual skills discussed in Chapter 3

would seem to be relevant here. Consideration of the task from a decision making standpoint suggests that different types of training may be appropriate for different stages of the task.

It was felt that attention should be given to two aspects of the quality standards employed. The first of these was their degree of correspondence to those of the customers, and the second the possibility of drifts over time of these standards.

There seems to be a strong possibility of vigilance effects being important in this situation and it was felt that an investigation of these factors should be made experimentally. The visually demanding nature of the task also suggested that the visual skills of the inspectors should be examined.

These considerations formed the basis of the experimental studies described in the next section.

5.3 Experiment 2

The first experiment was intended to provide basic data on the overall efficiency of the inspectors and to investigate whether there were any performance decrements in time. Differences in performance between inspectors were also of interest.

5.3.1 Procedure

As the aim of the experiment was to gather data on the system under conditions as authentic as possible, six sample films were chosen by

the Quality Control Manager, which represented an average cross-section of the current work. Each of the six films was presented one after another, in randomized order, to each subject. The films were 2000 feet in length and each film lasted for twenty minutes. Reloading films meant that there was a gap of approximately three minutes between each film. The total experimental session lasted approximately $2\frac{1}{4}$hours. There were no breaks apart from the two or three minutes between film changes.

Each subject was given a set of printed instructions in which he was told to inspect the film in the usual way, noting any defects that occurred on the moving tape. To maintain authenticity, no attempt was made to keep the inspection room silent during the session, and hence there were several sessions in which extraneous noises and conversations occurred.

When all the experimental sessions had been completed, the tapes were collected from the subjects, and each film placed in turn on the viewing box. It was then examined very slowly by two senior inspectors, and a key chart prepared for each film, containing the locations and descriptions of each defect. The subjects' tapes were then examined and a similar chart prepared for each subject. Eight experienced examiners were used as subjects and they all had at least one year's experience and normal vision, according to a Snellen Chart.

5.3.2 Analysis of results

Considerable problems arose in scoring the experiment. Since the films contained a very wide variety of defects, it had been decided

to use one particular defect type, a 'fibre', to assess performance. After the experimental sessions had been completed and the key charts were being prepared, it was found that the overall density of defects differed considerably between some of the test films. The main reason for this was the very high density of 'coating bubble' defects on three of the test films. This led to a very high rate of response by the examiners and since it was impossible to resolve the responses marked on the tape to closer than 3 feet on the film, it was inevitable that the apparent detection efficiency on the high density films would be greater than that for the others.

A further difficulty was the variability in the numbers of fibres on the films, which ranged from two to twenty in number. It was therefore decided that the possibility of confounding together defects of differing discriminability was to be preferred to estimating probabilities with widely differing sample sizes. The performance index used was therefore the percentage of all non-coating bubble defects that were detected. These problems could have been avoided by preselecting the films to ensure approximately equal densities of a particular type of defect, but sufficient time was not available to do this. Additionally it would not have been representative of a typical selection of films. It was decided not to attempt to obtain measures of the incidence of false alarms, since the high incidence of coating bubbles made it impossible to tell if responses were in fact false alarms or correct detections of nearby coating bubbles. For this reason SDT measures could not be obtained.

5.3.3 Statistical analysis

The raw data consisted of the percentage detection efficiency for all non-coating bubble defects for each film examined by each subject.

These were analysed first as a two-way analysis of variance with repeated measures on subjects, the other variable being time intervals of twenty minutes, corresponding to individual films. The data were then de-randomized such that the factors were subjects and films and a similar analysis performed to detect any differences between films in the detectability of the defects they contained.

5.3.4 Results

Table 5.1 gives the percentage detection efficiencies for all defects excluding coating bubbles arranged in randomized order as presented in the experiment.

Subject/Time Interval	T1	T2	T3	T4	T5	T6	Mean detection efficiency for subject
B	72.0	40.0	55.0	32.25	26.30	82.50	51.34
E	47.4	55.0	30.0	69.5	53.0	80.0	55.81
A	92.0	65.0	96.0	74.0	70.0	47.4	74.06
G	52.0	69.50	26.3	64.00	23.5	35.0	45.04
C	70.5	60.0	78.0	31.5	88.0	35.0	60.50
H	52.0	52.0	15.8	48.0	30.0	29.5	37.88
F	35.0	16.0	65.0	36.0	26.3	53.0	38.54
D	44.0	26.0	26.3	20.0	17.6	20.0	25.64
Mean detection efficiency for time intervals.	58.11	47.93	49.05	46.9	41.83	47.8	Grand Mean 48.607

Table 5.1

Percentage detection efficiencies for all defects excluding coating bubbles arranged under the 6 test films used, i.e. de-randomized.

Subject/Test Film	F1	F2	F3	F4	F5	F6
B	55.0	26.3	40.0	32.25	72.0	82.5
E	55.0	47.4	30.0	53.0	80.0	69.5
A	70.0	47.4	92.0	65.0	96.0	74.0
G	35.0	26.3	52.0	23.5	64.0	69.5
C	35.0	31.5	60.0	70.5	88.0	78.0
H	30.0	15.8	52.0	29.5	52.0	48.0
F	35.0	26.3	16.0	53.0	36.0	65.0
D	20.0	26.3	44.0	17.6	20.0	26.0
Mean detection efficiency for each film	41.87	30.91	48.25	43.04	63.5	64.06

Table 5.2

Analysis of Variance for Table 5.1

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F	Sig.
T	1122.95068	5	224.59011	<1	N.S.
S	9629.47072	7	1375.63843	3.9	$p<0.05$
TS	12287.94338	35	351.08404		
TOTAL	23040.36333	47			

Table 5.3

Analysis of Variance for Table 5.2

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F	Sig.
F	6801.32423	5	1360.26489	3.9	$p<0.05$
S	9629.47072	7	1375.63843	3.9	$p<0.05$
FS	6609.57325	35	188.84494		
TOTAL	23040.36333	47			

Table 5.4

5.3.5 Discussion

The results indicate that there were significant differences between subjects in their detection efficiency, significant differences between the films in the detectability of the defects they contained but no significant effects of time on task on defect detection.

In view of the differing difficulty of the test films, the results cannot be taken as indicating unequivocally that there are no time effects on performance. These could have been obscured by the film differences.

The reasons behind the considerable inter-subject variability are worth considering in detail. Subsequent discussion with the examiners revealed that there were large differences in the type of experience they possessed. Although all of the inspectors had been examining for at least a year, the experience of some of them went back over 15 years. It was noticeable that the very experienced inspectors had a detailed knowledge of the mechanics of film manufacture and were able to utilize this to aid them in the inspection. For example, if a defect in the coating machinery caused a distinctive variation in the density of the film, the inspector would then be looking out for other defects associated with the malfunction. Other examiners were accustomed to inspecting mainly colour film, and hence were less familiar with the appearance of defects occurring on the monochrome films used in the experiment. A rigorous pre-selection of the subjects would have provided a more homogeneous sample, but would not have provided the realistic assessment of overall group efficiency that was required by management. Management were in fact surprised at the rather low overall detection probability for non-coating

bubble defects of 49%. Some possible reasons for this finding will be discussed in the conclusion.

5.4 Experiment 3

Experiment 2 had left several questions unanswered. The differences between films had made it difficult to detect any performance decrements with time, and the high incidence of coating bubbles had meant that false alarm rates could not be calculated. SDT measures could therefore not be applied.

For experiment 3, attempts were made to overcome these difficulties by obtaining four films which had an incidence of defects more typical of the norm. These films had also been selected on the basis that they were of equal difficulty and that they contained no distinguishing features that would lead to them being recognized as being different from the normal test references.

The object of the experiment was to test again for decrements in performance with time and also to obtain a wider range of performance measures than in experiment 2. It was also hoped to apply SDT measures to the data.

5.4.1 Procedure

An experimental session consisted of the alternation of the four test films with four normal references. This procedure was replicated a second time during another examining session to provide an increased estimate of experimental error. For ethical reasons, the examiners were told that test films had been introduced into the screening

session, but were not informed of how many films there were, and of their position. Each film lasted for twenty minutes, and hence they were viewed during the following time intervals from the beginning of the session:

test	1	20 - 40 minutes
"	2	60 - 80 "
"	3	100 -120 "
"	4	140 -160 "

Because of the interuptions in the normal flow of work necessitated by the experimental requirements, it was only possible to perform the study on three of the most experienced examiners. These were told to perform film examination in the normal way but to give the tapes produced, indicating the positions of the defects, to the investigator.

The results were scored in a similar manner to experiment 1. The positions of the defects as indicated by the tapes were compared with a key which had been prepared by a special examination of the films under optimal conditions. The detection of 'fibres' was again used to assess performance.

To investigate the visual acuity of the examiners in more detail, subjects were given the Bausch and Lomb Orthorater test for visual skills subsequent to the experiment. Only six of the original eight subjects agreed to take this test.

5.4.2 Statistical design

The experiment was analysed as a 3-way repeated measures analysis of variance, the four presentations of the film being regarded as sampling four time intervals, replications being the second factor and subjects the third. All effects were assumed fixed, by the same logic as was discussed in Chapter 4 section 4.10.3.

5.4.3 Results

Problems were again encountered when attempting to score the results. The main difficulty was the fact that it was impossible to be certain that a particular response was in fact a false alarm, when it occurred in close proximity to another non-fibre defect on the key chart. In fact virtually all false alarms appeared to be misidentifications of this type. This result would be expected if the subjects were able to discriminate readily between defects and spurious stimuli on the film, but were less sensitive to differences between categories of defects. On this assumption, the results can be analysed as a discrimination problem between the defect category under consideration ('fibres') and all other types of defect that occur on the film.

Where no apparent false alarms occurred, the approximation used in Chapter 4 was employed to obtain an estimate of false alarm probability.

Tables 5.5 and 5.6 give the correct detection and false alarm probabilities calculated as described earlier.

		TIMES				
Subjects	Replications	T1	T2	T3	T4	Subject means
S1	R1	0.455	0.273	0.636	0.909	0.568
	R2	0.364	0.455	0.818	0.636	
S2	R1	0.273	0.182	0.182	0.091	0.182
	R2	0.273	0.091	0.182	0.182	
S3	R1	0.182	0.455	0.273	0.455	0.398
	R2	0.455	0.636	0.455	0.273	
time means		0.333	0.348	0.424	0.424	grand mean 0.383

Table 5.5 Correct detection probabilities for Experiment 3

		TIMES				
Subjects	Replications	T1	T2	T3	T4	Subject means
S1	R1	0.083	0.056	0.083	0.083	0.108
	R2	0.056	0.250	0.167	0.083	
S2	R1	0.333	0.250	0.056	0.167	0.160
	R2	0.167	0.083	0.167	0.056	
S3	R1	0.056	0.056	0.083	0.056	0.070
	R2	0.083	0.056	0.083	0.083	
time means		0.130	0.125	0.107	0.088	grand mean 0.112

Table 5.6 False alarm probabilities for Experiment 3

The statistical Appendix A section 2, gives the analyses of variance for the arcsine transforms of the correct detection and false alarm probabilities in Tables 5.5 and 5.6. There are significant differences between subjects in correct detections ($p < 0.001$) but no other effects. All effects are non-significant for false alarms.

		Phoria			Acuity		Depth	Colour
		Vertical	Lateral	Both	Right	Left		
	Range of values from Orthorater tests	1 - 9	1 - 15	0 - 15	0 - 15	0 - 15	0 - 9	0 - 6
FAR	Subject A	5	8	11	9	9	7	1
	Subject B	9	15	11	8	12	8	6
	Subject D	5	11	11	9	8	4	6
	Subject E	7	7	7	7	10	5	6
	Subject F	5	11	10	9	8	5	6
	Subject H	5	5	8	8	8	3	5
NEAR	Subject A	5	7	7	5	7		
	Subject B	9	15	6	7	8		
	Subject D	4	2	12	10	11		
	Subject E	6	10	10	5	9		
	Subject F	6	7	10	8	9		
	Subject H	4	4	7	8	7		

Table 5.7 "Orthorater" results of visual screening of subjects

The efficiency indices A1 and A4 both confirm the significant subject effects for correct detections. The SDT parameters d' and log beta both showed significant differences between subjects ($p < 0.01$ and $P < 0.05$) which were reflected in the corresponding nonparametric Norman-Pollack and Hodos-Grier indices ($p < 0.01$ and $P < 0.05$ respectively).

The SDT indices need to be considered in the light of the fact that the plot of the z-scores for correct detection and false alarm probabilities shows no sign of being fitted by a straight line, thus bringing into doubt the validity of the SDT assumptions in this experiment.

The visual profiles from the Orthorater visual test are given in Table 5.7. An attempt was made to assess the adequacy of the visual skills using the templates provided with the Orthorater which defined minimum visual standards for various jobs. The 'visual inspection' profile provided, however, made no reference to the far visual skills required for film examination. Subjects F and D show the most adequate visual standards, their only deficiency being some degree of lateral phoria (an inability to fuse images from each eye). Subject H has both near and far visual deficiencies whilst subject A has below standard near vision and impaired colour vision. There was no correlation between these scores and performance on the task however.

5.4.4 Discussion

The salient feature of the results is the very low overall detection probability for defects obtained in this experiment. In view of the

fact that the subjects were three of the most experienced examiners, the overall detection probability of 0.38 needs to be accounted for.

The most likely explanation would seem to be in terms of the way in which the inspectors view the task of examining. As has been mentioned before, the prime function of examining is seen as globally establishing the acceptability or otherwise of a film rather than of obtaining a numerical count of defects, except for special purposes. For this reason the information on the tape will usually be more in the form of an aide memoire indicating the approximate disposition of defects rather than a precise record. Although in the experimental situation the subjects obviously made a more determined attempt than normally, to document the occurrence of defects on the tape, it must be recognized that the act of noting down every defect would be unfamiliar to them. This explanation is confirmed to some extent by considering the detection results for experiments 2 and 3. Experiment 2 was far more of a special situation where the examiners could concentrate on attempting to note and detect defects, knowing that none of the sample films was actually a 'real' reference which had to be assessed for acceptability. In the case of experiment 3, they would not be certain which of the references was a test film, and would therefore have to treat all of them in the same manner. This would involve concentrating on the more general aspects of film quality to the detriment of noting down specific defects. These contrasting strategies are reflected in the higher apparent detection probabilities of experiment 2 compared with experiment 3.

In view of these considerations, it is not surprising that SDT does not appear to fit the data according to the Z-ROC curve. From the

preceding discussion it is clear that from the data available it is not possible to decide unambiguously whether or not SDT applies in this situation. However, it is interesting to note that the significant differences in log beta between the subjects reflects the appropriate changes in correct detections and false alarms that would occur if the theory did apply, i.e. decreasing log beta is accompanied by increases in correct detection and false alarms. Log beta will always provide an index of such changes, whether or not SDT applies.

With regard to the changes in performance over time, no significant effects were found. It is notable, however, that the changes in log beta over time intervals, although not statistically significant, showed a clear decline similar to that observed in experiment 1 in the last chapter. The absence of significant vigilance effects can be ascribed partly to the short breaks between each film and partly to the wide variety of non-target stimuli that occurred during the session, which can be regarded as maintaining arousal.

5.5 Conclusions

This study exemplifies the difficulty of using the SDT approach in situations where obtaining accurate estimates of correct detection and false alarm probabilities is difficult, and in which the standards for acceptability are not rigidly defined. In spite of these difficulties, the two stage decision model described would seem to be applicable in this and similar situations where an accumulation of evidence over time is utilized to make an accept/reject decision.

The practical difficulties of scoring the data for the experiments described in this chapter made a valid test of the applicability of SDT at the defect detection stage impracticable. Although the results quoted do not provide direct support for SDT, this is not surprising in view of the small sample size and the rather gross assumptions made in scoring the data. The changes in correct detections and false alarms over time, although not statistically significant, are consistent with a criterion change in a more lax direction, as observed in experiment 1, Chapter 4.

Another experiment was planned to test the stage II decision making part of the model, by asking the inspectors to rate a number of films in terms of their degree of acceptability, from completely acceptable, to completely unacceptable. This would enable the generation of ROC curves and provide a more meaningful test of the stage II part of the model than a yes-no experiment. Unfortunately sales and production pressures meant that the experiment had to be deferred indefinitely.

No support was found for the hypothesis that there would be performance decrements with time on the task, despite subjective reports of occasional sleepiness by some of the examiners when normally performing the task. The suggestion was made that the variety of stimuli which occur on the film provide sufficient stimulation to maintain arousal.

The investigation provided the impetus for several lines of research which are described in subsequent chapters. The first of these was the possibility of training the inspector to be able to modify his criterion readily in the face of changing defect rates or payoffs.

The study suggested a need for this ability at both stage I and stage II of the decision making process. The need to provide a rapid means of training the examiner to discriminate between defects and noise stimuli, and between different categories of defects, was also identified as being an important research area. Finally the absence of any effective selection procedures suggested the possibility of work on this topic.

The analysis of the inspection system in terms of some of the dimensions proposed in Chapter 3 produced a number of specific recommendations for management.

It was suggested that a library comprising examples of the various types of defect be established, which could be referred to readily. Similarly a proposal was made to keep a range of film reference examples which provided clear examples of the quality standards required for particular markets and products. It was felt that a uniform system of nomenclature for both types of defects and levels of acceptability would further aid the establishment of clear quality standards.

In conjunction with these latter suggestions, the practice of regular calibration sessions was proposed as a means of ensuring that quality criteria, especially those defined largely subjectively, did not drift over time and that there was a high degree of concordance between inspectors. The relationship between customer standards and inspection criteria was felt to be an important one. Investigation of this relationship would, however, require an extensive market research exercise.

Returning to the training area, it was suggested that the results of further research be fed back to Ilford to aid in the specification of a training scheme. In addition to the perceptual skills discussed earlier, it was felt that information on film manufacture should be given to trainees, in view of its apparent importance in the identification of defect types.

Finally there was a recommendation that all potential film examiners should receive some form of industrial eye test such as the Orthorater. Although there was no apparent correlation between visual abilities and performance on the task, the examiners with the most impaired vision were in fact the most experienced. It was likely therefore that this would provide compensation for any visual deficiencies. Adequate visual skills were felt to be particularly important during the training stage.

CHAPTER 6 INVESTIGATIONS INTO THE EFFECTS OF DEFECT PROBABILITY CHANGES ON INSPECTION PERFORMANCE

6.0 INTRODUCTION

The effects of changes in the probability of defects on inspection performance has been referred to at several points as being an important consideration in inspection situations found in industry. An example has been quoted in Chapter 5, where quality levels of a product became unacceptable because inspectors were unable to modify their criteria to take into account the changes in defect probability that had occurred.

In this chapter, a brief survey will first be made of some of the relevant theoretical approaches. Two experiments will then be described in which some of the factors affecting performance in the changing defect density situation were investigated.

6.1 Theoretical considerations

The modification of performance in a changing defect probability situation can be seen as consisting of two aspects: perception of the fact that the probability has in fact changed, and the adaptation of the inspector's response strategy to the new conditions. We shall consider these two areas separately in the following discussion.

6.2.1 Probability learning

A large literature exists on the ability of subjects to learn probabilities. The classical probability learning situation differs, however, from that which occurs in inspection. Usually the subject is asked to predict which of two mutually exclusive events will occur, or

has occurred, on each trial of a series in which two events occur with fixed, but unequal, probabilities. For example, the subject may be asked to predict which of two lights will blink on at each trial of a series. He is allowed to witness the outcome of each event and is therefore completely reinforced. He may, in addition, be given monetary rewards for correct predictions and/or be penalized in some manner for incorrect performance.

A number of theories exist for analysing behaviour in these situations. One of the earliest, from game theory considerations, was that after a few trials the observer will distinguish the more frequent event, and predict it on all succeeding trials, thus maximizing his total number of correct predictions. This 'pure strategy' occurs very infrequently. Another theory, called the probability matching hypothesis, predicts that the subject will learn to match his response ratios to the probabilities of occurrence of the two events. This behaviour was first noted by Grant, Hake and Hornseth (1951). They observed that over a series of trials in which two alternative reinforced events occurred with fixed probabilities, the subject's probability of predicting a given event tended to approach or "match" the actual probability of the event. A large number of other studies, e.g. Bush and Mosteller (1955), subsequently showed that the subject begins by predicting the two events equally often, then after a slow initial rise he comes to predict the two events with the actual probabilities of their occurrence. Later work by Edwards (1961) produced evidence suggesting, however, that if a large number of trials is given, i.e. > 300, then the subjective probability estimate becomes more extreme than the actual probability.

Siegel and Goldstein (1959) produced an analysis of probability learning from decision theory considerations which has obvious affinities with SDT. They suggested that the maximization of subjective expected utility will account for both probability matching and pure strategy (i.e. always responding the most likely alternative) behaviour. Since the utility of an outcome is its subjective value to the subject, his behaviour will be a function of the particular conditions and reinforcements inherent in the task. The general hypothesis is that he will maximize his subjective utility in whatever terms he perceives it.

Siegel and Goldstein suggest that in a task with no external payoffs, (monetary rewards etc.), the subject receives satisfaction from predicting and having confirmed the rarer of the two events. This might be expected to relieve the monotony of the pure strategy of predicting the more frequent event on all trials. For whatever reason, it is hypothesized that the subject will adopt a mixed strategy which approximates to the matching hypothesis.

If however, the subject receives some monetary reward for a correct prediction, or a penalty for an incorrect decision, the theory hypothesizes that as these rewards are increased, the subjects prediction of the more frequent event will tend towards 100% i.e. will approach a pure strategy. Experimental evidence confirmed these predictions.

A number of other factors have been found to affect probability learning. These are reviewed in Lee (1971), but will not be discussed in detail here because of their limited relevance to inspection.

It is clear that there are affinities between the inspection situation and research on probability learning. The main difference is in the nature of the reinforcement received by the inspector. In the probability learning task, the subject is usually completely reinforced, whereas the inspector receives only partial reinforcement, depending on such factors as his ability to recognize defects. The occurrence of inspection error leads to the inspector having an incomplete knowledge of the true proportion of defects. A probability learning study by Estes and Johns (1958), involving ambiguity in the reinforcement, similar to that occurring in the inspection task, found reasonably close agreement between the frequency an event was predicted, and the frequency with which it was judged to have occurred.

It seems clear then, that the most fundamental difference between the probability learning situations usually studied and the inspection task is the amount of information the subject receives concerning the nature of the overall defect probability. In the probability learning task, the only information available is that obtained from the responses. On the other hand, this information is usually complete. In inspection tasks, the evidence from responses is partial, but there may be other sources of information available, as will be discussed in the next section.

6.2.2 Sources of information on defect probability

Two sources of information are available to the inspector, from which he is able to estimate the prevailing defect probability. These are external sources and evidence inferred from the task itself.

The first type of external evidence could be described as 'feed-forward' and consists of prior knowledge of the defect frequency, acquired either as a result of experience or from some other source. Examples of such information sources have already been quoted in Chapter 5, where film inspectors are provided with a card detailing the manufacturing history of a particular batch of film. In the glass industry, inspectors of float glass are given prior warning that the incidence of defects may change by the furnace operators working upstream from the inspection point (Gillies (1975)). The other form of external information is feedback, or knowledge of results concerning his accuracy, that the inspector receives subsequent to the actual inspection of the sample. In most real inspection situations such feedback is delayed and incomplete, and cannot readily be related to the separate decisions made during the task.

The most accessible source of evidence available to enable the inspector to modify his subjective probability estimate is from the task itself. The accuracy of the subjective probability estimate will depend primarily on the inspector's ability to distinguish signal from noise, and to a lesser extent on the appropriateness of his current criterion to the prevailing defect density. A dynamic situation exists, in that the more optimally the criterion is set, in relation to the actual defect density, the more accurate will be the information available to the inspector to modify its position still closer to the optimum. It should be noted that these considerations are strictly only of importance in a relatively low d' situation. Where the signal to noise ratio is high, changes in defect probability can be expected to have a relatively slight effect on performance. On the other hand most real inspection

situations involve discriminations between perfect and defective items which would normally be regarded as 'difficult'.

6.2.3 Modification of the subjective probability estimate and the criterion

In the preceding sections, we analysed the sources of evidence available to the inspector concerning the fault density. In considering how this information will be utilized by the inspector to modify his criterion we have to examine several further questions. First, it is necessary to know the extent to which the inspector will utilize the information to revise his subjective probability to correspond to the actual defect probability. Secondly there is the question of how large a sample of the new probability he needs (in a situation of change) in order to effect this change. Finally, we need to consider the ability of the subject to modify his criterion to the appropriate optimal position on the basis of his revised subjective probability.

Data on the first question is available from a number of studies in the area of the revision of subjective probabilities, e.g. Edwards (1962), Edwards and Phillips (1964), Stael von Holstein (1971). In general these studies suggest that observers do not revise their subjective probabilities to the optimal extent that Bayes' theorem would predict. There is very little evidence on the size of the sample needed by a subject to modify his subjective probability as a result of objective probability changes. The sample size might be expected to depend strongly on the discriminability of the signal employed. Nearly all SDT experiments have employed well practiced subjects and have used a fixed within-session probability. Within

these constraints, subjects have shown that they can use the appropriate beta value corresponding to the a priori probability (Swets and Green, (1966)).

One study does exist which investigated subjective probability considerations in an inspection context. Sims (1972) used a simulated inspection task in which subjects had to inspect printed facsimiles of printed circuit boards containing a range of defects. In the first experiment, prior to examining each item, the subject had to record whether or not he felt that it was going to be defective or not. Having made their predictions, the inspectors then examined the items and assigned them to accept or reject baskets. Three percentages of defects were used, i.e. 26, 14 and 2 percent and the subjects were either told the incidence in advance of defects or given no information at all. In the second experiment, the subjects were told that the defect level was 2%, but unknown to them there was a step increase in the defect rate to 14% midway through the session.

Unfortunately, the experiment was analysed purely in terms of the inspector's perception of the defect probability, and no detection data were given.

Considering the subjective probability estimates however, it was found that the subjects' estimates were significantly different for the differing probability levels within 30 trials of the start of each session. As expected, these estimates were in general more conservative than the actual probabilities. It was found that prior knowledge of the probability did not affect the accuracy of the subjective estimates. The results also suggested that the final

subjective estimates of probability were more closely related to the proportion of items classified as defective, than to the actual probabilities. The probability change experiment produced anomalous results. Only two subjects were employed, and in one case the subject accurately changed his subjective estimate of probability to the new probability level within sixty trials. The other subject however, did not show any significant changes. Unfortunately no investigation was made of possible differences in sensitivity which could have accounted for this finding. The result that subjects match their probability to the perceived incidence of defects is one which might have been predicted on intuitive grounds. The lack of any effect of prior knowledge is a surprising one and suggests that the subjects weighted intrinsic information from the task more than prior evidence. This could of course be a function of the degree of difficulty of the task. Presumably the harder the task, the more reliance would be placed on prior evidence.

6.3 Experimental objectives

In view of the foregoing discussion it is clear that there are several questions in the general area of detection performance under changing probability conditions that need to be investigated.

The first of these is the general question of the ability of an inspector to adjust his criterion to the on-going defect probability. The other considerations concern the efficacy of knowledge of results and prior warning of a defect change in aiding this criterion change. Although the Sims study suggests that the inspector can correctly estimate the new probability, this does not guarantee that

this will lead to the appropriate criterion being adopted. Another consideration was whether the inspector's performance strategy was different if the change occurred during a session, rather than from the commencement of the session. Two experiments were conducted to investigate these questions, and will be described in detail in subsequent sections.

6.4 Experimental design: general

It was decided to employ an experimental design which allowed the accurate evaluation of SDT parameters. The fact that a minimum of 500 trials are recommended to do this (Swets and Green (1966)) and that a number of experimental conditions were to be investigated, meant that the number of subjects that could be included in the study was limited by time constraints, particularly as it was felt that extensive practice at the task was necessary in order to minimize learning effects. In view of the exploratory nature of the study, it was felt that three subjects would produce meaningful results.

In order to facilitate obtaining SDT parameters, the experiment was performed using a rating scale approach (McNicol (1972), p.99) as will be described in detail subsequently.

The first experiment was designed to investigate the situation where the defect probability remained constant throughout a session, but varied between sessions. The presence or absence of feedback on performance was considered, and performance changes between blocks within the session were also included in the analysis.

The second experiment was concerned with within-session changes in defect probability. The probability of defect occurrence changed after the second block within each session. In the first session, no warning was given that a change was going to take place. In the remaining sessions, subjects were warned that a probability change would occur. In the one case they were given summary feedback every hundred trials, whereas in the other case no feedback was provided.

The conditions investigated in the two experiments are summarized as follows:

Experiment 4

1. Constant high probability of signal (p = 0.5)

a) Feedback

b) No feedback

2. Constant low probability of signal (p = 0.1)

a) Feedback

b) No feedback

Experiment 5

Change in probability during session (p = 0.5, 0.2)

a) No warning, feedback

b) Warning, no feedback

c) Warning, feedback

Five hundred trials were given under each condition and the order of presentation of the signals was randomized, subject to the constraint that the probabilities were constant within each block of 100 trials.

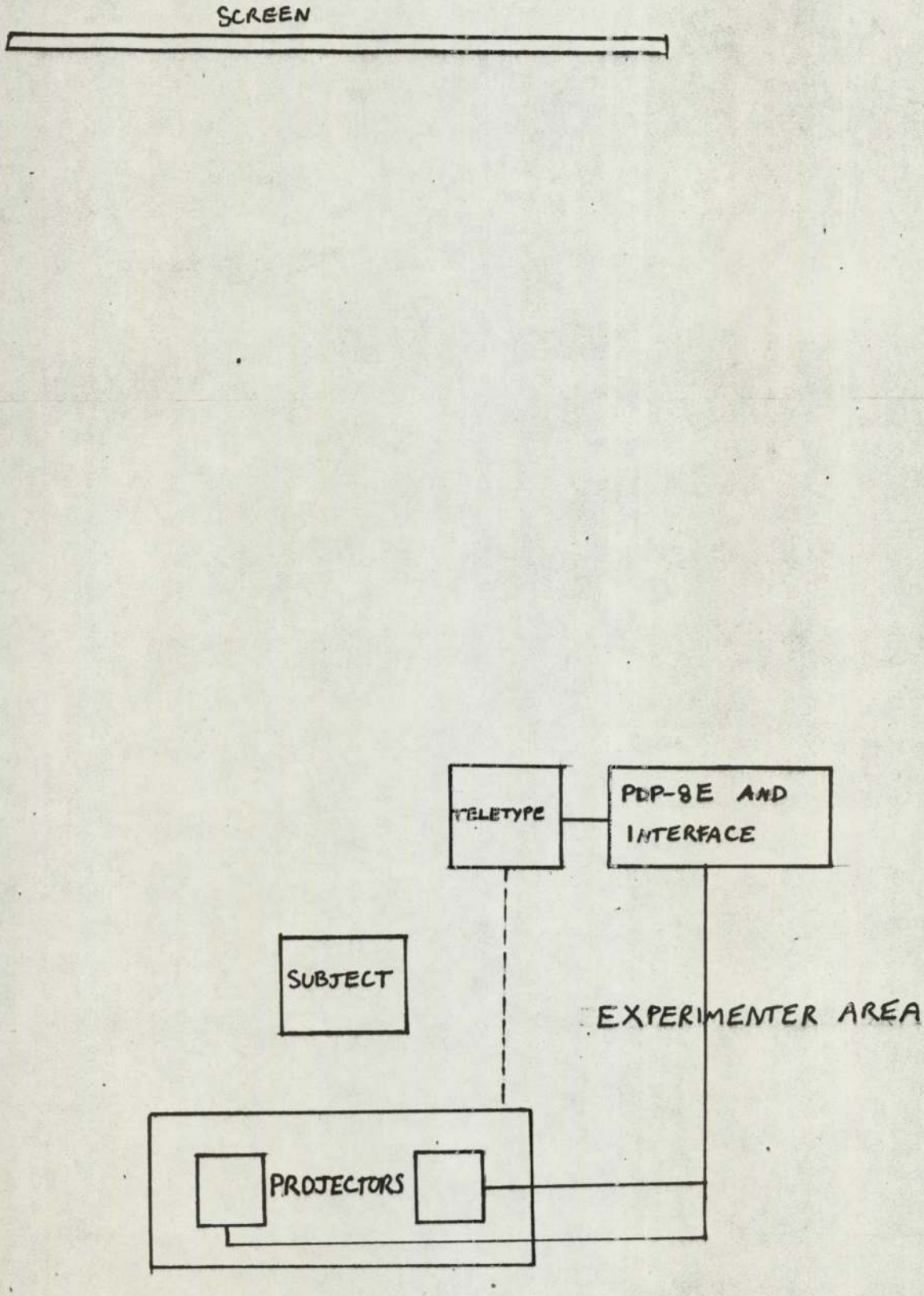
In view of the fact that vigilance effects would complicate the interpretation of the results, the experiment was designed specifically to minimize these effects. This was achieved by providing short rest periods every 100 responses, which meant that the subjects did not perform the task continuously for longer than about eight minutes at a time.

6.4.1 Apparatus

It was decided to build a general experimental facility that could be utilized for a wide variety of detection experiments in addition to those described in this thesis. The equipment was based on a PDP-8E computer and the stimulus material was projected by means of a pair of Kodak Carousel projectors, on to a screen in the darkened experimental room.

Essentially the requirement was for a device that could present signal and non-signal stimuli in a random sequence, would accept a series of responses corresponding to a rating scale and would provide various forms of knowledge of results for these and later experiments. It was also felt desirable to record the latency between the presentation of the stimulus and the response. The computer controlled the sequencing of stimuli, recorded the responses, and provided the feedback. A wide range of other facilities were also available which are described in detail in Appendix C . The equipment represents a highly general means of conducting a wide range of detection experiments. However, its design and construction, and the programming of the PDP-8E computer, which was the entire responsibility of the author, consumed a very considerable amount of

FIGURE 6.1 EXPERIMENTAL ARRANGEMENT



time, and in retrospect it is felt that a simpler if less elegant means of conducting the experiments would have been more appropriate.

The general experimental layout is shown in Figure 6.1. The first projector produced a constant display on the screen, consisting of a matrix of 9 equally spaced black discs, each 5 inches in size, giving an angular size of 2.6° from the table where the subject was seated, 10.5 feet from the screen. The second projector contained alternate opaque slides and slides with a small hole punched in them. When one of the latter slides was in the projector, the resulting pencil of light superimposed exactly on the central disc of the display. Filters were used with both projectors such that the increase in brightness produced was close to the threshold of detectability for each subject. A Compur magnetic shutter,which was controlled by the computer, operated to give a brief presentation of the stimulus. A teletype in the experimental area read in a steering tape which caused the second projector to either remain on the current slide, or to advance to the new slide, prior to the next operation of the shutter. This arrangement allowed a random sequence of any number of stimuli to be presented from a single Carousel magazine of 80 slides.

6.4.2 Procedure

At the beginning of the experiment the projector advanced to the first slide and the shutter opened for 0.25 seconds. This interval was chosen to approximate to the time of a single eye fixation, since it was originally intended to utilize the results in the analysis of subsequent search experiments. The subject then pressed

one of the response buttons and the response latency and the channel number of the response were output on paper tape by the computer. After a delay of three seconds from the response, the next stimulus was presented.

The six response buttons corresponded to six degrees of confidence that a signal had been presented on the preceding trial. From left to right, the ratings were: definitely non-signal, probably non-signal, possibly non-signal, possibly signal, probably signal, definitely signal. A symmetrical pay-off matrix was assigned to the responses in the following manner. Correct positive or negative responses were given 3, 2 or 1 points, depending on whether they were made at an extreme, intermediate, or low level of confidence. Incorrect responses gave minus these payoffs. Subjects were given 750 points at the beginning of the session and paid a bonus if they scored a further 750 points from their responses, i.e. if they scored half the total possible number of points from the 500 responses.

After each block of 100 responses the number of correct detections, correct rejections, false alarms and missed signals for that block was printed out on the teletype, together with the points scored, and the total cumulative number of points including the earlier blocks in that session. For the feedback conditions, the teletype was situated so that the subject could read this printout. A three minute break occurred between each block of 100 responses to allow the subject to do this. During non-feedback sessions the same length of break occurred. With both types of session, the experimenter made non-committal conversation with the subject during the break.

White noise was played at a low level through headphones for the whole of the session, apart from rest periods, to mask the sound of the shutter and projector operation. As the experiment was self-paced, its duration varied slightly, but most sessions were completed within about 45 minutes.

Prior to the main experimental sessions, all subjects had received at least 2000 practice trials using the same apparatus and stimuli. During the practice sessions, feedback similar to that administered during the test sessions was given, and the defect probability remained constant at 0.5. The last practice session had taken place one week before the first test session.

During the practice periods the nature of the scoring system was carefully explained to the subjects, with emphasis on the fact that both missed signals and false alarms would be equally penalized. The use of the rating scale response buttons was discussed, and any difficulties over what was meant by, for example, a 'probably signal' response were resolved. Immediately prior to each test session the subjects were told that the defect probability would be 'the same as during the practice sessions', or simply that it would be 'low', as appropriate. During the first session of the within-session probability change experiment, where no warning was given, the subjects were told that the defect probability would be the same as during the practice sessions. For the two subsequent probability change sessions, they were simply told that a change would occur at some point in the session. Those subjects who had not inferred that a change had taken place during the first session were explicitly informed of this immediately after the completion of the session.

All subjects received condition 1a first, which was identical to the practice sessions and was intended as a 'warmup'. The remaining conditions of the first experiment were presented in a random order, followed by the second experiment. Each subject performed one session per day. Attempts were made as far as possible to ensure that the subjects performed the task at the same time each day, but the exigences of lecture timetables meant that this was not always possible.

All subjects were final year undergraduates and were paid 50p per session, with a 25p bonus if they achieved the target score.

6.4.3 Statistical design

Experiment 4 was analysed as a 4-way factorial analysis of variance, the factors being blocks of 100 responses, presence or absence of feedback, signal probability of 0.5 or 0.1, and subjects. Experiment 5 was set out as a three way analysis of variance, with blocks of 100 responses, the three experimental conditions and subjects being the factors.

In both experiments all subjects received all combinations of conditions. As discussed in Chapter 4, the danger of carry-over effects exists in this type of design. In the present experiments, these effects were minimized by having a highly trained subject population and by including the blocks within the experiment as one of the specific factors. In view of the laboratory based nature of the experiments, a mixed model was utilized in the analysis of variance, with subjects being assumed a random effect and all other factors fixed.

6.4.4 Analysis of the results

The latency and response data which had been produced on paper tape by the PDP-8E computer, were fed into the PDP-15 computer for analysis. The responses were compared with the actual signal sequence and a number of performance measures produced.

The use of rating scale data allowed some performance indices to be calculated which had not been available for the yes-no experiments analysed previously. These indices have been discussed in detail in Chapter 2. They were the Altham-Hammerton sensitivity index, and P(A), the area under the ROC curve, which provides a sensitivity index independent of the underlying variances. The latter index was calculated using a simple numerical integration technique (McNicol (1972) p.115). Another quantity calculated was the score achieved by the subject obtained from the rating responses made and the payoff matrix. The remaining performance indices were those employed in previous experiments. These were the latencies for the various types of response, the false alarm and correct detection probabilities, d' and beta, and the corresponding nonparametric indices. The rating scale approach allowed the Grey-Morgan program to be utilized to fit ROC curves to each block of data. This in turn allowed correction of the beta values to take into account the ratio of the variances, and the calculation of d_e', the corrected value of d' discussed in Chapter 2. In most cases the unequal variance SDT model produced an acceptable fit to the rating data. In a small proportion of cases, an ROC curve could not be fitted because the subjects had used only the extreme response categories, effectively producing a single point on the ROC curve. In these instances, the slope for the ROC curve

fitted to the whole 500 points for the entire session was utilized. The mean ratio of signal to noise variance was 2.564, and there were no significant differences in the sigma ratio between experimental conditions.

6.5 Results - experiment 4

The analyses of variance are given in Appendix A and the raw data and condition means in Appendix B.

6.5.1 Signal detection results

The correct detection and false alarm probabilities for the five blocks of responses within the four experimental conditions of $P = 0.5$ and $P = 0.1$, feedback present or absent, are given in Figures 6.2 to 6.5.

The analysis of variance for the arcsine transforms of the two probabilities indicates significant differences between blocks and between probability conditions for the false alarm probability ($p < 0.05$ in each case) and significant differences between feedback conditions for the correct detection probability. Applying the Tukey multiple comparison test to the false alarm probability block means, indicates that this is significantly higher for the first block compared with the remaining four blocks ($p < 0.05$). With regard to the probability condition means the false alarm rate is significantly higher for the $P = 0.5$ condition than the $P = 0.1$ condition.

The analysis further indicates that correct detection probability is significantly higher under the feedback condition compared with the non-feedback situation ($p < 0.05$).

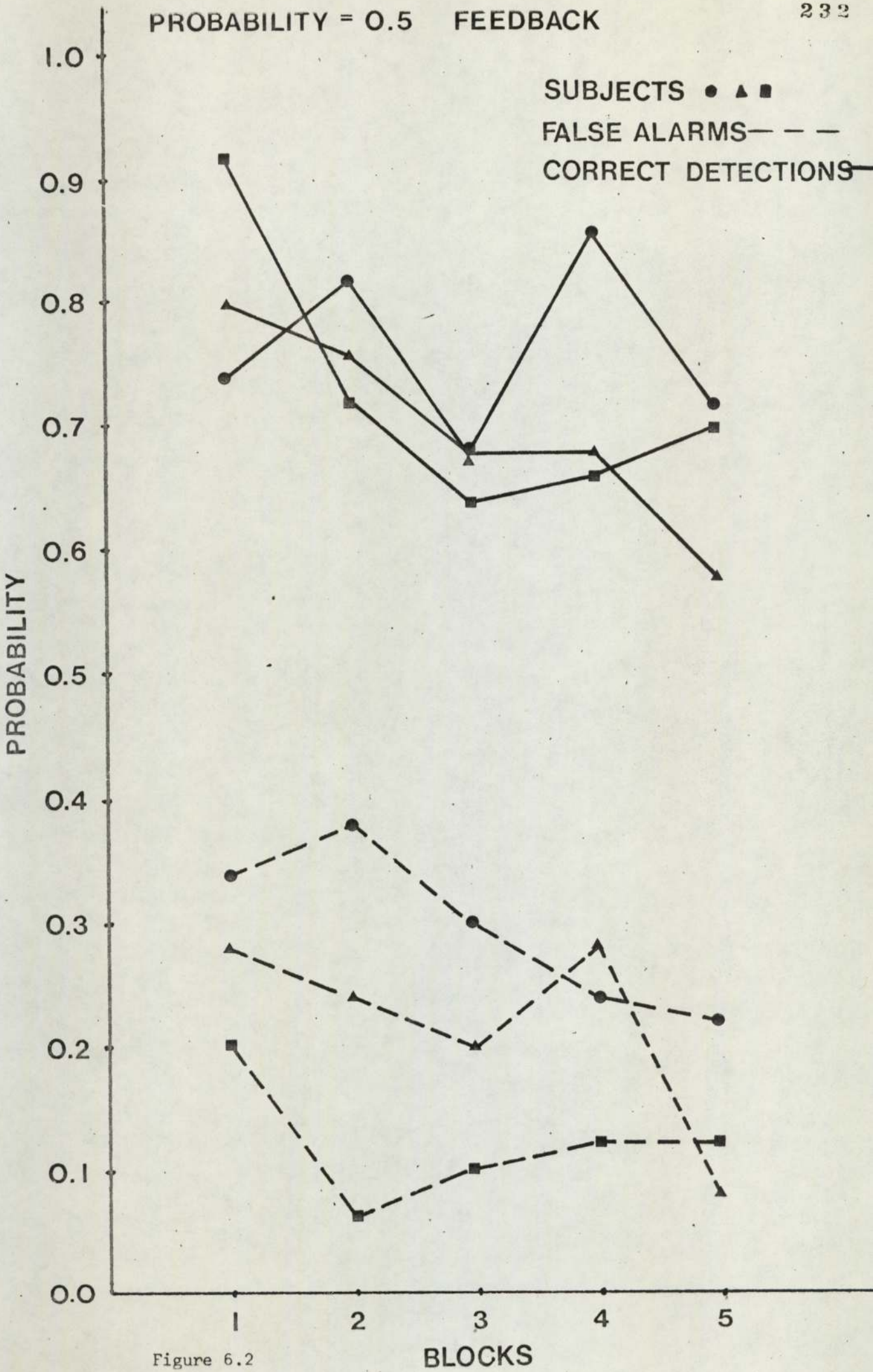

PROBABILITY = 0.5 FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS— — —
CORRECT DETECTIONS
PROBABILITY
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
BLOCKS

Figure 6.2

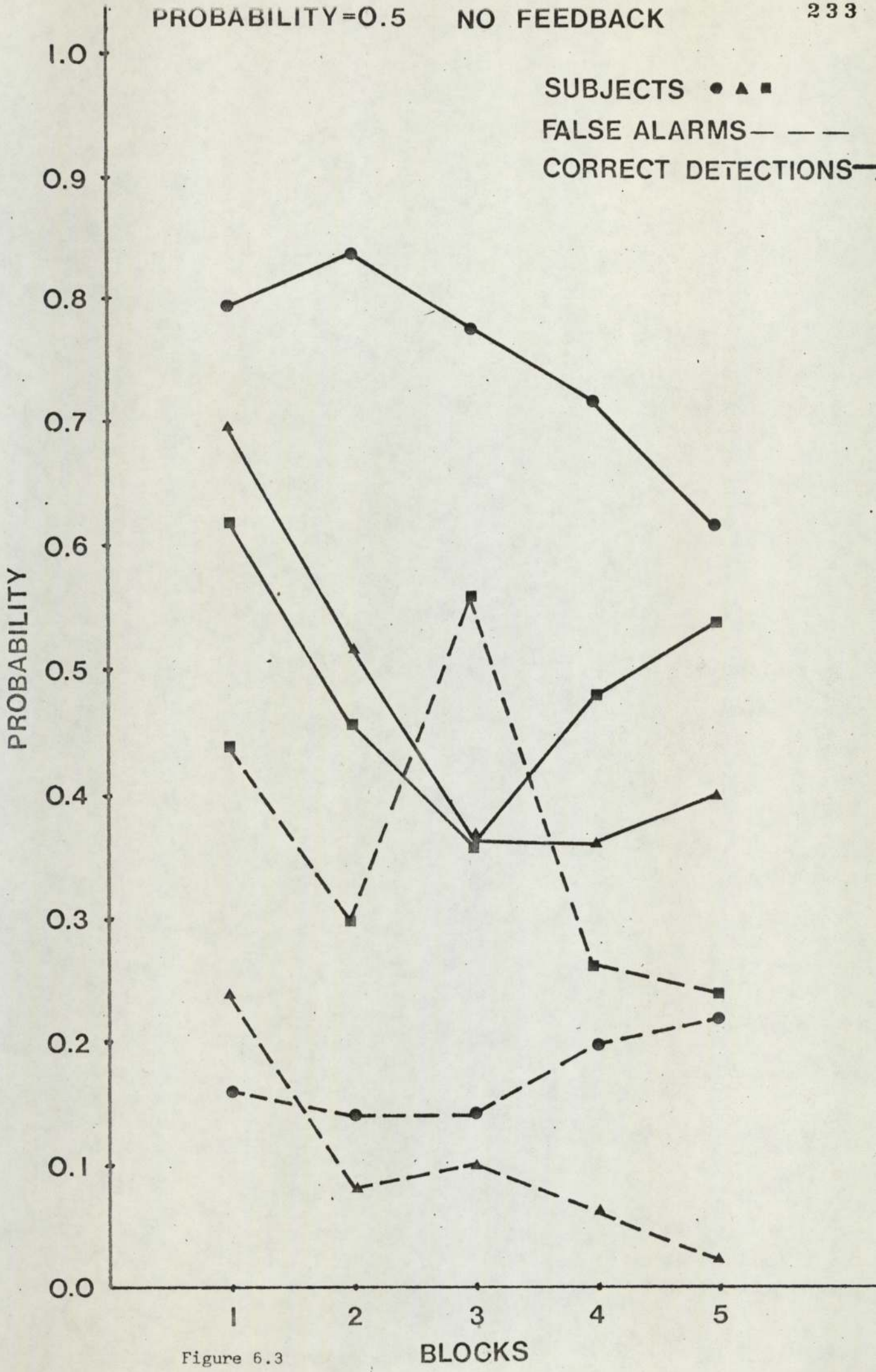

PROBABILITY=0.5 NO FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS— — —
CORRECT DETECTIONS——
PROBABILITY
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
BLOCKS

Figure 6.3

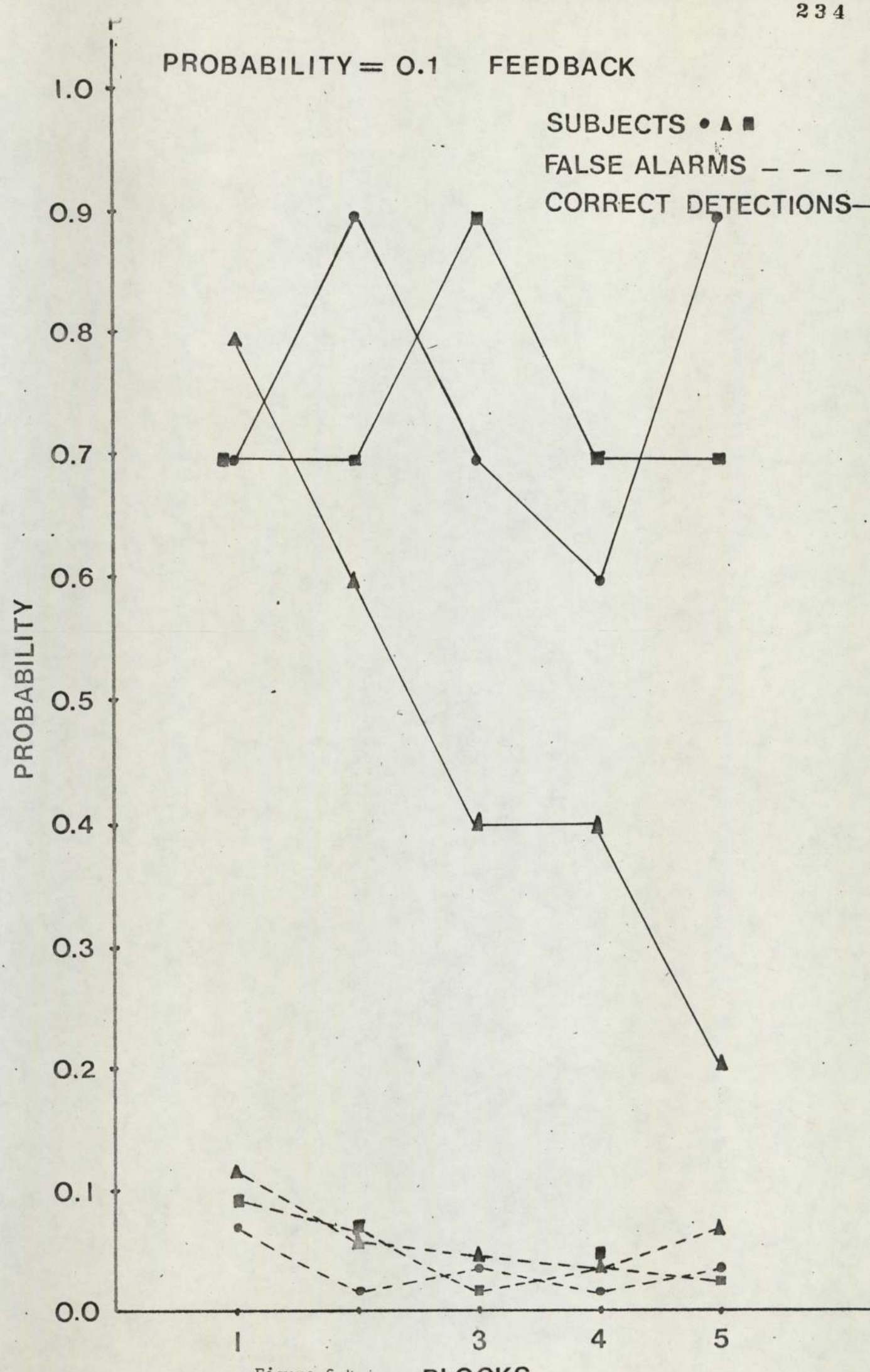

PROBABILITY = 0.1 FEEDBACK
SUBJECTS • ▲ ■
FALSE ALARMS – – –
CORRECT DETECTIONS—
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
1
3
4
5
BLOCKS

Figure 6.4

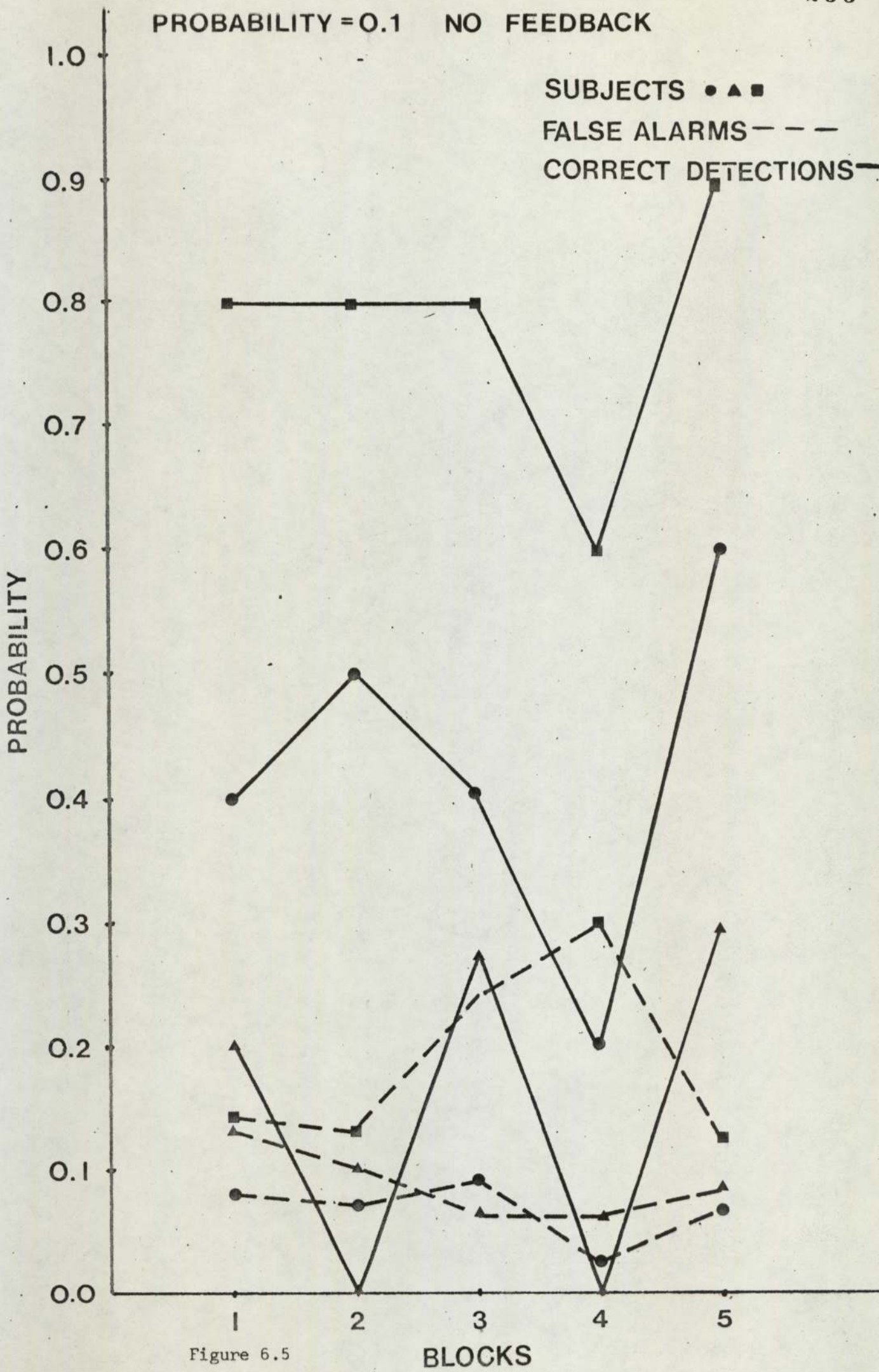

PROBABILITY = 0.1 NO FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS - - -
CORRECT DETECTIONS
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
1
2
3
4
5
BLOCKS

Figure 6.5

Considering the SDT measures, no significant differences are found for any of the sensitivity indices AH, P(A), d', de, or the Pollack-Norman index, under any of the experimental conditions. There are significant differences in bias as measured by log beta, between blocks ($p < 0.01$) and a significant feedback by probability condition interaction ($p < 0.05$). A similar pattern of results is observed for the Hodos-Grier bias index, the corresponding significance levels being $p < 0.05$ in each case.

Comparison of means for blocks shows that log beta increases significantly between the first and fourth ($p < 0.01$) and first and fifth ($p < 0.05$) blocks. This result accounts for the high false alarm rate observed during the first block. The correct detection probability was also highest during this block, as would be expected from the low log beta during this block.

The significant feedback condition x probability condition interaction is analysed in Table 6.1 below:

probability conditions	feedback conditions f/b	no f/b
P = 0.5	0.15	0.41
P = 0.1	1.35	0.87

Table 6.1 Comparison of probability and feedback condition means for log beta

A simple main effects analysis shows that log beta is significantly greater under the p = 0.1 condition, but only if knowledge of results is available.

The next variable analysed, the subjects' points score, is not strictly speaking a signal detection index, but is related to signal detection performance. It can be regarded as representing the value of the inspectors' performance to the inspection system, given the particular payoffs specified for the various degrees of response category which were described earlier.

There are highly significant differences between the probability conditions, the mean score for the low probability condition being very much greater than for the P = 0.5 condition ($p < 0.01$).

6.5.2 Latency data

The log transform of the total latency indicates significant differences between blocks. Multiple comparisons show that the first three blocks have a longer total latency than the last two ($p < 0.05$). The total latency is also significantly shorter ($p < 0.05$) under the P = 0.1 condition that the P = 0.5 condition. Analogous results are found with the correct rejection latencies, the significances being $p < 0.01$ and $p < 0.05$ in this case.

No significant effects are obtained for the correct detection and omission latencies. The false alarm latency however is very significantly longer for the P = 0.5 condition that the P = 0.1 condition ($p < 0.001$).

6.6 Discussion

The variables of particular interest in this experiment are the SDT measures of sensitivity and bias. The lack of any significant

effects for any of the sensitivity measures was not surprising, since none of the experimental factors was expected to affect sensitivity. The significant increase in beta found over blocks is unexpected, and closely resembles that found in vigilance tasks. In fact, some form of time related performance decrement is the only obvious explanation for this effect. If it occurred only with the P = 0.1 condition we could explain it in terms of the subjects attempting to adjust their criteria from the unbiased one appropriate to the P = 0.5 condition to a more stringent one for the lower probability condition. There is no obvious reason why vigilance effects should occur in view of the fact that frequent rests were provided. The effect also occurs in conditions where feedback is provided, which is contrary to almost all vigilance studies in which KR has been given. Additionally the subjects gave no verbal intimation that they found the experiment particularly boring. A possible explanation is in terms of the particular circumstances of the experiment.

The subjects were aware that the experiments were concerned with their ability to adjust to changes in defect probability, and may therefore have suspected that a change in defect density would occur within the session. There is some support for this notion in that the criterion, after reaching a maximum in session 4, declines again during session 5. Possibly the subjects had decided by that point that no change was going to take place.

The analysis of the feedback x probability interaction is of particular interest. It suggests that subjects are able to adjust their criteria in the optimum direction predicted by SDT, but only if they are provided with knowledge of results. The result suggests that in

this experiment, external sources of information were more important than evidence gained from the task itself, in shaping the criterion. It is difficult to compare this result directly with the Sims study, because no measurement of the criterion was obtained in that situation. In the current experiment we have no direct estimate of the subjective probability that the subject was employing. If we make the assumption that a change in subjective probability was translated directly into a change in criterion, in the present study, then the result is in direct contrast to the Sims study. The difference can be accounted for by the relative difficulty of discrimination of defects in the two studies. Although no measure of discrimination performance was given by Sims, it is clear from her description of the task, that it was considerably easier than the one under consideration. It is likely that Sims' task would provide far more intrinsic feedback for the subject than the present one.

These considerations lend support to the idea that the ability of a subject to adjust his criterion to the optimum for a particular defect probability is a function of the amount of information available on the defect density, whether it be from within the task or from external sources. Where the implicit task information is highly reliable, i.e. the signals are readily discriminable, the Sims study suggests that external evidence on the defect density is redundant.

The fact that the subjects obtained a higher net score under the $P = 0.1$ condition can be accounted for by the fact that in spite of extensive training, it was easier for the subjects to recognize a non-signal than a signal, as indicated by the greater variance of the

the signal distribution obtained from most of the ROC fits. This allowed them to make a high proportion of 'definite' (i.e. extreme rating category) responses when most of the trials were known to be non-signals. It is also possible that the subjects were adopting a nearly pure strategy with the low probability signal condition, as proposed by the Siegel-Goldstein probability learning hypothesis discussed earlier. Most of the subjects were aware that they could obtain a high payoff by responding 'definitely non-signal' on a high proportion of the trials, regardless of the evidence from the task. It seems possible that in a difficult discrimination task, with no a priori information as to defect probability, the criterion will be set according to the subjective probability generated by considerations discussed in the probability learning literature.

Most of the latency results can be accounted for by the SDT model. If the situation is as set out in Figure 6.6 below, with the signal variance greater than the noise variance, then with the assumptions set out in Chapter 4, the following predictions can be made.

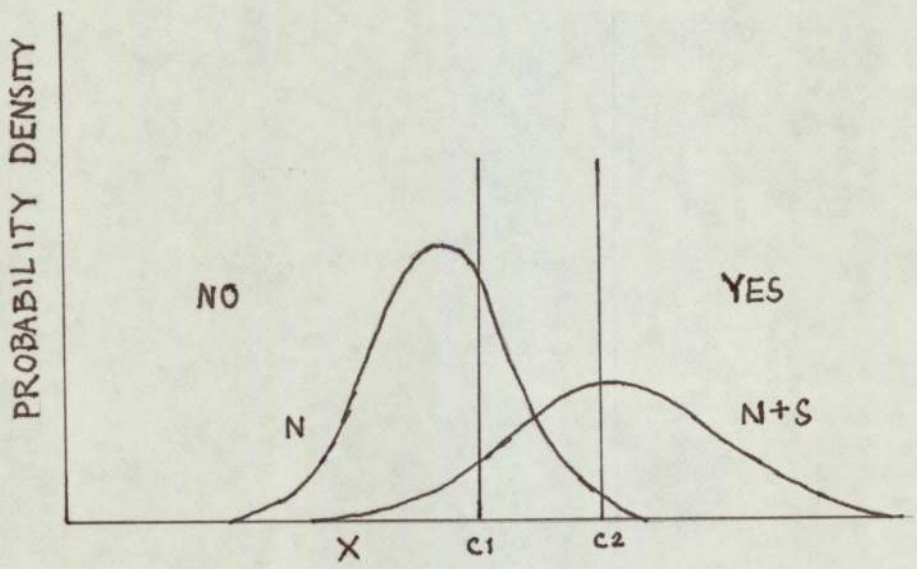


Figure 6.6 Response latencies under the unequal variance SDT model.

If the criterion moves from position C1 to a more stringent position C2, we would expect NO response latencies to decrease, since these would on average be distributed further from the criterion. In the equal variance situation, YES latencies would increase, because they would be distributed closer to the criterion and hence represent more difficult decisions. However, with the large signal variance situation illustrated above, the changes in both correct detection and omission latencies might be expected to be non-significant.

The results generally bear out these predictions. The criterion increases significantly with time on task and is overall greater for the P = 0.1 condition than the P = 0.5 condition. The correct rejection (i.e. correct 'no signal' decision) latency decreases significantly with both increases in criterion. The omission and correct detection latencies show no significant changes. The false alarm latencies are anomalous, since they should increase with the increase in log beta for the P = 0.1 condition, whereas they actually show a significant decrease compared with the P = 0.5 condition. This result can be seen as a consequence of the subjects habituating to a particular mode of response. In the P = 0.1 condition, since most of the responses were 'non signal' and the non-signal stimuli were more recognizable as such than the signal trials, the rapid high confidence responses made to non-signals probably tended to spread to other categories of response. This effect would be likely to dominate the effects of the greater bias. In the P = 0.5 condition, the equiprobable signal occurrence would prevent a style of response developing which was dominated by one type of trial.

The total response latency results are simply a reflection of their predominant component, the correct rejection response latencies.

6.7 Results - Experiment 5

The correct detection and false alarm probabilities for the blocks and experimental conditions are set out in Figures 6.7 to 6.9. The transformed correct detection probability showed significant differences between blocks ($p < 0.05$), the first two blocks differing significantly from the remainder. The transformed false alarm probability showed significant effects between blocks, with a highly significant blocks x conditions interaction, ($p < 0.001$). This interaction will not be analysed in detail at this point, since it is a direct consequence of the criterion effects to be discussed subsequently.

Considering the sensitivity indices, the results are somewhat ambiguous. The analysis for the Altham-Hammerton, Pollack-Norman indices and d' all indicate a significant blocks x conditions interaction, whereas d_e', and P(A) show no significant effects. This anomaly will be discussed in more detail subsequently. Log beta gives significant effects for blocks, conditions, and a highly significant condition by blocks interaction ($p < 0.001$). The simple main effects analysis is set out in Table 6.2 below:

Source	SS	df	MS	F	Significance
B at C1, no warning feedback	0.094	4	0.0235	0.28	ns
B at C2, warning, feedback	1.55	4	0.39	4.59	$p < 0.05$
B at C3, warning, no feedback	8.09	4	2.022	23.79	$p < 0.01$
error	1.355	16	0.0847		

Table 6.2 Simple main effects analysis of BxC interaction for log beta

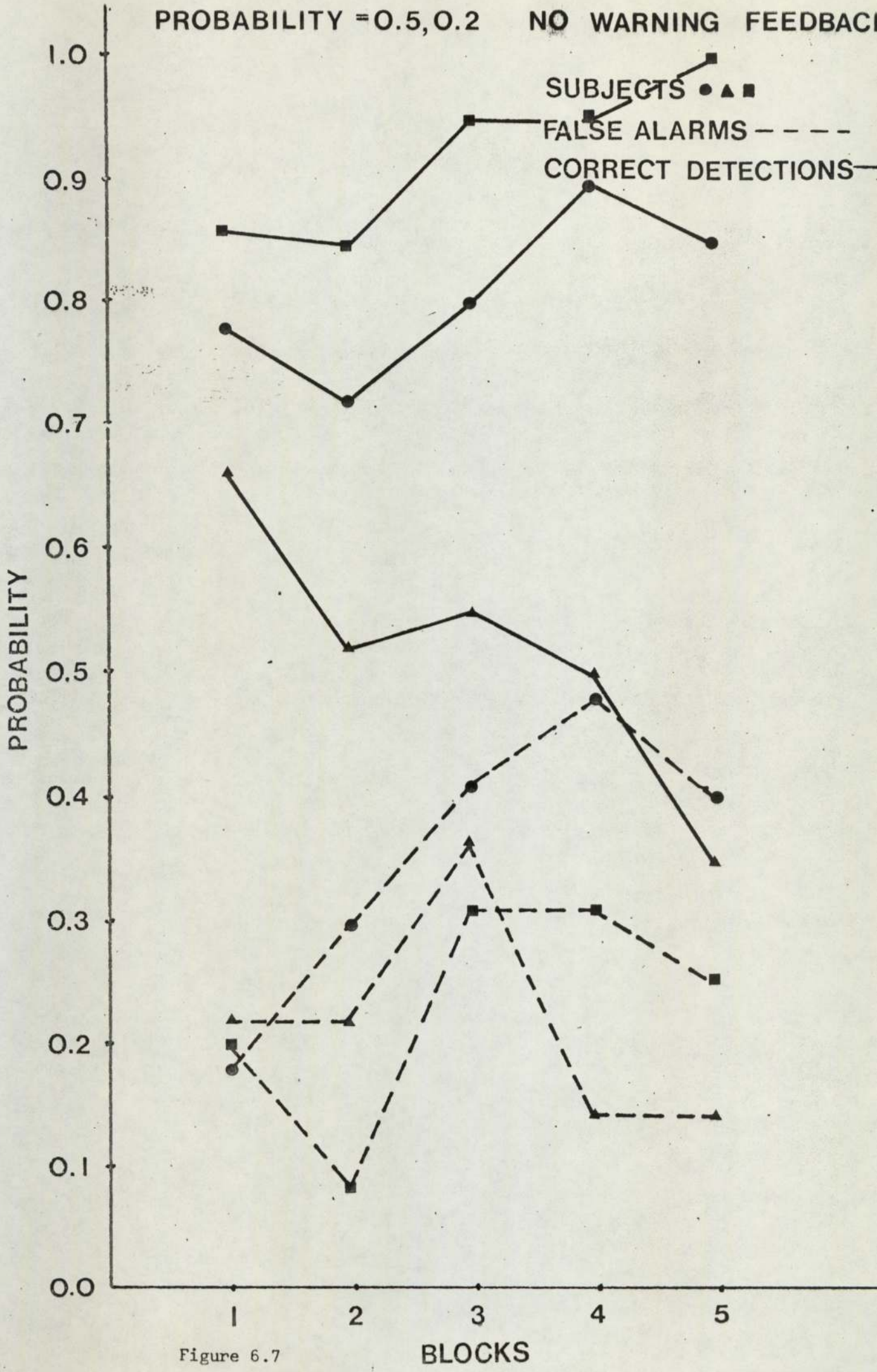

PROBABILITY =0.5, 0.2 NO WARNING FEEDBAC
SUBJECTS
FALSE ALARMS
CORRECT DETECTIONS
PROBABILITY
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
BLOCKS

Figure 6.7

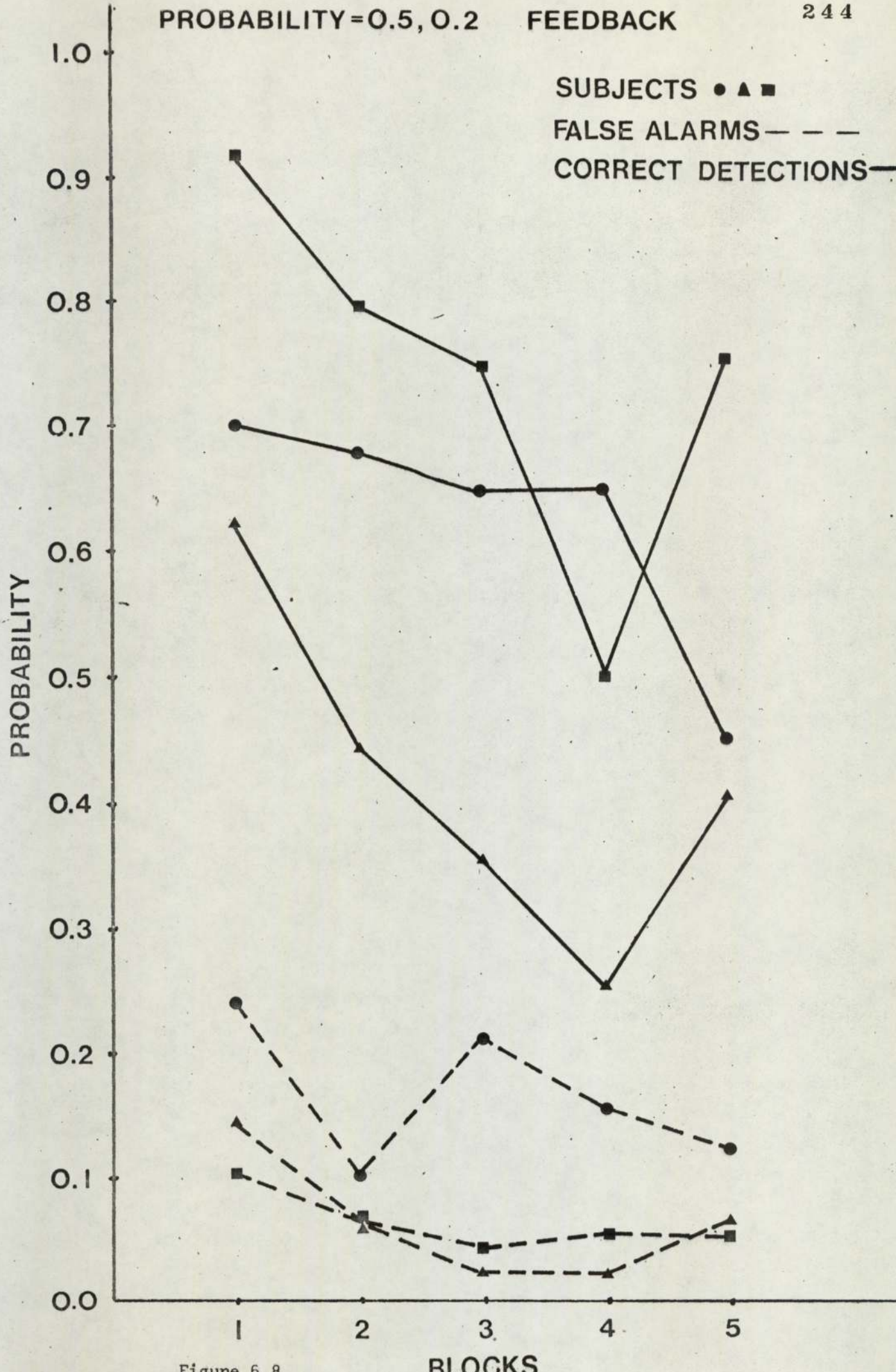

PROBABILITY=0.5, 0.2 FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS — — —
CORRECT DETECTIONS ——
PROBABILITY
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
BLOCKS

Figure 6.8

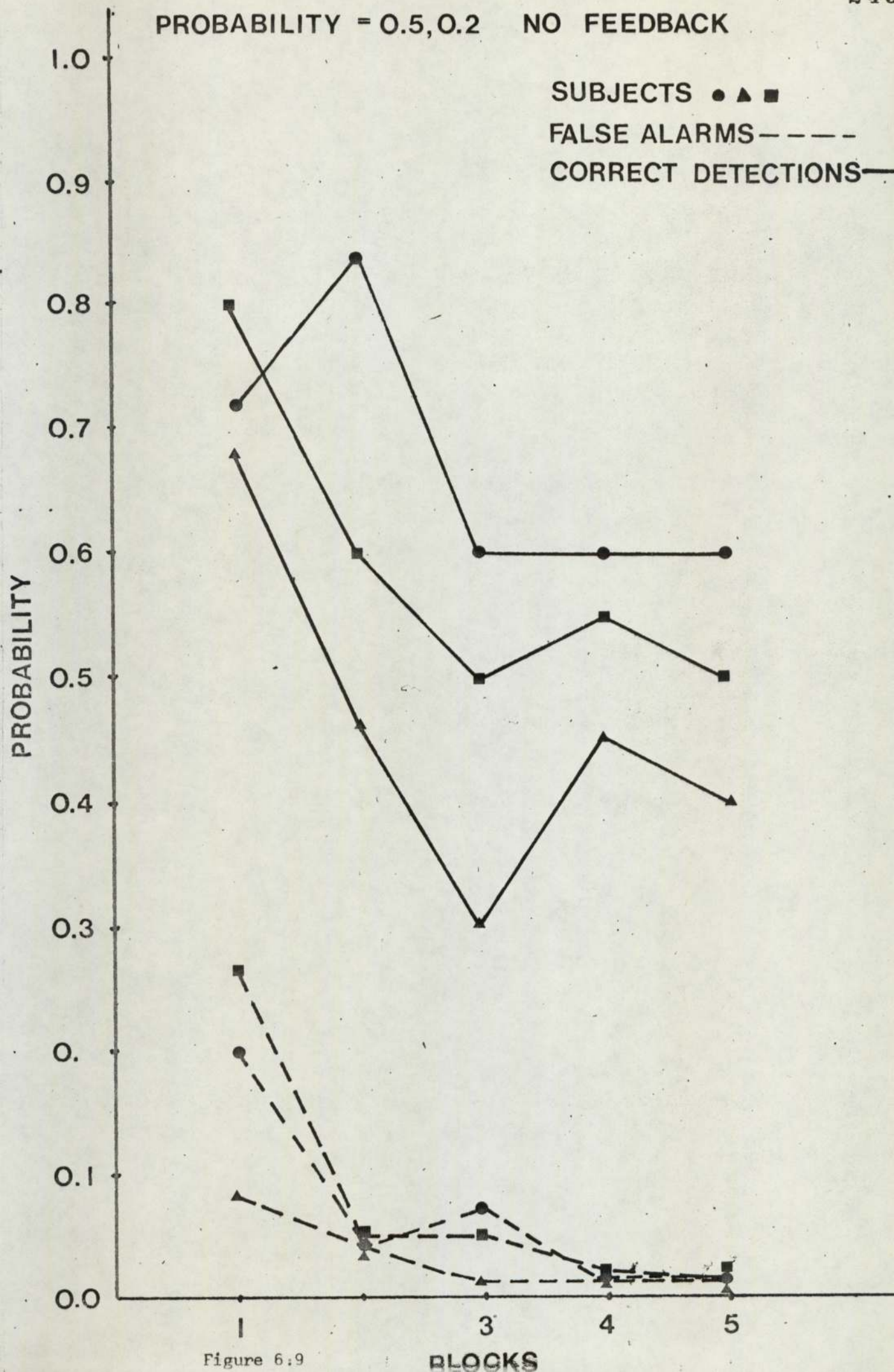

PROBABILITY = 0.5, 0.2 NO FEEDBACK
SUBJECTS ● ▲ ■
FALSE ALARMS - - - -
CORRECT DETECTIONS —
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.
0.1
0.0
PROBABILITY
1
3
4
5
BLOCKS

Figure 6.9

The simple main effects analysis indicates that there are significant differences between blocks for the two conditions in which warnings of change were given, but not for the no warning condition. Tukey multiple comparisons tests for C2 and C3 indicate that there are significant differences between blocks 1 and 2 combined together, and blocks 4, 5 and 6 combined, ($p < 0.05$, $p < 0.01$).

The latency results will not be considered in this experiment, since they do not provide any additional insights into the variables of interest.

6.8 Discussion

The first question to consider is whether any changes in sensitivity have occurred as a result of the within session change in probability. In view of the fact that both P(A) and de, which are known to be independent of the nature of the underlying variances in SDT, show no significant changes, the validity of the changes suggested by the other indices must be viewed with scepticism. This is particularly the case in view of the fact that sensitivity measures are not normally affected by signal probability changes. It seems likely, therefore, that the apparent change in sensitivity is an artefact in this instance.

The detailed results for the effects of the within session changes in probability on log beta provide very interesting insights. It appears that prior warning of a probability change is more effective in producing an appropriate criterion change than the provision of full feedback from the task. Also, when a warning is given, the subjects

are able to adjust their criterion in the appropriate direction, even in the absence of feedback. An interesting feature of the no feedback, warning condition is that the criterion continues to increase to the end of the session, whereas with feedback it remains fairly stable. The KR information can be regarded as encouraging a stable, moderately high criterion after the probability change, because even given a precise knowledge of the signal probability, from the KR, the subjects will always adjust their criteria in a conservative manner. Where KR is not provided, the subjects are likely to assume that the defect probability has changed to an even lower level than is actually the case. Adopting a more stringent criterion will lead to fewer defects being detected which will in turn produce an even higher criterion, as a consequence of the perceived signal probability.

6.9 Conclusions

These experiments have provided general support for the hypothesis that the degree of criterial change produced by a subject in a changing defect probability situation depends on the amount of information available on the probability from external sources, and from the task itself. The relative weighting of the internal and external sources of information would appear to be a function of the reliability of these sources. For between session changes in defect probability, even when the subject was aware that a reduced incidence of defects could be expected, significant criterion changes only occurred when knowledge of results was available. This could be explained by the fact that although the subject knew that the defect incidence was less, he did not know to what degree he needed to alter his criterion. The precise evidence obtained via feedback provided this information and encouraged the appropriate criterion change. In a situation of

high uncertainty, as with the within session probability change experiment, the importance of prior information is seen in enabling a criterion change to occur. Even when the subjects had full KR, they were unable to modify their criteria unless they had been alerted to the possibility of change. The fact that warning of change seems to be a more important factor in criterion adjustment, than feedback from the task, where within session probability changes occur, suggests further investigation is needed in this area. In both experiments the importance of external evidence of some form can be explained by the difficult nature of the discrimination task, and hence the unreliability of evidence from this source.

No attempt has been made to analyse the degree to which the subjects were able to achieve the actual optimum beta predicted by SDT. This is because, as shown by Green and Swets (1966) p. 92, there are a range of beta values about the optimum which will achieve a high proportion of the theoretical payoffs. This does not however, invalidate the practical importance of the inspector at least being able to modify his criterion in the direction appropriate to the current defect probability. As discussed earlier, an inability to do this can produce a very inefficient quality control system, if defect probabilities are liable to fluctuation.

Although these experiments were exploratory in nature and utilized a laboratory situation and a limited number of subjects, the results seem sufficiently interesting to suggest further work in an industrial context.

CHAPTER 7 TRAINING AND SELECTION FOR INSPECTION

7.0 INTRODUCTION

In this chapter experimental work arising from some of the issues discussed in the review of perceptual training techniques and selection methods in Chapter 3 will be presented. In the final phase of the study, two experiments were performed in which various types of training for perceptual skills were investigated. The performance measures from these experiments were used in conjunction with a number of tests of various cognitive skills, in order to establish the usefulness of such tests for the purposes of selecting individuals for inspection work.

7.1 Theoretical considerations in perceptual learning

It is useful at this point to briefly recapitulate the conclusions of the earlier review in the area of perceptual training.

The two basic methods of training that had been used to train for perceptual skills were cuing and knowledge of results (KR). Enhancement of signal detection ability by cuing was seen as a result of the simple paired contiguity in time of a signal and its name. This form of training promotes perceptual learning because it provides the subject with the maximum information concerning both the characteristics of the signal and its distribution in time.

It was suggested that the other form of perceptual training, KR, trains recognition skills primarily via the reinforcement of a simple S-R link. An important aspect of KR is its motivational effect in maintaining arousal during prolonged sessions. Improvements in performance using both techniques can be ascribed partly motivational effects,

partly to learning the characteristics of the signal, and partly to an increased knowledge of the statistical distribution of the signals. The applicability of the SDT approach to these latter two areas is clear. Increased knowledge of the signal characteristics implies an increase in d', whereas knowledge of the signal distribution produces an appropriate expectancy. This has been analysed in previous chapters as an accurate subjective estimate of the signal probability, leading to an optimization of the response bias, as measured by beta.

Early work by Annett and Wiener suggested that cuing enhanced d', whereas KR improved detection performance at the expense of increased false alarms mainly by producing a more lax criterion. Subsequently it was found that the latter result was to some extent an artefact of a free response situation, in that subjects appear to make a higher number of affirmative responses, and hence produce an apparently lowered criterion, in an attempt to gain more information about the signals and their distribution. Recent studies, e.g. Annett (1971) have suggested that in a situation where a subject has to respond to a series of successive trials, the differences between the two training techniques are slight, because they both provide essentially the same information.

Another aspect of perceptual skills training that was discussed was Wallis' (1963) suggestion that learning to recognize complex signals was best accomplished using an analytic-synthetic approach where the salient features of the stimulus are first learnt separately and then synthesized into a wholistic 'Gestalt'. The importance of the withdrawal of cues at an appropriate time was mentioned in this study, and the whole question of the dangers of a subject becoming dependent on cues or KR, to the detriment of learning, was further

emphasized by Abrams and Cook (1971). These workers employed a gradual reduction in the amount of KR to reduce dependence on feedback, and suggested that this technique facilitated the development of the internal referents necessary for identification skills.

From the point of view of its applicability to inspection, much of the existing work on perceptual skills training suffers from several disadvantages. The stimuli employed have usually been unrepresentative of those found in inspection tasks, and much of the work has been in areas such as sonar where auditory signals are employed. These factors were considered in the design of the experiments described in the following sections.

Experimental work - general

In view of the points discussed earlier it was decided to attempt to simulate the critical features of a real inspection task in both experimental studies. The task chosen was one that had already been analysed in some detail, the Ilford film inspection task described in Chapter 5. It will be recalled that considerable problems had been experienced in scoring experiments with the task in its original form, because of the difficulty in resolving the position of the defects on the film with a sufficient degree of accuracy to prevent ambiguity. It was therefore decided to alter the task in the following manner. A particular type of defect, known as 'insensitive spots' was chosen as being representative of those encountered by the inspectors in their everyday work. This defect type consists of a number of tiny spots on the film where the coating has not adhered adequately to the base. Because the film coating is

not darkened at these points, the appearance of the defects when the film is projected is that of a number of small points of light on the otherwise grey background of the correctly exposed film. Actual samples of defects were cut from films and turned into a series of slides. Similarly samples of 'perfect' film from other parts of the same roll were made into 'non defect' slides. This was done to ensure that the background density for both types of slide was of the same apparent brightness in both cases. These slides were then used in conjunction with the equipment described in Chapter 6, to simulate the appearance on the screen during film examination of a single frame containing insensitive spot defects. The speed of the shutter was adjusted to give a viewing time for each slide of 1/15th of a second. This was the closest speed available to the 1/18th second employed in the actual task.

A continuous overall background illumination of the screen was provided by a second Kodak carousel projector and was -0.4 log foot lamberts as measured by an SEI photometer. When a slide was projected on to the screen, the brightness increased to 0.2 foot lamberts, which approximated to that found in the actual inspection situation.

The basic operation of the equipment was similar to that described in Chapter 6. At the start of the session the computer operated the slide projector magazine to load the first slide, and then the Compur magnetic shutter to provide a presentation on the screen of 0.06 seconds. A software clock within the computer was started at the same time and this enabled the subjects response latency to be measured. The response arrangements were as in previous experiments. A series of six buttons was provided to enable the subject to make

the rating responses described previously. As before, a symmetrical payoff scheme was assigned to the responses such that 1, 2 or 3 points were gained for correct responses of increasing certainty and a similar number of points were subtracted for the corresponding incorrect responses. The computer outputted the response type and time on paper tape after each response.

The detailed arrangements for each experiment will be described separately in the following sections.

7.1 Experiment 6 - comparison of cuing and feedback techniques

The first experiment was designed to investigate the effectiveness of cuing versus knowledge of results in the simulated inspection task. To do this, three conditions were investigated. These were cuing, feedback and a control condition. All three groups were given a short practice session to familiarize them with the equipment and the method of response. The scoring scheme was carefully explained to them and twenty practice trials were given, at the end of which the individual slides were discussed to indicate what were the characteristics of the defective as compared with the non-defective slides. It was pointed out that even the 'perfect' slides contained many configurations which were confusable with defects and that they were only to categorize as defective those slides which contained 'insensitive spots'.

The subjects were assigned at random to one of the three experimental groups and all subjects first received five hundred trials as the pre-training phase. As in experiment 5, attempts were made to

minimize vigilance effects by providing the subjects with a three minute rest after each 100 trials. Low level white noise was played through headphones throughout the experimental sessions, but subjects were allowed to remove these and converse with the experimenter during the breaks.

For the training sessions, the control group received a further 100 trials similar to the pre-training session. Before each slide was presented to the cuing group, the computer typed a message on the teletype, which was visible to the subject, of the form: 'the next trial will be a defect' or 'the next trial will be a signal'. The subjects were told to read the message, and to use it to ready themselves to observe the characteristic features of the defect or non-defect. They were also told to make an appropriate response after the slide had been presented. For the KR group, a message was typed by the computer after each response and consisted of one of the following four types:

1. You have just missed a defect.
2. That was a false alarm.
3. That was a correct detection of a defect.
4. That was a correct detection of a non-defect.

The final post training session was a repeat of the first session, in which five hundred trials were given.

Twenty one subjects were employed in the experiment, seven being assigned randomly to each condition. They were all undergraduate students from various disciplines and were paid 50 pence per session. The pre-session and training session were administered on the same

day, and the post session, as far as practicable, at the same time on the following day. The probability of a defect remained constant within blocks of a hundred trials throughout the experiment and was 0.2.

Analysis of the results

As described for experiment 5, the paper tape output from the PDP-8 was transformed into magnetic tape files on the PDP-15 computer and the various performance measures employed in experiments up to this point calculated. The statistical design employed was a 'split plot' experiment (Keppel (1973) p.433). The between subjects factor was the training conditions, and there were repeated measurements on blocks of 100 responses within the 500 blocks for each subject. The quantity entered into the analysis of variance was the difference between the pre and post training sessions for the various performance measures. This method of analysis is generally regarded as being superior to including before and after training as a separate factor, because it compensates for the differing initial performances of the subjects.

Where the summary data for all 500 responses were used in the analysis, the design becomes a simple one-way completely randomized analysis of variance.

The rating data from the performance summaries for all 500 responses of the pre and post training data were run through the Grey-Morgan ROC curve fitting program to obtain the variance ratio, the sensitivity index d'_e and the five beta values employed corresponding to the six rating categories utilized by the subjects. A fit was obtained for 41 of the 42 sets of data, although in some cases signal to noise

variance ratios of 23 had to be assumed by the program to carry out the fitting procedure. This is not surprising in view of the fact that the first 500 responses the subjects were unfamiliar with the characteristics of the signal. The results in general suggest that the SDT model applies, however.

Results

All analyses of variance will be found in Appendix A, together with summary data.

The first analysis was conducted on the SDT information output from the Grey-Morgan program. Only the summary data for all 500 responses in the pre and post training sessions was considered. The sigma ratio was first analysed as an index of sensitivity change, because one might expect the variance of the signal distribution to decrease as the subject learns the characteristics of the signal. Such a decrease was only found for the cuing condition, but the differences between the conditions were non-significant. A similar result was obtained for d_e, the corrected sensitivity index. The next variables considered were the log beta values for each cutoff corresponding to the rating categories employed by the subjects. The only significant effect was a significant difference between subjects for cutoff 4 ($p < 0.05$).

For the remaining analysis, the separate blocks of 100 responses were considered as repeated levels of the same factor for each subject. No significant effects for any of the sensitivity measures were found apart from d', which indicated a significant change between blocks.

In view of the unreliability of d' in the unequal variance situation, and the lack of corroboration from any of the other sensitivity measures, this result must be regarded as artefact.

Considering the bias indices, a significant difference between training conditions was obtained for beta ($p < 0.05$). Tukey's test indicates a significant difference from the control, but not from the other training condition. The Hodos-Grier bias index indicates significant differences between blocks, ($p < 0.05$) the change in bias as a result of training being significantly greater for block 1 compared with blocks 3 and 5 (Tukey test, $P < 0.05$). The only other significant effect obtained was significant differences between blocks in the changes in false alarm probability due to training. The false alarm probability declined significantly for block 1 compared with blocks 3, 4 and 5.

Discussion

The overall absence of significant changes in sensitivity produced by either of the training conditions could be due to the relatively small number of training trials employed. The increases in bias observed as a result of training are in the opposite direction to those found with KR in free responding situations. The direction of change was in the correct direction in this experiment, since the final value of beta was closer to the optimum than the original. We can conclude therefore, that cuing seems to promote a greater change in bias in the correct direction than KR, although the difference is not statistically significant in this experiment.

The changes in bias indicated by the Hodos-Grier index, and reflected by the false alarm probability changes, probably occur because during the earlier blocks of the pre-training session, the subjects were utilizing a more inappropriately low criterion than during the later blocks, as a result of their initially limited knowledge of the signal distribution. The effects of the training in increasing bias would therefore be more marked for the earlier than for the later blocks.

7.2 Experiment 7 - further training techniques

Although the last experiment yielded some interesting results it was felt that a wider variety of training techniques should be investigated, particularly from the point of view of improving performance in situations where the probability of a defect occurring varied within sessions. It was also felt that a greater amount of training might produce changes in sensitivity which had not been revealed by the first experiment.

The first two conditions that were used in the final experiment were feedback alone and cuing alone. These conditions were included to replicate the first experiment, but to provide a greater number of training trials. The next condition was a combination of cuing and feedback, in which alternate training trials were either cued or KR was supplied after the response. The object of this technique was to investigate whether cuing and KR utilized together were more effective than either used alone. The fourth condition provided an even greater amount of information to the subject by giving him alternate cuing and KR as in condition 3, and also supplying him with a summary of his performance in terms of correct detections, false alarms

etc. at the end of each block of 100 responses. The final condition was similar to condition 1, the feedback only situation, but in addition to the feedback information, the subject was provided with the points score he had obtained as a result of his responses, at the end of each block of 100 trials. This condition was included to establish if the provision of information concerning the payoffs associated with the subject's response strategy would enable him to optimize this strategy more effectively than if feedback alone were provided.

In addition to the five training conditions, two other factors were included in the experiment. As described in the introductory section of this chapter, and in the full review in Chapter 3, one of the difficulties associated with training for perceptual skills is that subjects may become dependent on the presence of the particular training aid to maintain performance. This may prevent the performance improvement obtained during training being transferred to post training sessions. Abrams and Cook (op. cit.), had found that gradually reducing the amount of feedback during training improved the retention of any detection skills acquired. It seemed worth investigating if this effect was also obtained with cuing. All the training conditions described earlier were therefore performed under two levels of this factor. In the one case, feedback and/or cuing was provided on every trial. At the other level of this factor, the amount of information provided by cuing or feedback was gradually reduced throughout the session.

The other question of interest which was investigated in the experiment was whether the experience of a varying probability of defect occurrence during training would facilitate performance in post

training sessions where the within session defect probability also varied. It was hypothesized that the ability to adjust the criterion optimally in a changing defect probability situation would be enhanced if the subject had prior experience of a range of defect probabilities. The training conditions were therefore performed both with a constant level of defects and also with a defect incidence that varied from block to block during the training sessions.

Procedural details

The equipment utilized was essentially the same as for experiment 6. Certain changes were made, however, in the way that the feedback and cuing information was conveyed to the subject. It was found that during the first training experiment, the use of the teletype to type messages to the subjects had considerably slowed down the experiment. Also, subjects become irritated at having to read a long message before or after each response. An alternative means of conveying information was therefore designed and built. This consisted of four coloured lights on a block painted panel. The lights were operated by the computer to provide the appropriate cuing or feedback information. For the cuing conditions a red light was illuminated prior to the presentation of a defect, and a green light before each non-defect slide. For the KR condition four lights were employed. A green light was used to indicate a correct detection, and white, red and yellow lights were illuminated following correct rejections, false alarms, and missed signals respectively.

Each subject performed six sessions, each of 500 responses in blocks of a 100 with rest periods between each block, as in experiment 6.

Prior to session 1, a short familiarization session was given as before. Session one was the pre-training session, in which the defect probability remained constant at 0.2 within blocks of 100 responses. The second session, which was performed on the same day as the first session, was designed to investigate detection performance in a varying probability environment. The defect probabilities for blocks 1 to 5 respectively were : 0.15, 0.15, 0.40, 0.20, 0.10. The subjects received prior warning that the probabilities would vary from session to session. The next two sessions, which occurred the following day, were both training sessions. The training sessions were separated by a break of at least 45 minutes to minimize any effects of fatigue. In the variable probability training sessions, the subjects were told the probability prior to each block of 100 responses. The probabilities were chosen to be representative of those found in the test sessions.

On the day following the training the two final sessions were administered. The first of these was identical to session one, and the second similar to session 2, but with a different sequence of defect probabilities for each block, i.e. 0.25, 0.15, 0.20, 0.10 and 0.30. The differing sequence was chosen because the event probabilities employed in session one had also been utilized during the training period, and there was some possibility that the subjects would learn the sequence of probabilities. It will be noted that the probabilities used for each block of 100 responses, in both the pre and post training variable probability sessions, summed to give the same overall probability of 0.2 as in the fixed probability sessions. This was done to ensure that any performance differences between the fixed and varying probability sessions was a result of the variability itself rather than because of an overall difference in defect

probability. Prior to session 6, the subjects were told that the defect probability would vary, but that it would be different than during the training sessions.

The subjects were recruited from advertisements in schools and their ages ranged from 16 to 19. All had normal or corrected vision.

Statistical design

The basic structure of the experiment is set out in Figure 7.1 below.

training conditions	constant KR and cuing		fading KR and cuing	
	fixed defect probability	varying defect probability	fixed defect probability	varying defect probability
feedback alone cuing alone feedback + cuing feedback + cuing + summary feedback + score	B1 B2 B3 B4 B5			

Figure 7.1 Experimental design

The results were intended to be analysed as a four-way analysis of variance. The factors were training conditions, constant or fading cuing or KR, fixed or varying defect probability, and a repeated factor for each subject of 5 blocks of a hundred responses within sessions. The scores to be entered in the analysis of variance were the differences between pre and post training in the case of sessions 1 and 5. Sessions 2 and 6 were to be analysed separately to assess the degree to which the subject was able to modify his criterion to take into account the changing probabilities. It was originally

intended to use at least two subjects within each cell of the analysis. Unfortunately circumstances beyond the control of the author meant that the original design had to be modified. There was a delay of six months in obtaining essential interface circuitry from the computer manufacturers and this, together with other unforeseen delays, meant that the design had to be changed after the experiment had begun.

It was decided to use a strategy discussed in Kirk (1968) p.227 and Winer (1962) p.216 and p.267. These authors suggest that by assigning one subject only to each cell of a completely randomized analysis of variance, an estimate of the error mean squares necessary to conduct the analysis of variance can be obtained from the highest order interaction, by making certain assumptions which will be discussed in the next section.

The use of the additive model in the analysis of variance

If we consider the summary data for each session and do not analyse the within subject variable of blocks within sessions, the experiment can be regarded as a three-way analysis of variance, with training types, fixed or varying probability and fixed or fading feedback or cuing as the factors.

There are two alternative models for the analysis of variance that can be postulated. Using the usual notation for factorial experiments, these are:

1. $X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \epsilon_{ijk}$
2. $X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijk}$

Model 1 assumes that all sources of variation other than the main effects and first order interactions are part of the experimental error ϵ_{ijk}. In model 2 a second order interaction is assumed to occur. This interaction term may be considered as a measure of the non-additivity of the main effects and first order interactions, i.e. the extent to which the observation cannot be predicted from a knowledge of three main effects and interactions, and the experimental error. Model 1 is known as the additive model and model 2 the nonadditive model, using Winer's terminology.

On a priori grounds, in the present experiment, there did not seem to be any reason to suppose that the second order interaction would be significant. A test is available, however, due to Tukey (1949) which enables one to decide which of the models is appropriate. If the first model is applicable, the variation due to sources other than main effects and first order interactions can be regarded as consisting of two components. The first of these is that due to the linear X linear component of the ABC interaction, and measures the degree of nonadditivity. The remainder is known as the balance. If the additivity component is significantly larger than the balance, model 1 is rejected.

An F test that will test this hypothesis is given by:

$$F = \frac{MS \text{ nonadditive}}{MS \text{ balance}}$$

The computational procedures necessary to obtain this quantity for a three way analysis of variance are given in Winer (1962) p.267. The F test is usually made at a high critical value ($p < 0.25$) in

order to reduce the probability of a type 2 error. If the F value obtained does not exceed the critical value, then the additive model is appropriate for the analysis, and the error term used is the ABC interaction. A program was written to perform the calculations, and used to provide a nonadditive F test for each block of data prior to performing the analysis of variance. Twenty student subjects were assigned randomly to the experimental conditions.

Results - fixed probability sessions

The assumption of the additive model was upheld with most of the data analysed. In the cases where the additive model was rejected it was sometimes possible to obtain a transformation of the variable concerned which did fit the model.

The first data analysed were the differences between sessions 1 and 5, the fixed probability sessions, for a range of variables. The Grey Morgan programme was used to check the validity of the SDT assumptions and to produce the first variables to be analysed, the sigma ratios, the sensitivity index d'_e and the beta values corresponding to the rating categories employed.

The only significant effects obtained were for the beta value corresponding to cutoff 2. Unfortunately, this was one of the cases for which the nonadditive hypothesis was rejected, and hence no conclusions regarding these effects could be drawn.

The analysis was then carried out on the range of variables considered in previous experiments, including the latency measures, the

parametric and nonparametric indices of bias and the net changes in payoffs obtained by the subjects as a result of the training given. The actual inputs to the analysis of variance were the differences between these quantities for the summary data for the whole of sessions 1 and 5. No significant effects were obtained for any of the resulting 22 analyses of variance.

It was therefore decided to repeat these analyses for each separate block of 100 responses. The latency data were not included in these analyses. For the first block of data significant effects were obtained for the changes in payoffs and for d'. In both of these cases however, the additive model was rejected by the F test. A significant interaction was obtained for the arcsine transformation of the false alarm latency, between the fixed or variable probability during training factor and the training conditions ($p < 0.05$). Simple main effects analysis of this interaction indicated that the feedback alone training condition produced a significant reduction in false alarms, but only when a varying defect probability was employed during the training period. A significant effect was obtained in block 4 for beta, which indicated that the increase in beta due to training differed significantly between conditions in which fixed and varying probability was employed. Examination of the means indicated that the fixed probability conditions produced an increase in beta whereas the varying probability conditions showed a slight decrease, the changes being +349 and -0.88 respectively. This effect was not, however, reflected by significant changes in the correct detection or false alarm probabilities, or the nonparametric bias measure. In block 5 there was again a significant effect of the use of fixed or varying probability in the training session, this

time on the sensitivity as measured by the Altham-Hammerton index. The means suggested that the decline in sensitivity as measured by this index was greater for the fixed than for the varying probability conditions.

Considering the results in general, for the fixed probability sessions, the overall pattern that emerged was that the criterion increased slightly to a value closer to the optimum. This was reflected in a slight increase in the overall payoffs obtained. An anomalous result was that all the sensitivity indices indicated a slight decline in sensitivity between the pre and post training sessions.

Varying probability sessions

With the modified experimental design which was adopted, it was not possible to make direct statistical tests on the ability of the subjects to modify their criteria to match the changing within session probabilities. By considering the overall changes between sessions 2 and 6, however, indirect evidence could be obtained regarding this question.

Considering the results in general terms, there was an overall slight increase in all the sensitivity indices and in beta, after the training sessions. There was some evidence that the subjects were generally adopting a more optimal strategy, as evidenced by the increase in payoffs obtained.

There was a significant interaction between the fixed or varying probability during training factor and the fixed or fading factor for d'

($p < 0.05$). Although none of the other sensitivity indices reached significance, high F ratios of 4.58 and 3.25 were obtained for the same interaction for the Altham-Hammerton and Pollack-Norman indices, which suggested that effect was a genuine one. Examination of the interaction in detail indicated that a combination of fixed probability with fading feedback or cuing produced the greatest increase in sensitivity.

Both the false alarm probability and the correct detection probability showed significant decreases for the fixed or varying probability during training factor ($p < 0.05$). In the case of the false alarm probability, this factor interacted significantly with the training conditions, ($p < 0.05$). The simple effects analysis for this interaction is given below:

Source	SS	df	S	F	Significance
C at A1	0.083	4	0.02075	11.99	$p < 0.05$
C at A2	0.029	4	0.00715	4.085	N.S.
error	0.00693	4	0.00173		

Table 7.1 Simple effects analysis for AxC interaction for false alarm probability.

The analysis indicated that the false alarm probability decreased significantly after certain types of training, but only when a fixed probability was employed during the training session. The comparison of means for the training conditions indicated that there were significant differences between training condition 4 and conditions 1, 3 and 5. Training condition 4 is the maximum information session, where feedback, cuing and a summary were all provided. Similar

results were obtained when the data for the correct detection probability was considered in detail. The decrease in correct detections were more marked in the fixed probability condition. As with the false alarms, the greatest decrease in correct detection probability was found with a combination of training condition 5 and a fixed probability during training.

Discussion

Considering the fixed probability sessions, the overall picture is of an increase in the criterion, but a slight decline in sensitivity. This latter finding is very difficult to account for. With the considerable amount of experience that the subjects had received of examples of both defective and non-defective slides, it is very difficult to understand why at least some improvement in sensitivity did not occur.

The fact that the significant effects which did emerge from the analysis appeared in separate blocks, and were usually unsupported by changes in other related variables, leads one to suspect their validity. By performing separate analyses of variance on each block of data, a total number of 360 F tests were performed. It seems likely that at least some of these would achieve significance by chance, and this may account for the absence of any coherent pattern.

The varying probability session results produced more definite conclusions, which were more in accord with previous work. An overall increase in sensitivity occurred between the pre and post training sessions. Certain combinations of conditions seemed to promote a

greater increase in criterion than others. In particular the fixed probability version of the feedback + cuing + summary training condition promoted a greater degree of change towards the optimum than the other training methods. This was corroborated by the significant changes in false alarm and correct detection probabilities.

The finding that the training condition which gave the maximum amount of information about the signal distribution, also produced the greatest criterion change towards the optimum is in accord with the previous work on perceptual training reviewed earlier. Why this change should be greater when a fixed probability was utilized during the training session is not clear. Support for the suggestion that this combination of training variables enables the subject to optimize his criterion most effectively comes from a consideration of the analysis of variance for the total score obtained. (Appendix A). Although it does not achieve significance, an F of 3.53 for the interaction of training conditions x probability type used during training was obtained. Consideration of the summary table for this interaction shows that the magnitude of the increase in payoff for the fixed probability, feedback + cuing + score training condition was considerably greater than for any other combination. Since the payoff obtained is strongly influenced by the response bias of the subject, this provides reasonable indirect evidence that this form of training enables the criterion to be adjusted effectively to the changing probabilities.

The result that d' appears to be increased by fading the cuing or feedback employed during the training sessions confirms the finding of Abrams and Cook (op.cit.) and Wallis' suggestion that supplementary

cues and information must be withdrawn during training in order to promote effective learning of the signal characteristics.

If we consider the results in general, the most difficult aspect to account for is the relative absence of significant effects for the fixed probability sessions. This can be partly ascribed to the insensitivity of the design employed. Only 4 degrees of freedom are available for the error estimate, and an F value of at least 6.91 is required to achieve significance at the 5% level. On the other hand significant effects were obtained for the variable probability conditions. There seems to be no obvious reason for this effect.

In summary, the results have the following implications for training inspection and other perceptual skills. Where a variable signal probability situation occurs, the form of training which appears to produce the most effective optimization of the criterion is one in which KR, cuing, and knowledge of the payoffs resulting from the responses is given, in conjunction with a fixed signal probability. Sensitivity appears to be enhanced when the feedback or cuing employed in the training technique is gradually reduced.

7.3 The use of tests of cognitive skills in the selection of inspectors

In Chapter 3 it was proposed that various tests of cognitive skills might be suitable as selection tests for industrial inspectors. Three types of cognitive skills were proposed as having relevance to the inspection situation. The first of these was known as field independence, the ability of a subject to selectively disembed wanted items from a confusing background. In view of the fact that

most inspection situations involve the recognition of the characteristics of a defect in a perceptually ambiguous background, it was proposed that some of the embedded figure tests which measure field independence might prove suitable for selecting inspectors.

The next type of perceptual skill which seemed to be relevant to the inspector situation was the ability of the individual to resist distraction, where a distracting context may be regarded as one which tends to obscure a wanted signal without changing its nature. This cognitive skill was regarded as being related to, but distinct from, the dimension of field independence measured by embedded figure tests. The final cognitive skill of interest was known as shifting, and could be described as the ability to change one's attentional focus at will. In an inspection context, this ability would be important in allowing the inspector to readily shift attention as each item was presented.

In the study to be described in this section, an attempt was made to investigate the extent to which performance on the detection task described in the earlier sections of this chapter correlated with various pencil and paper tests designed to measure individual differences in terms of these three dimensions of perceptual skill. If sufficiently high correlations were obtained between the tests and the various performance measures, this would suggest that they could, after suitable validation, be employed for selection purposes.

Procedure

The first groups of subjects consisted of the 20 participants in experiment 7, the last training experiment described. For various

reasons, three of these subjects were unable to take the tests, and hence the final number of subjects employed was 17. It was felt that the subjects used in experiment 7 were, however, unrepresentative of a real inspector population. For this reason, a number of inspectors from the Quality Control Department of Ilford Ltd., which was analysed in detail in Chapter 5, were invited to take part in the study. Fifteen inspectors agreed to participate, which constituted nearly the whole of the examiner population. These subjects travelled to Birmingham on successive days, and took part in sessions 1 and 2, the fixed and variable probability conditions of experiment 7.

Subsequently, each group of subjects were administered the tests described in the next section.

Description of tests

The tests fell into three broad categories, corresponding to the three dimensions of cognitive skills discussed earlier.

The first two tests were designed to measure selectivity of attention, or field independence. The Group Embedded Figure Test (Witkin et al., (1971)) is an adaptation of Witkin's original Embedded Figures Test for group administration. The test consists of 18 complex figures, within which are embedded simpler figures. The complex figures are shaded, to emphasize large, organized Gestalts which serve to increase the difficulty of the disembedding task. The subject is prevented from simultaneously viewing the simple form and the complex figure containing it, by printing the simple forms on the back of the test booklet, and the complex figures on the booklet pages. The test

is divided into three sections, the first containing 7 very easy items for practice and each of the remaining two sections containing 9 more difficult items. Three minutes were allowed for the practice session and 5 for each of the remaining sections. The other test employed was the Closure Flexibility test, Thurstone and Jeffrey (1965). This is an adaptation of the Gottschaldt Figures from which most of the embedded figure test work originated. The test measures the ability of a subject to form a closure in the face of some distraction and was developed as a result of factor analytic studies by Thurstone (1944) and Pemberton (1951). Each item in the test consists of a figure, presented to the left of the page, followed by a row of four more complex drawings to the right. Some of the more complex drawings contain the given figure in its original size and orientation, and subjects are required to check the appropriate drawings. Ten minutes were allowed for this test. For both tests the subject's score was the number of correct answers given in the allotted time.

To assess distraction, five tests from Karp's Kit of Selected Distraction Tests (1962) were used. The Distracting Contexts Test I (DCT1) involves the subject locating a simple geometric figure within a matrix of extraneous lines and figures. In the distracting contexts test 2A (DCT 2A) the subject is required to locate a series of simple geometric figures within a large matrix of such figures. In the second version of this test (DCT 2B) as an added source of distraction, coloured overlays were superimposed over the simple figures. Two minutes were allowed for each of these tests. The arithmetic operations test consisted of 24 simple arithmetic problems spaced evenly on horizontal rows, interspersed with a series of irrelevant jokes, instructions and pictures. The subjects were given one

minute to complete as many of the problems as possible. The final distraction test was a cancellation task in which subjects were required to cross out the letters a, t and c each time they appeared on a page of randomly arranged single spaced letters. Three minutes were allowed for this test.

Two tests were employed to assess the shifting variable. The first of these had been employed in a study by Sack and Rice (op.cit.) and consisted of the first and last pages of an anagram test developed by Gardner, Lohrenz and Schoen (1968). After a practice list of anagrams, the subjects were given two minutes to solve as many of the 20 anagrams as possible. Their score was the number correct in this time. Mendelsohn et al., (1966) amongst others have postulated that anagrams require voluntary shifting in attention. The second test alleged to measure the shifting variable is the Reversed Triangles Test, Sanguiliano (1951). In this test the subject is given one minute to draw as many triangles as he can, each separate and with the apex upwards. This is then repeated with the apex downwards, and in the final one minute session he is required to alternate the triangles with points upwards and downwards. The score is the total number of triangles drawn during the last session.

In the test session the tests were presented in the order described, i.e.:

1. Group embedded figures test
2. Closure flexibility
3. DCT 1
4. DCT 2A
5. DCT 2B

6. Arithmetic operations
7. Cancellation
8. Triangles
9. Anagrams

Results

The raw scores for the various tests are in Appendix B. A multiple regression analysis was first performed using all 9 test scores as independent predictor variables, for the first fixed probability and the first variable probability session for all 36 subjects. The dependent variables considered were those that had been utilized in previous analyses i.e. the parametric and nonparametric measures of sensitivity and bias, the overall score obtained by the subjects and the false alarm and correct detection probabilities. An analysis of variance was performed on each multiple regression equation to test its significance.

For the first analysis, high multiple correlation coefficients were obtained, but because of the large number of predictor variables utilized, relative to the number of subjects, none of the regression analyses reached significance, although significant correlations were obtained for some of the individual tests. The analyses were therefore repeated, but prior to each multiple regression, those independent variables which did not account for a significant degree of the total variance were deleted. This procedure, which is a standard one for multiple regression analyses, produced a considerable improvement in the fit of the regression equations, virtually all of which were now significant. The results for the fixed probability session are shown below in Table 7.2. All the multiple

correlation coefficients are significant at $p < 0.05$.

A similar procedure was carried out for the variable probability session, and the best combination of predictor variables found as before. Again, all the multiple correlation coefficients are significant at $p < 0.05$. These results are given in Table 7.3. In both tables, the independent variables used in the regression are set out, and any significant correlations of those variables with the dependent variable are indicated. The significance test employed was a two-tailed t test in this case. The independent variables referred to by number in the tables are the following tests:

2. Group Embedded Figures Test
3. Closure Flexibility
4. Distracting contexts test 1
5. Distracting contexts test 2A
6. Distracting contexts test 2B
7. Arithmetic operations
8. Cancellation
9. Triangles
10. Anagrams

It will be recalled that tests 2 and 3 are related to the field independence variable, tests 4 - 8 to the ability to resist distraction variable, and tests 9 and 10 to the shifting of attention variable.

Discussion

In general the results are encouraging in that they provide reasonable support for the assumption that at least some of the cognitive skills

dependent variable	independent variables included in the regression	multiple correlation
Altham-Hammerton index	3, 4, 5, 7*, 8*	0.636
Pollack-Norman index	2, 5, 7, 8	0.516
P(A)	3, 4, 8, -10	0.554
d'	2, 5, 8	0.481
log beta	2**, -3, -4, 5	0.550
Hodos-Grier bias index	2*, -3, -4	0.503
Score	2*, -7*, 8*, 10	0.500
False alarm prob.	-2*, -3, -8	0.486
Correct detection prob.	2, 3*, 4, 8, -10	0.596

* = $p < 0.05$

** = $p < 0.01$

\- = negative correlation

Table 7.2 Summary of multiple regression results for fixed probability session.

dependent variable	independent variables included in the regression	multiple correlation
Altham-Hammerton index	2*, 5, -7, 8	0.555
Pollack-Norman index	2**, 5, -7*	0.571
P(A)	2*, 5	0.557
d'	2*, 5, -7*	0.525
log beta	2*, -3, -4*, -7, 9	0.587
Hodos-Grier bias index	2*, -3, -4, -7*, -9, 10	0.630
Score	2, -7**, 9, 10*	0.625
False alarm prob.	-2*, 7**, -10	0.639
Correct detection prob.	3*, 5, 9, 10	0.550

* = $p < 0.05$

** = $p < 0.01$

\- = negative correlation

Table 7.3 Summary of multiple regression results for varying probability session.

measured by the tests utilized in this study are related to performance on the simulated inspection task.

If we consider the sensitivity indices for both fixed and varying probability tasks certain general patterns emerge.

In both situations, sensitivity seems to be consistently related to the field independence dimension, as measured by tests 2 and 3 in the first session and test 2 in the second. Some of the tests of distractability also show clear relationships with the various sensitivity indices, tests 5 and 8 being positively related in both fixed and variable probability sessions. However, test 7, the Arithmetic Operations test, shows a positive correlation for the fixed probability session and a negative correlation for the variable probability situation. This will be discussed subsequently (n.b. a high score in the distractibility tests indicates low distractibility). The indices of distractibility apart from variable 7, show the expected relationship, i.e. the less distracted the inspector is by non-signal stimuli, the higher his apparent sensitivity is likely to be.

Examination of the bias measures indicates a negative correlation between the dependent variable and measures of distractibility, this time tests 3 and 4 in the fixed probability session and 3, 4 and 7 in the variable session. Since subjects who have a low criterion will be those who have not raised it to the level appropriate to the overall defect density in both sessions of 0.2, we can infer that the ability to modify the criterion in the appropriate direction seems to be negatively correlated with distractibility as measured by tests 3, 4 and 7. The results for test 7 for the sensitivity indices

can now be explained. In the fixed probability session, the ability to ignore irrelevent stimuli that is characterized by high scores on distractibility measures, leads to the positive correlation of variable 7 with sensitivity. It could be hypothesized that another aspect of low distractibility, at least as measured by test 7, might be an inability to readily change strategies in the light of a change in defect density. This would give rise to the negative correlation with apparent sensitivity in the variable session.

These conclusions are complicated by the intercorrelations that can be expected between the bias and sensitivity measures in the unequal variance case, and it is probably better to concentrate on the performance measures which do not exhibit these complications. Considering the overall score, this can be regarded as a function of both the sensitivity of a subject and the appropriateness of his response strategy. The significant correlations obtained with the Embedded figures test are clearly related to the strong relationship of the field independence measures to sensitivity. The negative correlation of test 7 with the score, confirms the earlier suggestion that a high score on this test implies a lack of flexibility of strategy. This could be due to an inability of the subject to observe the 'irrelevant' evidence of a change in defect probability. The presence of the shifting variables in connection with the score is also presumably related to the ability of the inspector to shift his attention from the primary focus of the task, to the secondary aspect of the defect density, and its possible changes.

The final dependent variables of interest are the correct detection and false alarm probabilities. In both fixed and varying probability

sessions the correct detection probability shows a significant correlation with field independence as measured by the Closure Flexibility Test. In view of the earlier correlation of this cognitive dimension with sensitivity, this result is not surprising. This also accounts for the negative correlation of variable 2 with the false alarm probability. Significantly, the false alarm probability shows a strong positive correlation with the arithmetic operations test 7, which lends weight to the earlier suggestion that a high score on this test is associated with an inflexible strategy.

To summarize these findings, it appears that sensitivity for defects on tasks similar to the one utilized in this study, is predicted fairly well by performance on field independence tests. The ability to modify the criterion optimally seems to be associated with a low score on the arithmetic operations test.

It should be emphasized at this point that the analytical techniques employed in this study have been somewhat crude. For example the selection of variables to include in the multiple regression may not have been optimal, since it was based primarily on the significance of these variables in the first analysis. A better procedure is to use a step-wise multiple regression, which includes variables in the regression in the order that they reduce the sum of squares of the variability. Work is continuing along these lines, but the results could not be included in this thesis. Finally, a much more sophisticated approach such as a factor analysis might have been appropriate. This was considered, but the limited sample size of 36 meant that this technique could not be meaningfully employed.

In spite of the crudity of the methods the degree of significant correlation found in this essentially exploratory study suggest that the approach is a fruitful one, particularly in comparison with the other attempts at predicting inspection performance, as discussed in Chapter 3. The next step involves testing the validity of these findings in a large scale industrial study.

7.4 Conclusions

The overall conclusions that emerge from experiments 6 and 7 are in line with the perceptual learning principles that have been established in laboratory tasks using simple signals. To this extent, the studies have performed a useful function in extending the results to a situation more typical of a real inspection task. Certain previously unreported findings have, however, emerged.

In terms of training an inspector to modify his criterion optimally in a changing defect density situation, the use of a fixed probability during the training session appears to produce the best results. A possible explanation for this is that the fixed defect density gives the trainee sufficient experience of a single probability that he is able to utilize this as an anchor point on the subjective probability continuum. A particular defect density would then be recognized as being greater or less than the anchor probability, and this would presumably facilitate the modification of the criterion in the appropriate direction. This model suggests that the appropriate training technique changes the recognition of the prevailing probability from an absolute to a comparative judgement task. All the available evidence suggests that the latter form of judgement is performed more

effectively than the former, e.g. Guilford (1954)

The finding that the most effective criterial adjustment occurred after training with the feedback + cuing + score condition is a logical consequence of regarding effective training as providing as much experience as possible of an anchor probability. The provision of the score information would in addition enable the trainee to develop appropriate strategies to maximize his payoffs. Such strategies are identical to those which optimize the criterion setting.

The general conclusions from the existing literature, that there is little to choose between cuing and KR as a means of enhancing sensitivity, where these two techniques produce the same information, is generally confirmed by the results. Sensitivity was increased significantly by training in which the cuing or KR was gradually reduced, as found by Abrams and Cook (op.cit.) and predicted by Wallis (op.cit.). The finding that a fixed probability during training is also necessary for an increase in sensitivity is at first sight difficult to account for, since the subject receives the same number of signal and noise samples under both fixed and varying probability conditions. One possibility is that the fixed probability sessions allow a more evenly spaced occurrence of the defect samples, and learning theory in general suggests that spaced practice is preferable to the massed practice that would be represented by the high defect probability blocks during the training sessions.

The results of the correlational study between cognitive skills as measured by certain pencil and paper tests and performance on the inspection task, suggest that this is a promising approach to the

question of selecting inspectors. Overall, the results suggest that sensitivity for defects can be predicted to some extent by tests which measure field independence, the ability to be able to perceive wanted configurations embedded in background noise which contains perceptual elements similar to the signal. This result is in accord with the predictions of SDT, that sensitivity is a function of the distance apart of the signal + noise and the noise distributions, whether this separation is due to external attributes of the signal, or to internal characteristics of the perceiver.

The finding of a consistent negative correlation between a test measuring the dimension of distractibility and the ability to maximize the payoffs in the inspection task was unexpected. If distractibility is regarded as the ability or otherwise to maintain a fixed focus of attention, however, the results become more comprehensible. It seems reasonable that such a quality might correlate negatively with the flexibility of strategy required to change the criterion to match the prevailing defect probability. The rigidity of attention that would be an asset in a situation where many external distractions are present, could be a disadvantage where more subtle aspects of the task needed to be noted, such as changes in defect density.

Although the present results are of considerable interest, they need to be replicated with a larger sample size, and with an industrial task, before the tests could be utilized as part of a selection procedure for inspectors.

CHAPTER 8 GENERAL CONCLUSIONS

8.0 INTRODUCTION

In this chapter the findings from the previous chapters will be presented and overall conclusions drawn. Finally, directions for further research will be discussed.

8.1 The literature review and its application to the analysis of inspection tasks

Chapter 2 provided a general overview of the SDT literature and established basic guidelines for applying SDT to inspection studies. Consideration of the studies that had utilized SDT suggested that the usefulness of SDT in the context of inspection tasks needed to be further investigated. The possibility of using SDT to examine the ability of inspectors to change their strategies emerged as a further experimental goal.

Problems were encountered when attempting to classify the inspection literature as a whole. Originally it had been hoped that the informal inspection model proposed in Chapter 1 would provide the basis for a classification scheme. It was found, however, that although this model provided a useful conceptual summary of the psychological and other areas relevant to inspection, it did not allow the major external factors influencing performance to be readily included.

Although the number of studies on inspection as such was limited, a very large number of theoretical areas were relevant to analysing such tasks. The review was, therefore, divided into two parts. The first part considered the major theoretical areas apart from SDT,

i.e. vigilance and visual search, and the second part utilized a simple classification scheme. This consisted of four main headings, i.e. task characteristics, environmental factors, organizational factors and individual factors. These main categories were divided into sub-categories as indicated below:

1. Task characteristics

a. Pacing and movement of the item being inspected

b. Magnification, lighting and other aids to enhance defect discriminability

c. Complexity

d. Display organization

e. Signal rate

f. Number of inspectors

g. Repeated inspection

2. Environmental factors

a. Heating

b. Lighting

c. Noise

d. Workplace layout

3. Organizational factors

a. Management and social aspects

b. Motivational variables

4. Individual factors

a. Selection

b. Visual abilities

c. Age

d. Sex

Although this scheme was somewhat ad hoc in nature, it encompassed most of the available literature and provided a means of structuring the subsequent analysis of the inspection tasks considered in the case studies.

8.2 The case studies

8.2.1 The data analysis group

The first case study was a re-analysis of an inspection system described by the author in an earlier study. The much broader range of literature available, and the insights obtained from the use of SDT enabled a much more sophisticated analysis of this task to be performed. The first goal of the study was to investigate if the SDT model applied to the data, and to examine the interrelationships between the parametric and non-parametric measures of sensitivity and bias. A number of tests of the SDT model were applied, most of them employing the detection and false alarm data. In general, the model, in its unequal variance form, fitted the results reasonably well. Corroborative evidence was obtained, from a consideration of the latency data, that the inspectors were employing a likelihood ratio criterion.

The parametric and non-parametric sensitivity and bias indices showed a high correlation, although it was demonstrated that they did not produce equivalent results if utilized in statistical analysis. It

was shownthat other commonly used inspection measures were correlated with both sensitivity and bias.

There were very few significant effects of the experimental variables on the SDT indices, apart from log beta, which showed a significant decline with time on task. This effect, which was in the opposite direction to that generally found in vigilance tasks, was explained in terms of the subject's perception of the experiment as a 'risky' situation, inducing an initially abnormally high criterion. With habituation to the experimental situation, the criterion declined to its usual level. No significant main effects were found for the noise variable, although some interactions with subjects occurred.

Analysis of the latency data was not attempted in detail, because scanning and decision times could not be separated. The mean latencies for the various categories of response were however, in accord with a simple extension of the SDT model to the latency situation. The unexpected lack of any significant differences in discriminability between the two types of signal employed, was explained as being due to the self-paced nature of the task. The inspectors were able to take longer to sample more attributes of the inherently less discriminable signals.

In general it was felt that the use of SDT in the analysis of this task provided insights which could not have been obtained by conventional measures.

8.2.2 The Ilford quality control system

The other case study, although it was also concerned with targets on film, presented completely different problems from the first. The difficulty of obtaining accurate estimates of the correct detection and false alarm probabilities limited the application of SDT in this study. A further difficulty was the highly subjective nature of the standards of product acceptability. A decision making model was proposed for the task, which suggested that SDT principles operated at two levels. The first of these determined the quality of the evidence that was obtained from the film in terms of the number of defects that were observed. Although the number of defects occurring was not usually precisely noted, the particular sensitivity and bias being employed by an inspector would determine his subjective estimate of the incidence of defects, which would in turn be used as evidence at the second stage of decision making. It was hypothesized that the inspector utilized a criterion at this stage which determined whether or not the film was acceptable as a whole. This criterion could be set at a different position than that employed at the first stage.

Two experiments were performed in an attempt to apply SDT to the task. In both cases considerable difficulty was experienced for the reasons discussed earlier, and the SDT parameters could only be obtained for the second experiment. Even in this case, the estimation of these quantities involved a number of untested assumptions. The most obvious characteristic of the results were the relatively low detection scores for defects which occurred, particularly in view of the considerable experience of the examiner subjects. This was felt

to be because the examiners normally made global estimates of the acceptability or otherwise of a film sample, rather than noting the incidence of individual defects. No statistically significant effects were obtained from the experiments, apart from subject differences, although a decline in log beta over time in experiment 3 similar to that observed in experiment 1, was noted.

Although the SDT parameters could not be easily determined in this situation, it was felt that the general principles of the SDT model provided useful insights into ways of improving the system. In particular, it was proposed that a form of training which would enable the examiners to modify their criteria in the optimal direction, in the event of a change in defect incidence, was desirable. The indications from management, that the ability to detect a change in quality was important, provided a spur for experimental investigations in this area in Chapter 5. Various suggestions of improvements were made to management as a result of the investigation. In order to facilitate the stabilization and standardization of the definitions of acceptable quality, it was proposed that a library of reference samples of defects be provided, together with examples of entire films that illustrated the required standards of quality. Regular 'calibration sessions' were also suggested to ensure homogeneous standards between different inspectors. It was felt that the relationship between the quality standards utilized by the inspectors and the criteria of acceptability of the customers should also be investigated.

8.3 The laboratory studies

8.3.1 The effects of between and within session defect probability changes

The first group of laboratory studies, comprising experiments 4 and 5, were concerned with the reaction of subjects to changes in the incidence of defects, both within and between sessions.

In the first experiment, performance was compared under four conditions. In the first two conditions, the subjects were required to detect signals which occurred at the same probability as they had been accustomed to during extensive practice sessions. In one case feedback in the form of a summary of performance was provided every hundred trials, whilst feedback was absent under the second condition. The other experimental conditions were similar to the first two, except that a much lower probability of signal was utilized. The results indicated that subjects were able to increase their criterion in the correct direction for the changed probability, but that this increase was only significant where feedback was provided. These results were considered from the standpoint of the amount of information available to the subject concerning the defect density. It was suggested that there were two sources of information available, which allowed the subject to revise his subjective probability estimates of the actual defect probability. These were information from the task and information from external sources. Where the defects were of low discriminability, as in this task, the subjective probability estimate, and hence the ability of the subject to modify the criterion, was primarily dependent on the external evidence available. There was some evidence that under the low probability conditions the subjects were

attempting to maximize their score by adopting a more nearly pure strategy in game theory terms. The latency data were in agreement with the unequal variance SDT model which had been found to fit the data as a whole. The extension of SDT to the latency scores which had been utilized in experiment 1 seemed to provide a reasonable description of the results.

Experiment 5 considered the subjects' reaction to within session changes in defect probability, with and without the provision of prior warning and feedback. It was found that prior warning that a change in defect probability would occur, was more effective in producing a criterion change in the required direction, than feedback from the task. Even in the absence of feedback, the subjects were able to modify their criteria appropriately. Without feedback, the degree of adjustment was more extreme than when feedback was provided.

Both experiments provided general support for the hypothesis that the degree of criterial change produced by a subject in a changing defect density situation was primarily a function of the amount of evidence available, from both within and outside the task, concerning the actual defect probability. Comparison of this study with others suggested that the greater the difficulty of the discrimination required, the more important external evidence became. In industrial situations, the provision of 'feedforward' information, giving the inspector prior warning of defect changes, was emphasized.

8.3.2 Training techniques for inspection

In experiments 6 and 7 two main questions were investigated. The first of these was concerned with testing the applicability of existing research on perceptual training, to tasks employing stimuli more representative of those found in real inspection tasks, than had hitherto been employed. The second objective was to investigate the possibility of devising training techniques which would both enhance the sensitivity of the inspector and also enable him to modify his criterion in a changing defect probability situation.

Both experiments 6 and 7 utilized stimulus material obtained from the Ilford inspection task described in Chapter 5, and the experiments were designed to simulate the critical features of this task.

Experiment 6 was a straightforward comparison between cuing and KR as training techniques, and a control condition. There were no significant changes in sensitivity as measured by the SDT indices, but the cuing training produced a significantly greater increase in the criterion towards the optimum, compared with the control condition. It was not, however, significantly different from the KR condition. The reason for the absence of a change in sensitivity was thought to be due to the complex nature of the signals employed, compared with those utilized in most perceptual training experiments. It seemed likely that many more training trials would be necessary to produce a significant change in sensitivity. The adjustment of the criterion in the appropriate direction after training sessions can be regarded as being due to the additional information on the signal distribution that these sessions provided.

The final experiment was designed to provide more evidence concerning the effectiveness of various types of training, and utilized a considerably greater number of training trials than experiment 6, in the light of the non-significant results obtained in that experiment. Two of the training conditions investigated were identical to those in the previous experiment, and the remainder utilized combinations of cuing, KR and/or the provision of the score information which indicated the subjects overall payoff from the task, given the payoff matrix assigned to the various response possibilities. Two further factors were included in the experimental design. These were the use of a fixed or VARYING defect probability during the training sessions, and either gradually reducing or keeping constant the amount of cuing or KR being provided. The variable probability factor was included to test the hypothesis that the inspector would be better able to adjust his criterion to a variable defect probability situation if he had previously experienced a range of known probabilities. The inclusion of the fading or otherwise condition was intended to establish whether the previous findings of Abrams and Cook (op.cit.), and Wallis (op.cit.), that cues or KR needed to be removed during training to produce a sensitivity increase, applied in this situation. The experiment employed two pre and two post training sessions, one with a fixed and the other with a varying defect probability.

Very few significant effects of training were found with the fixed probability session. No obvious reason was apparent for this finding, apart from the fact that the post training session may have been affected by the absence of the training aids that had been present during the immediately preceding two sessions.

With the variable probability sessions, a number of interesting effects were obtained. The combination of a fixed probability during training, with the training condition that provided both feedback, cuing, and the score information, appeared to produce the most optimal performance in a situation of changing defect density. It was suggested that the fixed probability condition enabled the inspector to build up an accurate perception of a particular defect incidence. This then provided an anchor point on the scale of subjective probabilities, that allowed him to accurately assign other observed probabilities as being greater or less than this probability. Consequently the inspector was able to modify his criterion more accurately in accord with the changing probabilities of defects. The particular training condition found to be most effective, provided the maximum information on the defect density, and also the score information that would allow the inspector to develop the response strategies that would maximize his payoffs.

Considering the sensitivity changes, it was found that, as predicted by the workers cited earlier, significant sensitivity changes were obtained in training sessions where the cuing or feedback was gradually reduced. It was also found that a fixed probability during the training session promoted the greatest increase in sensitivity. A possible explanation for this effect was the more even spacing of the defect samples that would occur during the fixed probability trials.

8.4 Cognitive skills as factors in the selection of inspectors

The scores from experiment 7 were used in conjunction with a number of tests of cognitive skills to determine if such tests could be used in selection procedures for inspectors. In order to make the

subject population more representative of those employed in inspection work, 16 inspectors from the Ilford quality control system performed the first two sessions of experiment 7 and their scores were included in the correlation analysis.

It was found that sensitivity, as measured by a number of parametric and non-parametric indices, seemed to be significantly related to performance on tests of the field independence dimension of cognitive skills. There was some indication that the ability of the subject to modify his strategy in a changing defect density situation was negatively correlated with performance on one of the distractibility tests, the arithmetic operations test. It was emphasized that the correlational study as it stood was exploratory in nature, and that further analysis could be performed on the data. The findings did indicate however, that with proper validation in an industrial context, this particular approach shows promise as a potential selection aid.

8.5 General conclusions

Both the industrial and the laboratory based studies have shown that SDT, as a general conceptual standpoint, offers unique advantages in suggesting ways in which industrial inspection systems can be optimized. It is suggested that a basic initial step in the analysis of any ongoing system is to perform an experimental study of its effectiveness which can be subsequently analysed in SDT terms, as shown in the industrial case studies. A particular strength of the SDT approach is that even where, as in the Ilford situation, precise measurements of the SDT parameters cannot be made, the insights that

can be gained by considering the system from this standpoint are still valuable. SDT has often been applied in a rather casual manner in many previous studies. An attempt has been made to show that the simple equal variance form of the model cannot be applied to any situation without initial tests to ensure that its underlying assumptions are fulfilled. The procedures that need to be observed in applying SDT to industrial situations have been spelt out in some detail. The use of the non-parametric indices of bias and sensitivity using real inspection data, has allowed some assessment of their usefulness. In general it is felt that their main application is in adding weight to conventional measures, provided these have been obtained using the appropriate model.

The survey of the literature produced a simple classification scheme which provides at least an initial approach to structuring the analysis of an inspection system. Used in conjunction with the SDT paradigm, it should provide a useful source of data in the design of new quality control systems. It is hoped, in the long run, to produce a more comprehensive handbook and classification scheme which should prove to be a useful aid to ergonomists working in this area.

The first laboratory experiments emphasized the importance of prior warning that changes were going to occur in the incidence of defects, in allowing the inspectors to adjust their criteria to maintain optimal performance. It could be argued that the function of a quality control system is to monitor sudden changes of this type and that prior warning does not usually occur. Although this is true in some situations, there is very often some indication at the production stage, that a higher incidence of defects can be expected. If this

information is effectively communicated to the inspection system, a far higher detection efficiency can be expected. The existence of 'feedforward' links of this type is of course a function of the organizational structure of which the quality control system is a part. The other result from the first two laboratory experiments was that feedback during the task was an effective means of maintaining an optimal criterion. Although most inspection systems do not provide such knowledge of results, in a direct way, there seems to be good grounds for recommending that re-inspection be carried out far more frequently, by senior inspectors, in order to provide such feedback. An additional advantage of this procedure would be that a greater degree of consensus would be produced between all the inspectors, as to what the appropriate criterion should be for a particular product. This is particularly important in a system such as at Ilford, where the levels of acceptability are essentially subjective, and are determined by a complex combination of factors, such as product, customer, and market conditions.

In the analysis of the experiments under discussion, the possibility that insights from probability learning theory might explain performance in changing defect density situations was put forward. Although the approach in this study has largely been from the stance of SDT, this should not lead us to ignore other relevant viewpoints. There seems to be many common areas between the two orientations that could usefully be explored in an inspection context.

The final experiments confirmed the general principle from previous studies of perceptual learning, that up to a point, the greater the information presented to the subject concerning the characteristics

of the defect and its distribution in time, the more effectively he was able to increase his sensitivity and adapt his criterion to the prevailing defect density situation. The experiments were limited in size and scope and therefore care needs to be taken in generalizing from them to other situations. Nevertheless the suggestion that subjects detect changes in defect probabilities by comparing them with a previously established single subjective probability, seems to be a reasonable one which fits the experimental data. Further work is needed to establish its validity more generally. The same considerations apply to the finding that a gradual reduction in the amount of cuing and feedback during training enhances sensitivity. This result apparently conflicts with the earlier finding concerning the importance of information in enhancing sensitivity. It seems clear that the provision of feedback or cuing is only effective up to a point in increasing sensitivity. Beyond that point, further information is counter productive, because it hinders the development of the internal referents necessary for true learning. The exact point at which the supplementary information should be reduced is probably a function of the individual task concerned, and the provision of general guidelines on this point would require more detailed research.

8.6 Directions for further research

As indicated in the introduction to the thesis, the research philosophy adopted has been to approach the area of quality controm from a number of directions, which were unified by the general orientation of SDT. The outcome of this research has been that a number of results of direct practical applicability have been obtained. The practical utility of the approach has to be weighed

against the fact that some degree of speculation has been necessary, and the size of the experiments has been smaller than might be found in the traditional academic investigation of a single clearly defined research hypothesis. It is argued that the orientation adopted has generated testable hypotheses that have relevance to real world problems and can therefore be verified in an industrial environment. Most of the recommendations for further research are concerned with the validation of these findings.

8.6.1 Validation of the two stage inspection model

In Chapter 5 a model was proposed which postulated two distinct decision making phases which might be expected to occur in inspection situations involving the aggregation of information over time. It is proposed that this model be validated both at Ilford and in other similar inspection systems.

8.6.2 Factors affecting the modification of the criterion

The discussion of Chapter 6 considered the sources of information that the inspector might utilize in modifying his subjective probability estimates, and the degree to which his response strategy actually changed, given that he perceived the on-going defect density accurately. Two experimental investigations are required to clarify these points. The first would involve changing the amount of evidence available from the two sources of information and manipulating its reliability. Studies of this type have been carried out by Ingleby (1974) in the context of auditory detection, but no corresponding studies have been performed for inspection. The second study

would involve performing an experiment similar to that described by Sims (1972) and discussed in detail in Chapter 6. In that experiment subjects were required to give their subjective probability that an item would be defective before inspecting it. It would be of considerable interest to investigate the relationship between this subjective probability, the value of the criterion adopted, and the actual defect density.

8.6.3 Verification of the perceptual training findings

As mentioned earlier, the findings from experiments 6 and 7 although intuitively reasonable, require further validation, using both laboratory simulation and real-life inspection tasks. In this case, it would be necessary to utilize a task which would allow the unambiguous calculation of the SDT parameters, and which would allow extensive trials to be taken.

Ilford have already expressed their willingness to incorporate the results of work on the perceptual training problem in their standard inspector training procedures. This would allow a longitudinal study of their usefulness to be performed.

8.6.4 Further work on the cognitive skills approach to selection

As discussed earlier, the current analyses of the cognitive skills tests and the inspection data must be regarded as provisional. Further analysis is under way, and additional correlational studies using industrial performance data are planned.

8.7 Concluding remarks

Although it is clear that we are only just beginning to be able to specify all the human factors requirements of an optimal inspection system, it is hoped that the data and techniques reported in this research will be of direct applicability in the design and analysis of quality control systems.

In addition to this objective, an attempt has been made to achieve a more general goal. This is to show that given an appropriate research strategy, the theoretical models of experimental psychology, exemplified in this case by SDT, can make a significant contribution to solving practical industrial problems. The approach adopted in this study, of combining field work with simulation and laboratory studies, unified by a common theoretical orientation, seems to be of potentially wide application.

There seems to be a considerable need at the present time for a bridge building operation between the concerns of much research in the behavioural sciences and the practical problems of society and industry. It is hoped that the research methodology presented in this study, whilst far from being a blueprint, will at least suggest some ways in which the construction of such bridges might begin.

REFERENCES

ABRAMS, A.J., & COOK, R.L., (1971) Information feedback: Contributions to learning and Performance in Perceptual Identification Training. Technical Bulletin STB 72-5 Naval Personnel and Training Research Laboratory, San Diego.

ADAMS, S.K., (1975) Decision making in quality control: some perceptual and behavioural considerations. In: Human Reliability and Quality Control (London: Taylor & Francis).

ALPERN, M., & BARR, L., (1962) Durations of the after images of brief light flashes and the theory of the Broca and Sulzer phenomenon. Journal of Optical Society of America, 52, 219-221.

ALTHAM, M.E., (1973) A non-parametric measure of signal discriminability. The British Journal of Mathematical and Statistical Psychology, 26.

ANNETT, J., (1959) Some Aspects of the Acquisition of Simple Sensori-motor Skills (D.Phil. Thesis, Oxford University).

ANNETT, J., (1961) The Role of Knoweedge of Results in learning: A Survey (NAVTRADEVCEN Technical Report No. 342-3, U.S. Naval Training Device Centre, New York).

ANNETT, J., (1966) Training for perceptual skills. Ergonomics, 9, 459-468.

ANNETT, J., (1969) Feedback and Human Behaviour (Harmondsworth: Penguin Books).

ANNETT, J., (1971) Sonar recognition training: An investigation of whole versus part and analytic versus synthetic procedures. (Technical Report No. 67-L-0105-L: U.S. Naval Training Devices Center, Orlando).

ANNETT, J., & CLARKSON, J.K., (1964) The use of cuing in training tasks. (U.S. Naval Training Device Center: Technical Report 3143-1).

ANNETT, J., & PATERSON, L., (1966) The use of cuing in training tasks: Phase II. (U.S. Naval Training Device Center: Technical Report 4119-1).

ANNETT, J., & PATERSON, L., (1967) The use of Cuing in Training Tasks: Phase III (Technical Report, NAVTRADEVCEN 4717-1, Orlando, Florida).

ASTLEY, R.W., & FOX, J.G., (1975) The analysis of an inspection task in the rubber industry. Human Reliability in Quality Control (London: Taylor and Francis).

AYERS, A.W., (1942) A comparison of certain visual factors with the efficiency of textile inspectors. Journal of applied psychology, 26, 812-827.

BADALAMENTE, R.V., & AYOUB, M.M., (1969) A behavioural analysis of an assembly line inspection task. Human Factors, 11, 339-352.

BADDELEY, A.D., & COLQUHOUN, W.P., (1969) Signal probability and vigilance: a reappraisal of the 'signal rate' effect. British Journal of Psychology, 60, 169-178.

BAKAN, P., & MANLEY, R., (1963) Effect of visual deprivation on auditory vigilance. British Journal of Psychology, 54, 115-119.

BAKER, C.H., (1963) Further towards a theory of vigilance. Vigilance: A Symposium (New York: McGraw-Hill) 127-153.

BAKER, E.M., (1975) Signal Detection Theory Analysis of Quality Control Inspector Performance. Journal of Quality Technology, 7(2), 62-71.

BAKER, C.A., MORRIS, D.F., & STEEDMAN, W.C., (1960) Target recognition in complex displays. Human Factors, 2, 51.

BAKER, E.M., & SCHUCK, J.R., (1975) Theoretical Note: Use of Signal Detection Theory to Clarify Problems of Evaluating Performance in Industry. Organisational Behaviour and Human Performance, 13, 307-317.

BANKS, W.P., (1970) Signal Detection Theory and Human Memory. Psychological Bulletin, 74, 81-99.

BELBIN, R.M., (1957) New fields for quality control. British Management Review, 15, 79-89.

BELT, J.A., (1971) The Applicability of Vigilance Laboratory Research to a Simulated Inspection Task. U.S. Government Report No. A.D. 728-490.

BLACKWELL, H.R., (1959) Specification of interior illumination levels. Illuminating Engineering, 43, 906-931.

BLACKWELL, H.R., (1952) Journal of Experimental Psychology, 44, 306.

BLOOMFIELD, J.R., (1970) Visual Search. (Ph.D. Thesis, University of Nottingham).

BLOOMFIELD, J.R., (1975) Theoretical Approaches to Visual Search. Human Reliability and Quality Control, (London: Taylor and Francis)

BLOOMFIELD, J.R., (1975) Classifying studies of visual inspection. Human Reliability and Quality Control (London: Taylor and Francis).

BROADBENT, D.E., & GREGORY, M., (1963) Vigilance considered as a statistical decision. British Journal of Psychology, 54, 309-323.

BROADBENT, D.E., (1971) Decision and Stress (London: Academic Press).

BROCK, J.F., WELLS, R.G., & ABRAMS, M.L., (1974) Development and Validation of an Experimental Radiograph Reading Training Program. Navy Personnel Research and Development Center, San Diego, NPRDC-TR74-33.

BROWNE, R.W., (1965) On-the-Job Training of the Aerospace Nondestructive Test Inspector, Materials Evaluation, October 1965, 489-492.

BUCK, J.R., (1975) Dynamic Visual Inspection, Human Reliability and Quality Control (London: Taylor and Francis).

BUSH, R.R., & MOSTELLER, F., (1955) Stochastic Models for Learning. (New York: Wiley).

CAMPBELL, R.A., (1964) Feedback and noise-signal detection at three performance levels. Journal of Acoustic Society of America, 36, 434-438.

CHANEY, F.B. & HARRIS, D.H., (1966) Human Factors techniques for quality improvement. 20th Annual Conference A.S.Q.C.: Technical Conference Transactions, 400-413, New York.

CHANEY, F.B., & TEEL, K.S., (1967) Improving inspector performance through training and visual aids. Journal of Applied Psychology, 51(4), 311-315.

CHAPMAN, D., & SINCLAIR, M.A., (1975) Applications of Ergonomics in Inspection Tasks in the Food Industry. Human Reliability and Quality Control (London: Taylor and Francis).

COCHRAN, D., PURSWELL, J.L., HOAG, L., (1973) Development of a Prediction Model for Dynamic Visual Inspection Tasks. Proceedings of the Seventeenth Annual Meeting of the Human Factors Society, Santa Monica, California.

COCKRELL, J.T., & SADACCA, R., (1971) Training Individual Image Interpreters Using Team Consensus Feedback. U.S. Army Behaviour and Systems Research Laboratory, Technical Research Report 1171.

COLQUHOUN, W.P., (1960) Temperament, Inspection efficiency, and time of day. Ergonomics, 3, 377-378.

COLQUHOUN, W.P., (1959) The effect of a short rest-pause on inspection efficiency, Ergonomics, 2, 367-372.

COLQUHOUN, W.P., (1961) The effect of unwanted signals on performance in a vigilance task. Ergonomics, 4, 41-5.

COLQUHOUN, W.P., (1967) Sonar target detection as a decision process. Journal of Applied Psychology, 51, 187-190.

COLQUHOUN, W.P., & BADDELEY, A.D., (1964) Role of pretest expectancy in vigilance decrement. Journal of Experimental Psychology, 68. 156-160.

COLQUHOUN, W.P., & BADDELEY, A.D., (1967) Influence of signal probability during pretraining on vigilance decrement. Journal of Experimental Psychology, 73, 153-155.

COOMBS, C.H., DAWES, R.M., & TVERSKY, A., (1970) Mathematical psychology, (New Jersey: Prentice-Hall).

CORNSWEET, T.N., (1970) Visual Perception (New York: Academic Press)

CRAWFORD, W.A., (1960) Perception of Moving Objects, IV, The Accuracy of fixation required in the Perception of Detail in Moving Objects. Air Ministry, Flying Personnel Research Committee Memo ISOD - London.

DAVIES, D.R., & TUNE, G.S., (1970), Human Vigilance Performance (London: Staples Press).

DORFMAN, D.D., & ALF, E. JR., (1968) Maximum likelihood estimation of parameters of signal detection theory - A direct solution. Psychometrika, 33,(1).

DORFMAN, D.D., & ALF, E. JR., (1969) Maximum likelihood estimation of signal detection theory and determination of confidence intervals - rating method data. Journal of Mathematical Psychology, 6, 487-96.

DRURY, C.G., (1975) Inspection of Sheet materials: Model and Data. Human Factors, 17 (3), 257-265).

DRURY, C.G., (1975) Human decision making in quality control. Human reliability and quality control (London: Taylor and Francis).

DRURY, C.G., (1973) The effect of speed of working on industrial inspection accuracy. Applied Ergonomics, 4 (1), 2-7.

DRURY, C.G., (1973) The inspection of sheet materials - model and data. 17th Annual Meeting of the Human Factors Society, Washington, October 1973, 457-464.

DRURY, C.G., & ADDISON, J.L., (1973) An industrial study of the effects of feedback and fault density on inspection performance. Ergonomics, 16, 159-169

DRURY, C.G., & SHEEHAN, J.J. (1969) Ergonomic and economic factors in an industrial inspection task. International Journal of Production Research 7 (4), 333-341.

DUSOIR, A.E., (1975) Treatments of bias in detection and recognition models: A review. Perception & Psychophysics, 17 (7), 167-178.

EDWARDS, W., (1961) Probability learning in 1000 trials. Journal of Experimental Psychology, 62, 385-394.

EDWARDS, W., (1962) Dynamic decision theory and probabilistic information processing. Human Factors, 4, 59-73.

EDWARDS, W. & PHILLIPS, L.D., (1964) Man as a transducer for probabilities in Bayesian Command and control systems. In G.L. BRYAN & M.W. SHELLEY (Editors), Human Judgements and Optimality, (New York: Wiley).

EHLERS, H.W., (1972) Effects of low frequency continuous noise on an inspection task. U.S. Army Logistics Management Center, Texarkana, Texas, Intern Training Center Report No. USAML-IT6-2-72-02.

EILON, S., (1961) Recirculation of products through an inspection station. International Journal of Production Research, August 1961, 39-44

EGAN. J.P., & CLARKE, F.R., (1966) Psychophysics and signal detection . In: Experimental Methods and Instrumentation in Psychology, (New York: McGraw - Hill).

EGAN, J.P., SCHULMAN, I., & GREENBERG, G.Z., (1959) Operating characteristics determined by binary decisions and by ratings. Journal of the Accoustical Society of America, 31, 768-73

ELLIOTT, E., (1960) Perception & alertness. Ergonomics, 3, 357-364

EMBREY, D.E., (1970) The Bubble Chamber Data Analysis Group. A Study of the Organisation and Human Factor Aspects of a Large Research Group. Unpublished B.Sc. Thesis, Department of Physics, University of Birmingham.

EMBREY, D.E., (1975) Training of the Inspectors' Sensitivity and Response Strategy. In: Human Reliability in Quality Control. (Taylor & Francis, London pp 123-131.

ERICKSON, R.A., (1964) Relation between visual search time and peripheral visual acuity. Human Factors, 6, 165-177.

EVANS, R.N., (1951) Training improves micrometer accuracy. Personnel Psychology, 4, 231-242.

ESTES, W.K., & JOHNS, M.D., (1958). Probability learning with ambiguity in the reinforcing stimulus. American Journal of Psychology, 71, 219-228.

FARINA, A.J., & WHEATON, G.R.,(1971) Development of a Taxonomy of Human Performance. The task characteristics approach to performance prediction. American Institutes for Research, Washington Technical Report 7, Contract Nos. F44620-67-0116.

FAULKNER, T.W., & MURPHY, T.J., (1973) Lighting for difficult visual tasks. Human Factors, 15 (2), 149-162.

FECHNER, G.T., (1860) Elemente der Psychophysik (Leipzig, Breitkopf & Hartel).

FLEISHMAN, E.A., & STEPHENSON, R.W., (1970) Development of a Taxonomy of Human Performance: A review of the third years progress. Report No. AIR-726-9/70-TPR, American Institutes for Research, Washington, D.C.

FOX, J.G., (1964) The ergonomics of coin inspection. Quality Engineer, 28, 165-169.

FOX, J.G., (1973) Sustaining vigilance in inspection. Paper at E.R.S. Meeting Human Reliability in Quality Control - Birmingham.

FOX, J.G., & HASELGRAVE, C.M., (1969) Industrial inspection efficiency and the probability of a defect occurring. Ergonomics, 12, 713-721.

FREEMAN, H.A.., FRIEDMAN, M., MOSTELLER, F., & WALLIS, W.A. (Eds.), (1948), Sampling Inspection. (New York, McGraw - Hill).

FREEMAN, P.R., (1973) Tracts for Computers XXX: Table of d′ and Beta. Cambridge: Cambridge University Press).

FROOT, H.A., & DUNKEL, W.E., (1975) Visual inspection of integrated circuits: A case study. Human Reliability in Quality Control (London, Taylor & Francis).

GALE, A., BULL, R., PENFOLD, V., COLES, M., & BARRACLOUGH, R., (1972) Extraversion, time of day, vigilance performance and physiological arousal: failure to replicate traditional findings. Psychonomic Science, 29, 1-5.

GARDNER, R.W., LOHRENZ, L.J., & SCHOEN, R.B., (1968) Cognitive control of differences in the perception of persons and objects. Perceptual and Motor Skills, 26, 311-330.

GARDNER, R.W., & MORIARTY, A., (1968) Personality structure at pre-adolescence..(Seattle:University of Washington Press).

GIBSON, E.G., (1953) Improvement in perceptual judgements as a function of controlled practice or training. Psychological Bulletin, 50, 401-431.

GILLIES, G.J., (1975) Industrial applications in the glass industry. Human Reliability and Quality Control, (London: Taylor & Francis).

GRANT, D.A., & HORNSETH, J.P., (1951) Aquisition and Extinction of a verbal conditional response with differing percentages of reinforcement. Journal of Experimental Psychology, 42, 1-5

GREY, D.R., & MORGAN, B.J.T. (1972) Some aspects of ROC curve-fitting: normal and logistic models. Journal of Mathematical Psychology, Volume 9.

GRIER, J.B., (1971) Nonparametric indexes for sensitivity and bias. Psychological Bulletin Vol. No. 75.

GUILDFORD, J.P., (1954) Psychometric Methods (New York: McGraw-Hill).

HAMMERTON, M., & ALTHAM, P.M.E., (1971) A nonparametric alternative to d'. Nature, 234, 487-488.

HARRIS, D.H. (1964) Development and Validation of an aptitude test for inspectors of electronic equipment. Journal of Industrial Psychology, 2, 29-35.

HARRIS, D.H., (1966) Effect of equipment complexity on inspection performance. Journal of Applied Psychology, 50, 236-237.

HARRIS, D.H., (1968) Effect of defect rate on inspection accuracy. Journal of Applied Psychology, 52, 377-79.

HARRIS, D.H. (1969) The nature of industrial inspection. Human Factors, 11, (2), 139-148.

HARRIS, D.H. & CHANEY, F.B., (1969) Human factors in quality assurance. (New York: Wiley).

HEIMSTRA, N.W., ELLINGSTAD, V.S. & DEKOCK, A.R., (1967) Effects of operator mood on performance in a simulated driving task. Perceptual & Motor Skills, 25, 729-735.

HODOS, W., (1970) Nonparametric index of response bias for use in detection and recognition experiments. Psychological Bulletin, 74, (5).

INGLEBY, J.D., (1974) Further studies of the human observer as statistical decision maker. Organisational Behaviour and Human Performance, 12.

JACOBSON, H.J. (1953) A study of inspector accuracy. Engineering Inspection, 17, (2-10), 1953.

JAMIESON, G.H , (1966) Inspection in the telecommunications industry: A field study of age and other performance variables. Ergonomics, 9, 297-303.

JENKINS, H.M., (1958) The effect of signal rate on performance in visual monitoring. American Journal of Psychology, 71, 647-661.

KAPPAUF, W.E., CROWDER, W.F., McDIARMID, C.G., & DAVIES, J.D., (1955) Performance during prolonged watch-keeping at a visual detection task involving search. (University of Illinois Memor. Report 14-10, Urbana).

KAPPAUF, W.E., & POWE, W.E., (1959) Performance decrement on an audio visual checking task. Journal of Experimental Psychology, 57, 49-56.

KARP, S.A., (1962) A factorial study of overcoming embeddedness in perceptual and intellectual functioning. Unpublished doctoral dissertation, New York University.

KARP, S.A., (1962) Kit of selected distraction tests. Cognitive tests: State University of New York, Downstate Medical Center.

KARP, S.A., (1963) Field dependence and overcoming embeddedness. Journal of Consulting Psychology, 27, (4), 294-302.

KEPPEL, G., (1973) Design and Analysis: A Researcher's Handbook. Englewood Cliffs, New Jersey: Prentice-Hall Inc.).

KIBLER, A.W.,,(1965) The relevance of vigilance research to aerospace monitoring tasks. Human Factors 7, (2), 93-99).

KIRK, R.E., (1968) Experimental design: procedures for the behavioural sciences. (Monterey, California: Brooks/Cole.)

KIRK, R.E., & HECHT, E., (1963) Maintenance of vigilance by programmed noise. Perceptual and Motor Skills, 16, 553-560.

KRKOVIC, A., & SVERKO, B.,(1967) Characteristics of performance on a new variety of vigilance task. Acta Instituti Psychologici Universitatis Zagrebiensis, No. 54.

LAU, A.W., (1966) Doppler Discrimination as a function of variations in dimensions of the sonor echo. U.S. Naval Personnel Research Activity: Technical Bulletin STB 66-25.

LEE, W., (1969) Relationships between Thurstone category scaling and signal detection theory. Psychological Bulletin, 71, 101-7.

LEE, W., (1971) Decision Theory and Human Behaviour. (New York: Wiley).

LEVINE, J.M., ROMASHKO, T., & FLEISHMAN, E.A., (1971) Development of a taxonomy of human performance. Evaluation of an abilities classification system for integrating and generalising research findings. American Institutes for Research, Washington Technical Report 12, Contract No. DAHL 19-71-L-0004.

LINK, H.C., (1920) Employment Psychology, (New York: McMillan).

LION, J.S. (1964) The performance of manipulative and inspection tasks under Tugsten and fluorescent lighting. Ergonomics, 7, 51-61.

LION, J.S., RICHARDSON, E. & BROWNE, R.C., (1968) A study of the performance of industrial inspectors under two kinds of lighting. Ergonomics, Vol. 11, No. 1, 23-34.

LION, J.S., RICHARDSON, E., WEIGHTMAN, D., & BROWNE, R.C., (1975). The influence of the visual arrangement of material and of working singly or in pairs, upon performance at simulated industrial inspection. Ergonomics, 18, (2), 195-204.

LION, J.S., RICHARDSON, E., BROWNE, R.C., WEIGHTMAN, D. (1975) The influence of the visual arrangement of material at simulated industrial inspection. Ergonomics, 11, (1), 23-34.

LUDVIGH, E.J., & MILLER, J.W., (1958) Study of visual acuity during ocular pusuit of moving test objects. Journal of Optical Society of America, 48, 799-802.

LUSTED, L.B., (1971) Signal detectability and medical decision making. Science, 171, 1217.

MACKWORTH, J.F.,(1969) Vigilance and habituation. (Harmondsworth- Penguin Books)

MACKWORTH, J.F., (1970) Vigilance and attention. (Harmondsworth: Penguin Books)

McCann, P.H., (1969) The effects of ambient noise on vigilance performance. Human Factors, 11, 251-256.

MCCORMACK, R.L., (1961) Inspector Accuracy: A study of the literature (SCTM-53-61 (14), Sandia Laboratories, Albuquerque, New Mexico.

MCFARLING, L.H., (1974) Noise effects, sex differences, and task pacing in simulated inspection. (Ph.D., Thesis, University of South Dakota).

MCKENZIE, R.M., (1958) On the accuracy of inspectors. Ergonomics pp 225-272.

MCKENZIE, R.M., & PUGH, D.S., (1957) Some Human Aspects of Inspection Journal of the Institute of Production Engineers Vol. No. 36.

MCNICOL, D., (1972) A primer of signal detection theory. (London, George Allen and Union Ltd.)

MARTINEK, H., & SADACCA, R., (1965) Error Keys as reference aids in image interpretation. U.S. Army Personnel Research Office, Washington, Technical Research Note 153.

MENDELSOHN, G.A., GRISWOLD, B.B., & ANDERSON, M.L., (1966) Individual differences in anagram-solving ability. Psychological Reports, 1966, 19, 799-809.

MILLER, R.B., (1962) Task description and analysis. Psychological Principles in System Development (New York: Holt, Rinehart and Winston).

MILLER, R.B., (1971) Development of a taxonomy of human performance: Design of a systems task vocabulary. Technical Report II American Institutes for Research, Washington.

MITCHELL, J.H., (1935) Subjective stands in inspection for appearance. Human Factor 9, 1935.

MITTEN, L.G., (1957) Research team approach to an inspection operation. In - Introduction to Operations Research, (New York: Wiley).

MORAAL, J., (1975) Analysis of an industrial inspection task. Human Reliability and Quality Control (London, Taylor & Francis).

MORRISSETTE, J.O., HORNSETH, J.P., & SHELLAR, K. (1975) Team organisation and monitoring performance. Human Factors 17, (3), 296-300.

NAVON, D., (1975) A simple method for latency analysis in signal detection tasks. Perception and Psychophysics 18 (1), 61-64.

NEAL, G.L., & PEARSON, R.G., (1966) Comparative effects of age, sex, and drugs, upon two tasks of auditory vigilance. Perceptual and Motor Skills, 23, 957-974.

NELSON, J.B., & BARANY, J.W., (1969) A dynamic visual recognition test for paced inspection tasks. AIIE Transactions, 1, (4), 327-332.

OGILVIE, J.C., & CREELMAN. C.D., (1968) Maximum likelihood estimation of Receiver Operating Characteristics curve parameters. Journal of Mathematical Psychology, 5, 377-91

PASTORE, R.E., & SCHEIRER, C.J., (1974) Signal Detection Theory: Considerations for General Application. Psychological Bulletin, 81, No. 12, 945-958.

PEMBERTON, C.L., (1951) A study of the speed and flexibility of closure factors. Unpublished Ph.D. dissertation, Department of Psychology, University of Chicago.

PERRY, G., (1968) Lighting for Inspection. (Technical Note No. 115., British Glass Industry Research Association).

PIKE, A.R., (1971) The latencies of correct and incorrect responses in discrimination and detection tasks: Their interpretation in terms of a model based on a simple counting. Perception & Psychophysics, 9, pp 455-460.

PINNEO, L., (1966) On noise in the nervous system. Psychological Review, Vol. No. 73, No. 3.

POLLACK, I., & HSIEH, R., (1969) Sampling variability of the area under the ROC curve and of d'. Psychological Bulletin Vol. 71, No. 3.

POLLACK, I,, & NORMAN, D.A., (1964) A nonparametric analysis of recognition experiments. Psychonomic Science, Vol. No. 1.

POLLACK, I., NORMAN, D.A., GALANTER, E., (1964) An efficient non-parametric analysis of recognition memory. Psychonomic Science 1, 327-328.

POULTON, E.C., (1969) Bias in experimental comparisons between equipments due to the order of testing. Ergonomics, 12, (4), 679-687

POULTON, E.C., (1973) The effect of fatigue upon inspection work. Applied Ergonomics, 4.2, 73-83.

POWERS, J.R., BRAINARD, R.W., ABRAM., R.E., & SADACCA, R., (1973) Training techniques for rapid target detection. U.S. Army Research Institute for the Behavioural & Social Sciences, Technical Paper 242.

PURSWELL, J.L., etal. (1972) An inspection task experiment. Proceedings of the 16th Annual Meeting of the Human Factors Society, Vol. 1., 297-300.

RAPHAEL, W.S., (1942) Some Problems of Inspection. Occupational Psychology, 16, (4), 157-163.

RIZZI, A.M., BUCK, J.R., & ANDERSON, V.L., (1974) Effects of some task variables on conver-paced visual inspection accuracy. (Working Paper, School of Industrial Engineering, Purdue University.)

ROETHLISBERGER, F.J., & DICKSON, W.J., (1939) Management and the Worker. (Cambridge: Harvard University Press).

SACK, S.A., & RICE, C.E., (1974) Selectivity, resistance to distraction and shifting as three attentional factors. Psychological Reports, 34, 1003-1012.

SAKGUCHI, T., & NAGAI, H., (1973) Studies on relation between various light sources and visual fatigue. Journal of the Illuminating Engineering Institute of Japan, 57, (5), 4-13.

SANGUILIANO, I.A., (1951) An investigation of the relationship between the perception of the upright in space and several factors in personality organization. Unpublished doctoral dissertation, Fordham University.

SARTAIN, A.Q., (1945) The use of certain standardized tests in the selection of inspectors in an aircraft factory. Journal of Consulting Psychology, 9, 234-235.

SCHLEGEL, R E., BOARDMAN, D.W. & PURSWELL, J.C., (1973) A comparison of single and multiple inspection systems. Proceedings of the 17th Annual Meeting of the Human Factors Society, Santa Monica, California.

SCHOONARD, J.W., & GOULD, J.D. (1973) Field of view and target uncertainty in visual search and inspection. Human Factors, 15, (1), 33-42.

SCHUMAN, J.T., (1945) The value of aptitude tests for factory workers in the aircraft engine and propller industries. Journal of Applied Psychology, 29, 156-163.

SCOTT BLAIR, G W , & COPPEN, F.M.V., (1942) The subjective conception of the firmness of soft materials. American Journal of Psychology, 5, 127-139.

SEALE, S.J., (1972) Some psychometrics in relation to target acquisition. British Aircraft Corporation Ltd., Guided Weapons Division: Bristol Works, Target Acquisition Research Group B15-1-5.

SHEEHAN, J.J., & DRURY, C.G., (1971) The analysis of industrial inspection. Applied Ergonomics, 2, 74-78.

SHEFT, D.J., JONES, M.D., BROWN, R.F., & Ross, S.E. (1970) Radiology, 94, 427.

SIEGEL, S., & GOLDSTEIN, D.A., (1959) Decision-making behavior in a two-choice uncertain outcome situation. Journal of Experimental Psychology, 57, 31-42.

SIEGEL, A.I., & WOLF, J., (1969) Man Machine Simulation Models. (New York: Wiley).

SIMS, L.P., (1972) "An analysis of inspector perception of quality levels". M.Sc. Thesis, Auburn University, Alabama.

SINCLAIR, M., (1971) Paced and unpaced inspection of a bakery product. Paper at E.R.S. Annual Conference.

SLOVIC, P., FISCHHOFF, B., & LICHTENSTEIN, S., (1975) Cognitive Processes and Societal Risk Taking. Oregon Research Institute.

SMITH, G.L. JR., (1975) Signal detection theory and industrial inspection. In, Human Reliability & Quality Control (London: Taylor & Francis)

SMITH, G.L., & ADAMS, S.K. (1971) Magnification and microminiature inspection. Human Factors, 13, (3), 247-254.

SMITH, L.A., & BARANY, J.W., (1970) An elementary model of human performance on paced visual inspection tasks. AIIE Transactions, Volume II, No. 4., 298-308, 1970.

SMITH, L.A., & BARANY, J.W., (1971) An elementary model of human performance on paced visual inspection. A.I.I.E. Transactions, 4, 298-308.

SMITH, R.L., LUCACCINI, L.F., GROTH, H, & LYMAN, J., (1966) Effects of anticipatory alerting signals and a compatible socondary task on vigilance performance. Journal of Applied Psychology, 50, 240-246

SMITH, R.L., & LUCACCINI, L.F., (1969) Vigilance Research: Its application to industrial problems. Human Factors, 11, 149-156.

SOSNOWEY, J.K., (1967) An investigation of the effects of incoming quality and inspection rate on inspector accuracy. (M.S.I.E. Thesis, Texas Technological College).

STAEL von HOLSTEIN, C.A.S., (1971) The effect of learning on the assessment of subjective probability distributions. Organisational Behaviour and Human Performance, 6, 304-315.

SURY, R.J., (1964) An industrial study of paced and unpaced operator performance in a single stage work task. International Journal of Production Research 3, (1), 91-98.

SWAIN, A.D., (1972) Design techniques for improving human performance in production.(London: Industrial and Commercial Techniques Ltd.)

SWETS, J.A., (1963a) Control factors in auditory frequency selectivity. Psychological Bulletin, 60, 429-440.

SWETS, J.A., (1964) Signal Detection and recognition by human observers. (New York: Wiley).

SWETS, J.A., (1973) The relative operating characteristics in Psychology Science, (Vol. 182).

SWETS, J.A., & GREEN, D.M., (1966) Signal Detection Theory and Psychophysics). (New York: Wiley)

SWETS, J.A., MILLMAN, S.H., FLETCHER, W.E., & GREEN, D.M., (1962) Learning to identify non verbal sounds: An application of a computer as a teaching machine. Journal of Acoustical Society of America, 34, 928-935.

SWETS, J.A., TANNER, W.P., & BIRDSALL, T.G., (1961) Decision processes in perception. Psychological Review, 68, 301-340.

TANALSKI, T.G., (1956) The eyes have it - At Convair. Industrial Quality Control, 12, 9-10.

TAYLOR, M.M., (1967) Detectability theory and the interpretation of vigilance data. Acta Psychologica, 27, 390-399.

THEODOR, L.H., (1972) Some comments on 'A table for the calculation of d' & beta'.

THOMAS, L.F., & SEABORNE, A.E.M., (1961) The socio-technical context of industrial inspection. Occupational Psychology Vol. 35, pp 36-43.

THOMAS, L.F., (1962) Perceptual organization in industrial inspectors. Ergonomics, 5, 429-434.

THOMPSON, L.W., OPTON, E., & COHEN, L.D., (1963) Effects of age, presentation speed, and sensory modality on performance of a vigilance task. Journal of Gerontology, 18, 366-369.

THORNTON, C.L., BARRETT, G.V. & DAVIS J.A., (1968) Field dependence and target identification. Human Factors, 10, (5), 493-496.

THURSTONE, L.L., (1944) A Factual Study of Perception. Psychometric Monographs No. 4 (Chicago: University of Chicago Press).

THURSTONE, L.L., & JEFFREY, T.E., (1965) Closure Flexibility (Concealed Figures) Test (Industrial Relations Center, University of Chicago).

TIFFIN, J., & ROGERS, H.B. (1941) The selection and training of inspectors. Personnel, 18, (1), 14-31.

TRIMBY, H. (1959) The development of industrial vision screening in the United States, with special reference to work done with the B & L Ortho-Rater. "The Optician".

TUKEY, J.W., (1949) One degree of freedom for non additivity. Biometrics, 5, 232-242.

VIRSU, V., (1972) Inspection and presentation of radiographs and micrographs: An applied review of basic visual parameters. Reports from the Institute of Psychology, University of Helsinki, Report No. 3.

WAAG, W L., HALCOMB, C.G. & TYLER, D.M., (1973) Sex differences in monitoring performance. Journal of Applied Psychology, 58, 272-274.

WALLACK, P M., (1967) An experimental investigation of industrial inspector accuracy under varying levels of product defectiveness. (Unpublished Doctoral Dissertation, Oklahoma State University, Stillwater, Oklahoma).

WALLACK, P.M., & ADAMS, S.K., (1969) The utility of signa-detection theory in the analysis of industrial inspector accuracy. AIIE Transactions, 1, (1), 33-44.

WALLACK, P.M., & ADAMS, S.K., (1970) A comparison of inspector performance measures. AIIE Transactions, II,(2), 97-105.

WALLIS, D., (1963) Occupational Psychology in the sixties: Some implications of recent studeies of perceptual training and skill. Occupational Psychology, 37, No. 4, 237-253.

WELFORD, N.T., (1952) S.E.T.A.R. A sequentail event timing and recording apparaturs. Journal of Scientific Instruments, 29, 1-4.

WHITFIELD, D., (1975) Man-computer symbiosis: A 1975 review. A.P. Report 57, Applied Psychology Department, Unviersity of Aston in Birmingham.

Whittenberg, J.A., & ROSS, S., (1953) A study of three measures of perceptual efficiency during sustained vigilance (University of Maryland Tech. Rep. 14, College Park.

WIENER, E.L., (1963) Knowledge of results and signal rate in monitoring: A transfer of training approach. Journal of Applied Psychology, 47, 214-222.

WIENER, E.L., (1967) Transfer of training from one monitoring task to another. Ergonomics, 10, 649-58

WIENER, E.L., (1968) Training for vigilance - repeated sessions with knowledge of the results. Ergonomics, 11, 547-556.

WIENER, E.L., (1969) Money and the monitor. Perception and Motor Skills, 29, 627-634.

WIENER, E.L. (1975) Individual and group differences in inspection. In: Human reliability and quality control (London:-Taylor & Francis).

WIENER, E.L., & ATTWOOD, D.A., (1968) Training for vigilance. Combined cueing and knowledge of results. Journal of Applied Psychology, 6, 474-479.

WILKINSON, R.T., (1960) The effect of lack of sleep on visual watch keeping. Quarterly Journal of Experimental Psychology, 12, (1), 36-40.

WILKINSON, R.T., (1964) Artificial 'signals' as an aid to an inspection task. Ergonomics, 7, 63-72.

WILLIAMS, L.G., & BOROW, M.S., (1963) The effect of rate and direction of display movement upon visual search. Human Factors, 5, (2), 139-146.

WILLIGES, R.C. & STREETER, H. (1971) Display Characteristics in Inspection Tasks. Journal of Applied Psychology 55, (2). 123-125.

WINER, B.J., (1962) Statistical Principles in experimental design. (New York: McGraw-Hill).

WITKIN, H.A., (1950) Individual Differences in Ease of Perception of Embedded figures. Journal of Personality 19, 1-15.
Perception

WITKIN, H.A., LEWIS, H.B., HERTZMAN, M., MACHOVER, K., MEISSNER, P.B., & WARNER, S., (1954) Personality through perception. (New York: Harper).

WITKIN, H.A., DYK, R.B., FATERSON, H.F., GOODENOUGH, D.R., & KARP, S.A. (1962) Psychological Differentiation (New York: Wiley).

WITKIN, H.A., Oltman, P.K., RASKIN, E., KARP, S.A., (1971). A manual for the embedded figures test. (Palo Alto: Consulting Psychologists Press).

WYATT, S., & LANGDON, J.N., (1932), Inspection Processes in Industry. (Industrial Health Research Board Report 63, London).

ZUNZANYIKA, X.K. , & DRURY, C.G., (1975) Effects of information on industrial inspection performance. Human Reliability and Quantity Control (London: Taylor & Francis).

Appendices

Appendix A : Statistical Analysis

Experiment 1

" 3

" 4

" 5

" 6

" 7

Appendix B : Data

Appendix C : The experimental apparatus and its associated control program

Appendix D : Analysis programs

Because of space limitations only a selection of the statistical analyses referred to in the text are to be found in Appendix A. Appendix B contains the raw test scores of the cognitive tests. All other data and statistical analyses are available at the Department of Applied Psychology, The University of Aston in Birmingham. For similar reasons, Appendix D only contains the most important programs used.

APPENDIX A : EXPERIMENT 1

ANALYSIS OF VARIANCE. . . 2ASIN CORRECT DETECTION PROBABILITY

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 1.80137

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	0.20974	2	0.10487	2.3987	2,24	0.11069
P	0.01117	1	0.01117	0.2555	1,24	0.62322
NP	0.06263	2	0.03132	0.7163	2,24	0.50278
T	0.41057	2	0.20529	4.6954	2,24	0.01861*
NT	0.14627	4	0.03657	0.8364	4,24	0.51720
PT	0.07414	2	0.03707	0.8478	2,24	0.44404
NPT	0.18590	4	0.04647	1.0630	4,24	0.39701
S	1.18058	6	0.19676	4.5005	6,24	0.00371*
NS	0.24648	12	0.02054	0.4698	12,24	0.91313
PS	0.57977	6	0.09663	2.2101	6,24	0.07689
NPS	0.30784	12	0.02565	0.5868	12,24	0.83165
TS	0.45288	12	0.03774	0.8632	12,24	0.59187
NTS	0.70647	24	0.02944	0.6733	24,24	0.83041
PTS	0.35515	12	0.02960	0.6769	12,24	0.75703
NPTS	1.04929	24	0.04372	1.0000	24,24	0.50000
TOTAL	5.97888	125				

ANALYSIS OF VARIANCE. . . . A1 INDEX

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 92.92094

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	13.54423	2	6.77212	0.5235	2,24	0.60416
P	2.27998	1	2.27998	0.1762	1,24	0.68052
NP	30.01689	2	15.00844	1.1602	2,24	0.33090
T	13.06806	2	6.53403	0.5051	2,24	0.61490
NT	99.98139	4	24.99535	1.9322	4,24	0.13699
PT	2.92952	2	1.46476	0.1132	2,24	0.88801
NPT	72.09114	4	18.02279	1.3932	4,24	0.26559
S	323.81124	6	53.96854	4.1719	6,24	0.00542**
NS	137.25577	12	11.43798	0.8842	12,24	0.57357
PS	137.40765	6	22.90128	1.7703	6,24	0.14771
NPS	134.20350	12	11.18363	0.8645	12,24	0.59072
TS	326.31057	12	27.19255	2.1020	12,24	0.05846
NTS	337.39571	24	14.05815	1.0867	24,24	0.42016
PTS	107.54167	12	8.96181	0.6928	12,24	0.74328
NPTS	310.46919	24	12.93622	1.0000	24,24	0.50000
TOTAL	2048.30652	125				

ANALYSIS OF VARIANCE. . . D'

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 2.79303

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	0.53628	2	0.26814	0.8619	2,24	0.43815
P	0.02309	1	0.02309	0.0742	1,24	0.77751
NP	0.89781	2	0.44890	1.4430	2,24	0.25518
T	1.00675	2	0.50337	1.6181	2,24	0.21796
NT	1.96700	4	0.49175	1.5807	4,24	0.21096
PT	0.27749	2	0.13874	0.4460	2,24	0.65063
NPT	1.90961	4	0.47740	1.5346	4,24	0.22326
S	8.21668	6	1.36945	4.4021	6,24	0.00415**
NS	2.72701	12	0.22725	0.7305	12,24	0.71006
PS	3.60381	6	0.60064	1.9307	6,24	0.11641
NPS	2.37865	12	0.19822	0.6372	12,24	0.79081
TS	5.48413	12	0.45701	1.4691	12,24	0.20358
NTS	5.37898	24	0.22412	0.7204	24,24	0.78628
PTS	3.05035	12	0.25420	0.8171	12,24	0.63269
NPTS	7.46619	24	0.31109	1.0000	24,24	0.50000
TOTAL	44.92381	125				

ANALYSIS OF VARIANCE....LOG CORRECTED BETA

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 0.72958

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	1.84492	2	0.92246	2.4484	2,24	0.10614
P	0.47117	1	0.47117	1.2506	1,24	0.27405
NP	0.62900	2	0.31450	0.8348	2,24	0.44959
T	3.88696	2	1.94348	5.1584	2,24	0.01352*
NT	1.32725	4	0.33181	0.8807	4,24	0.49187
PT	0.49871	2	0.24936	0.6618	2,24	0.52934
NPT	0.82255	4	0.20564	0.5458	4,24	0.70646
S	1.17037	6	0.19506	0.5177	6,24	0.79019
NS	5.09648	12	0.42471	1.1273	12,24	0.38445
PS	3.29944	6	0.54991	1.4596	6,24	0.23354
NPS	5.47149	12	0.45596	1.2102	12,24	0.33115
TS	15.63257	12	1.30271	3.4577	12,24	0.00492**
NTS	16.02485	24	0.66770	1.7722	24,24	0.08409
PTS	5.01980	12	0.41832	1.1103	12,24	0.39613
NPTS	9.04219	24	0.37676	1.0000	24,24	0.50000
TOTAL	70.23776	125				

ANALYSIS OF VARIANCE. . . . ZFA (BIAS INDEX)

LEVELS OF FACTORS

N	3
P	2
T	3
S	7

GRAND MEAN 1.76842

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
N	0.04747	2	0.02373	0.2792	2,24	0.76165
P	0.13298	1	0.13298	1.5645	1,24	0.22108
NP	0.26647	2	0.13323	1.5675	2,24	0.22807
T	0.24574	2	0.12287	1.4456	2,24	0.25458
NT	0.58374	4	0.14594	1.7169	4,24	0.17844
PT	0.02311	2	0.01155	0.1359	2,24	0.86971
NPT	0.46622	4	0.11655	1.3713	4,24	0.27284
S	0.83669	6	0.13945	1.6406	6,24	0.17898
NS	1.45440	12	0.12120	1.4259	12,24	0.22113
PS	0.46697	6	0.07783	0.9157	6,24	0.50165
NPS	1.32573	12	0.11048	1.2998	12,24	0.28067
TS	2.78422	12	0.23202	2.7297	12,24	0.01743
NTS	3.38231	24	0.14093	1.6580	24,24	0.11143
PTS	1.29286	12	0.10774	1.2675	12,24	0.29801
NPTS	2.03995	24	0.08500	1.0000	24,24	0.50000
TOTAL	15.34886	125				

APPENDIX A : EXPERIMENT 3

ANALYSIS OF VARIANCE. . . D'

LEVELS OF FACTORS

T	4
R	2
S	3

GRAND MEAN 0.95118

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.69767	3	0.23256	0.9934	3, 6	0.45849
R	0.02642	1	0.02642	0.1128	1, 6	0.74237
TR	0.25841	3	0.08614	0.3679	3, 6	0.78026
S	8.43197	2	4.21599	18.0092	2, 6	0.00363**
TS	2.61940	6	0.43657	1.8649	6, 6	0.23313
RS	0.46714	2	0.23357	0.9977	2, 6	0.42457
TRS	1.40461	6	0.23410	1.0000	6, 6	0.50000
TOTAL	13.90561	23				

ANALYSIS OF VARIANCE. . . LOG BETA

LEVELS OF FACTORS

T	4
R	2
S	3

GRAND MEAN 0.61572

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.04966	3	0.01655	0.0763	3, 6	0.96596
R	0.00219	1	0.00219	0.0101	1, 6	0.88544
TR	0.95239	3	0.31746	1.4628	3, 6	0.31571
S	2.76177	2	1.38089	6.3627	2, 6	0.03306
TS	0.86914	6	0.14486	0.6675	6, 6	0.68284
RS	0.20630	2	0.10315	0.4753	2, 6	0.64663
TRS	1.30217	6	0.21703	1.0000	6, 6	0.50000
TOTAL	6.14362	23				

APPENDIX A : EXPERIMENT 4

ANALYSIS OF VARIANCE. . . . C. D. PROBABILITY

LEVELS OF FACTORS

T	5
F	2
P	2
S	3

GRAND MEAN 0.62367

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.19623	4	0.04906	1.9985	4, 8	0.18752
F	0.30817	1	0.30817	37.5508	1, 2	0.02196*
TF	0.03593	4	0.00898	1.4207	4, 8	0.31071
P	0.04374	1	0.04374	0.2790	1, 2	0.42384
TP	0.05609	4	0.01402	0.7575	4, 8	0.46190
FP	0.00417	1	0.00417	0.0328	1, 2	0.42384
TFP	0.02327	4	0.00582	0.8738	4, 8	0.46190
S	0.37529	2	0.18765	1.0000	2, 2	0.50000
TS	0.19637	8	0.02455	1.0000	8, 8	0.50000
FS	0.01641	2	0.00821	1.0000	2, 2	0.50000
TFS	0.05059	8	0.00632	1.0000	8, 8	0.50000
PS	0.31356	2	0.15678	1.0000	2, 2	0.50000
TPS	0.14811	8	0.01851	1.0000	8, 8	0.50000
FPS	0.25441	2	0.12721	1.0000	2, 2	0.50000
TFPS	0.05325	8	0.00666	1.0000	8, 8	0.50000
TOTAL	2.07559	59				

ANALYSIS OF VARIANCE. . . . F. A. PROBABILITY

LEVELS OF FACTORS

T	5
F	2
P	2
S	3

GRAND MEAN 0.14552

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.04076	4	0.01019	3.7911	4, 8	0.05149
F	0.01157	1	0.01157	0.1123	1, 2	0.42384
TF	0.02068	4	0.00517	1.5042	4, 8	0.28793
P	0.30944	1	0.30944	28.0542	1, 2	0.03033
TP	0.01193	4	0.00298	0.9859	4, 8	0.46190
FP	0.01157	1	0.01157	0.5401	1, 2	0.42384
TFP	0.00927	4	0.00232	0.6158	4, 8	0.46190
S	0.04283	2	0.02141	1.0000	2, 2	0.50000
TS	0.02150	8	0.00269	1.0000	8, 8	0.50000
FS	0.20605	2	0.10303	1.0000	2, 2	0.50000
TFS	0.02749	8	0.00344	1.0000	8, 8	0.50000
PS	0.02206	2	0.01103	1.0000	2, 2	0.50000
TPS	0.02420	8	0.00302	1.0000	8, 8	0.50000
FPS	0.04286	2	0.02143	1.0000	2, 2	0.50000
TFPS	0.03010	8	0.00376	1.0000	8, 8	0.50000
TOTAL	0.83231	59				

ANALYSIS OF VARIANCE. . . LOG BETA

LEVELS OF FACTORS

T	5
F	2
P	2
S	3

GRAND MEAN 0.69509

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	3.10849	4	0.77712	7.1302	4, 8	0.00997**
F	0.15828	1	0.15828	0.0600	1, 2	0.42384
TF	0.94028	4	0.23507	1.6606	4, 8	0.25042
P	10.40656	1	10.40656	9.0236	1, 2	0.09470
TP	1.56394	4	0.39098	2.1261	4, 8	0.16883
FP	2.06128	1	2.06128	42.9889	1, 2	0.01885*
TFP	0.13338	4	0.03334	0.2125	4, 8	0.46190
S	1.55469	2	0.77734	1.0000	2, 2	0.50000
TS	0.87192	8	0.10899	1.0000	8, 8	0.50000
FS	5.27277	2	2.63639	1.0000	2, 2	0.50000
TFS	1.13248	8	0.14156	1.0000	8, 8	0.50000
PS	2.30652	2	1.15326	1.0000	2, 2	0.50000
TPS	1.47118	8	0.18390	1.0000	8, 8	0.50000
FPS	0.09590	2	0.04795	1.0000	2, 2	0.50000
TFPS	1.25505	8	0.15688	1.0000	8, 8	0.50000
TOTAL	32.33271	59				

ANALYSIS OF VARIANCE. . . SCORE

LEVELS OF FACTORS

T	5
F	2
P	2
S	3

GRAND MEAN 174.98333

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	6213.06665	4	1553.26666	2.2969	4, 8	0.14720
F	13771.35010	1	13771.35010	0.7560	1, 2	0.42384
TF	3266.40002	4	816.60001	0.9977	4, 8	0.46190
P	*****.*****	1	*****.*****	156.2018	1, 2	0.00445**
TP	4696.40002	4	1174.10001	1.7897	4, 8	0.22378
FP	1118.01666	1	1118.01666	0.1868	1, 2	0.42384
TFP	293.06667	4	73.26667	0.0715	4, 8	0.46190
S	21105.73340	2	10552.86670	1.0000	2, 2	0.50000
TS	5409.93335	8	676.24167	1.0000	8, 8	0.50000
FS	36433.20019	2	18216.60010	1.0000	2, 2	0.50000
TFS	6547.80005	8	818.47501	1.0000	8, 8	0.50000
PS	2172.93335	2	1086.46667	1.0000	2, 2	0.50000
TPS	5248.40002	8	656.05000	1.0000	8, 8	0.50000
FPS	11970.13330	2	5985.06665	1.0000	2, 2	0.50000
TFPS	8192.53345	8	1024.06668	1.0000	8, 8	0.50000
TOTAL	296146.96873	59				

APPENDIX A : EXPERIMENT 5

ANALYSIS OF VARIANCE. . . . ASIN C. D. PROBABILITY

LEVELS OF FACTORS

T	5
C	3
S	3

GRAND MEAN 1.40293

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.28675	4	0.07169	5.0032	4, 8	0.02585*
C	0.02973	2	0.01486	0.0709	2, 4	0.44610
TC	0.21039	8	0.02630	1.8576	8, 16	0.13881
S	0.91364	2	0.45682	1.0000	2, 2	0.50000
TS	0.11463	8	0.01433	1.0000	8, 8	0.50000
CS	0.83896	4	0.20974	1.0000	4, 4	0.50000
TCS	0.22653	16	0.01416	1.0000	16, 16	0.50000
TOTAL	2.62062	44				

ANALYSIS OF VARIANCE....ASIN F.A. PROBABILITY

LEVELS OF FACTORS

T	5
C	3
S	3

GRAND MEAN 0.70090

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	0.62792	4	0.15698	7.0379	4, 8	0.01034*
C	3.26910	2	1.63455	15.5388	2, 4	0.01487*
TC	0.73618	8	0.09202	7.5128	8,16	0.00055***
S	0.34964	2	0.17482	1.0000	2, 2	0.50000
TS	0.17844	8	0.02231	1.0000	8, 8	0.50000
CS	0.42077	4	0.10519	1.0000	4, 4	0.50000
TCS	0.19598	16	0.01225	1.0000	16,16	0.50000
TOTAL	5.77802	44				

ANALYSIS OF VARIANCE....LOG BETA

LEVELS OF FACTORS

T	5
C	3
S	3

GRAND MEAN 0.80606

SOURCE OF VARIATION	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARES	F-RATIO	DF	P
T	6.25840	4	1.56460	12.2851	4, 8	0.00222**
C	18.14703	2	9.07351	86.1783	2, 4	0.00157**
TC	6.71521	8	0.83940	9.9151	8,16	0.00017**
S	3.01589	2	1.50795	1.0000	2, 2	0.50000
TS	1.01886	8	0.12736	1.0000	8, 8	0.50000
CS	0.42115	4	0.10529	1.0000	4, 4	0.50000
TCS	1.35454	16	0.08466	1.0000	16,16	0.50000
TOTAL	36.93109	44				