1996

# OUTCOME BIAS IN JUDGEMENTS OF THE QUALITY OF EXPERIMENTAL DESIGNS

BRADON, PETER

# OUTCOME BIAS IN JUDGEMENTS OF THE QUALITY OF EXPERIMENTAL DESIGNS

by

**PETER BRADON**

A thesis submitted to the University of Plymouth
in partial fulfilment for the degree of

**Doctor of Philosophy**

Department of Psychology
Faculty of Human Sciences



August 1996

# OUTCOME BIAS IN JUDGEMENTS OF THE QUALITY OF EXPERIMENTAL DESIGNS

by PETER BRADON

## ABSTRACT

This thesis is concerned with the evaluation of experimental design. It reviews design research in general and notes some methodological limitations. The process of experimental design is considered in the light of this review, particularly in terms of the ability of designers to evaluate potential designs. Experiment 1, a prospective study involving the creative design of psychology experiments, reports inadequate assessment of experimental power and little or no evidence of explicit evaluation processes. Experiment 2 assesses the evaluation of existing experimental designs and demonstrates that judgements of the quality of experimental design are influenced by the presence of outcome information more than any other factor. Following this result a review of the hindsight and outcome bias literature is presented.

Experiments 3,4 and 5 demonstrate that outcome bias is a pervasive effect not mediated by task presentation, explicit definitions of quality or statistical expertise. Also, by manipulating subjects' perspective, Experiment 5 differentiates between the effects of outcomes themselves and their financial implications. The term outcome salience is defined as "the relative importance of the implications of an outcome from the point of view of the subject". It is shown that both the size and direction of an outcome bias are determined by (and could be predicted from) the associated outcome salience. Experiment 6 explores subjects' beliefs about the relevance of experimental factors, outcome information and the financial implications of outcomes to quality judgements. Ratings of relevance are shown to be in direct opposition to the actual use of these factors in judgements. Relevance ratings are also shown to be influenced by a subject's perspective.

Experiment 7 tests the links between outcome bias effects and traditional, memory based, hindsight bias effects using a memory based paradigm. Results show that, in addition to biases of judgement in foresight, the same outcome information will also bias the memory of earlier judgements and the memory of relevant task details in hindsight.

The practical implications of outcome bias are discussed. Using a motivational account based on the concept of outcome salience, hindsight bias is redefined as one particular form of outcome bias. This account unifies two previously separate research areas and is shown to explain a number of previously unexplained effects in the hindsight literature. Accounts of reasoning are reviewed, as are information processing and motivational accounts of hindsight bias. The theoretical implications of the present results are discussed in the light of these accounts.

## LIST OF CONTENTS

## LIST OF ILLUSTRATIONS AND TABLES

## ACKNOWLEDGEMENT

The author gratefully acknowledges the assistance and advice of
Prof. Jonathan Evans and Dr. Ian Dennis
and the support of the staff of the Psychology department at the University of Plymouth.

## AUTHORS DECLARATION

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

This study was financed with the aid of a studentship from the Science and Engineering Research Council.

Signed....P. Bradon.....

Date....6:12:96.....

## Table of contents - Chapter one

# CHAPTER ONE

# A REVIEW OF LITERATURE RELATED TO DECISION MAKING IN DESIGN TASKS

## 1.1 THESIS INTRODUCTION

This research utilised the design of experiments as a domain in which to explore the ways in which subjects used information in qualitative judgements. The study of the evaluation of experimental designs presented a number of advantages over research in other domains. In this domain normative analysis allows for the possibility of at least determining that in some cases experimental designs are unequivocally bad; although it is often more difficult to define those designs which are good. As it is possible to determine the factors which are relevant to the statistical reliability of experimental designs and those which are not, the study of experimental design allows for a more quantitative evaluation than that available in other design domains.

The thesis will first explore the evaluation of putative experimental designs as a necessary part of the overall process of experimental design. This will be followed by a number of experiments which will explore judgements of the quality of existing experimental designs. The inclusion of outcome information in these descriptions of experimental design will generate a paradigm in which biases related to the influences of irrelevant hindsight information can be clearly differentiated from the inclusion of relevant factors in judgements of quality. Because of this clear differentiation manipulations of this paradigm can be expected to lead to clearer and more consistent results than previous hindsight bias or outcome bias paradigms. From these results it should be possible to generate a clearer theoretical description of the underlying nature of these biases.

It is the intention of this thesis to show that judgements relating to the quality of experimental designs are inevitably biased by the nature of their outcomes where those outcomes are known. The findings of these studies will have important practical implications in such areas as peer reviewing of putative journal articles and the examining of

Ph.D. theses. Information about the outcomes of experimental procedures can be seen as a form of hindsight information as they are not available to the designer of an experiment at the time of design. Thus the studies presented in this thesis will be strongly linked to previous studies of hindsight bias and will attempt to differentiate between varying theoretical explanations of hindsight and related biases.

## 1.2 CHAPTER ONE INTRODUCTION

Chapter 1 is concerned with a review of studies related to the processes involved in experimental design. Within psychological research a number of persistent design problems have been reported; notably failures to correctly assess the power of experimental designs and problems with the interpretation of experimental results (Cohen, 1990, Sedlmeier and Gigerenzer, 1989). It is the intention of this thesis to determine to what extent bias in the evaluation of experimental design may be responsible for these shortcomings. There is at present little or no research into the cognitive processes involved specifically in the design of psychology experiments, however, there is a great deal of research available from other areas of design and from goal oriented problem solving within psychology. It is the intention of this chapter to determine some of the processes involved in designing a complete experiment by using relevant studies from these fields to gain an insight into design in general. In an attempt to overcome the problems inherent in making valid comparisons across areas as diverse as those in design related research and problem solving a system of categorisation of the different stages of design tasks is presented. This may help to clarify a number of contradictory results in this extensive and varied research area. This categorisation uses a conceptual division of the design process into a design stage and an evaluation stage. Some methodological problems inherent in the study of design are explored with reference to these proposed stages.

The second part of this chapter focuses on experimental design in general and the design of psychology experiments in particular. Experimental design is compared to other design tasks using the proposed categorisation scheme in an attempt to determine common

underlying psychological processes. Factors relating to evaluation of putative designs are explored in the light of these processes.

## 1.3 STUDIES OF DESIGN PROCESSES

### 1.3.1 History of design studies

In recent years there has been an increasing interest in the cognitive processes involved in design. In most cases this interest has stemmed from attempts to create intelligent systems to aid design processes. Subsequently a number of computer aided design (CAD) systems have been developed in such areas as architecture, mechanical engineering, electrical engineering and computer programming. At the present time there are, "hundreds of functioning expert systems and thousands more under development" (Shanteau and Stewart, 1992, p.101) yet despite this apparent success there has been no great breakthrough in understanding the cognitive processes used by experts in design.

Ullman (1991), in a review of design theory research; reports that despite their recent proliferation CAD systems are "best suited for drafting and *..are*...little help in design" (page 206) and thus have not lived up to their claims of being aids to the design process. Ullman points to the failure of this whole area of research to produce any consistent design theory and concludes that design research is still in the pre-theory stage.

This failure to find a consistent design theory must, in part, be due to the wide range of areas in which design is studied and the lack of connections between these areas. In an international survey of research and studies on design Tempczyk (1986) listed 191 projects completed or in progress. A brief analysis of this list reveals the very diverse areas under which design has been studied. There are 27 different departmental specialities represented. The majority of studies are in engineering followed by a large number in architecture. At the other end of the scale are departments such as philosophy and business administration. Notably only one study out of the whole 191 was within a psychology department. This seems a very strange state of affairs; design is clearly a complex psychological process and

yet it is studied with almost no reference to psychology itself.

Within psychology judgement and decision making research pre-dates expert systems research and has studied many allied areas. Specifically studies of cognitive processes in problem solving tasks which start with an initial problem state and attempt to achieve a final goal state seem to be particularly analogous to design tasks. However, there has been almost no consideration of this research in the design literature. As Chan (1990) points out, this may be because judgement and decision making research has focused on the deficiencies of experts whilst expert system research has focused on what experts do best. However this difference in focus alone should not be sufficient to prevent comparison between these allied areas.

Another clue to why these fields may have remained separate comes from the surprising lack of agreement within any single field of research. For example Stauffer and Ullman (1988) in a comparison of the results from six empirical studies of mechanical design found twenty seven different general conclusions across the six studies. None of these conclusions were agreed on by more than half the studies. The authors also found very few similarities in design task and methodology between the studies. The majority of the article was taken up with discussing these differences between studies and this led to the authors' final conclusion; that a great deal more research was needed before areas of agreement could be found. Similarly, Hubka and Eder (1987) in a review of engineering design conclude that "A broad analysis..(of design studies)..is sorely needed" (page 123).

Since this time a great deal more research has been completed and yet a comparison of this research reaching any firm conclusions has still not been made. As Ullman (1991) points out, "design research.....is still discipline and viewpoint fragmented" (page 207).

In trying to reach general conclusions from design studies the researcher is presented with a bewildering array of different problem solving strategies. This may be a result of designers using a huge number of different strategies. However it would be a mistake to reach this conclusion without considering the wide range of fundamental differences in the type of problem being studied and the methodological difficulties inherent in this field of research.

### 1.3.2 Categorisation of Design Tasks

The intention of this literature review is to assess the results from previous studies which relate to decision making in the design of psychology experiments. There have been very few studies of experimental design and those that are available generally refer to specific aspects of statistical decision making (Evans and Bradshaw 1986, Reagan 1989, Gill 1987, etc.). These experiments present subjects with an almost completed design scenario and require them to complete the missing attribute; for example an estimate of statistical power or necessary sample size.

Whilst this work is of great interest and will be considered later the focus on a specific aspect of design does not allow for an overview of the processes involved in designing a complete experiment. Therefore, initially it is necessary to analyse the small number of relevant studies from other fields in order to gain an insight into the overall design process selecting first those studies which focus directly on generative design processes where the design process is seen as a process of problem solving directed at the goal state of a completed design. As has been shown, due to the complexity and variability of different design tasks these studies often have little in common making general comparison difficult. Therefore this review will start by attempting to categorise the important factors within a design task and then go on to explore previous studies in the light of these factors.

A clear categorisation of design tasks is necessary to prevent the erroneous comparison of studies which may vary hugely in terms of design attributes, evaluation methods and associated semantic knowledge. Analysis of studies within this framework should enable more valid conclusions to be drawn from comparisons of research from very different design fields. This method should also allow for the inclusion of relevant psychological research not strictly in the domain of design. For example, this strategy should enable the inclusion of findings from judgement and decision making research in a framework covering the whole range of human problem solving related to design.

In order to achieve this synthesis it is necessary to have a clear understanding of differences in design tasks. These include various levels of multi-attribute problem, complexity of solution evaluation and variations in associated semantic knowledge. Despite

the possibilities for comparison of diverse research fields created by the use of a categorisation system it is not the intention of this review to cover the whole gamut of research on design. Even if this were possible the task would be immense thus criteria have been applied limiting the range of research included. These criteria are firstly relevance to the understanding of underlying cognitive processes in design, and secondly relevance to processes likely to be common to the design of psychology experiments.

For the purpose of this analysis of design tasks the process of design has been broken down into a design stage and an evaluation stage. This analytical method is not intended to suggest that these are discrete stages; in the real life design process there may be a great deal of overlap between the two stages.

The primary factors in a design task are the number of attributes to be considered in the design stage and the relationship between these factors. Thus at one end of the scale subjects may be presented with a task where they only have to determine the value of a single attribute; as in a simple mathematical task. Conversely a task might require the combination of values for large numbers of inter-related attributes as in large scale architectural design.

Secondly, studies vary in the level of semantic knowledge associated with successfully determining appropriate values for these attributes (and also in evaluating the final solution). This can also be seen as the level of expertise associated with the task. Semantic knowledge is, strictly speaking, a quality of the designer and not of the design task. In numerous studies relating to expertise the same task is given to both expert and novice subject groups; differences between the groups being related to the level of semantic knowledge brought to the task by the subject.

However a particular level of semantic knowledge can also be seen as a requirement of the task used in a design study. In this respect there is often a fundamental difference between the goal oriented problem solving tasks studied in psychology and the studies of realistic design tasks. For example, Kahney (1993) defines four different sorts of information present in well-defined laboratory problems such as the 'tower of Hanoi' or the 'missionaries and cannibals' problems. These four sorts of information completely define the problem in terms of; the initial state, the goal state, legal operators (things you are allowed to do in solving the problem) and operator restrictions (factors which constrain legal

operators). Given this information in the problem instructions no semantic information whatsoever is required for the solution. At the other end of the scale the chemical engineering thermodynamics design problems used by Bhaskar and Simon (1977), require a high level of semantic information in terms of detailed expert knowledge and these tasks are only relevant to a very specific subject group. Other realistic design related tasks such as the planning of meals used by Byrne (1977) may seem at first to be semantically poor as the low level of general knowledge they require makes them relevant to a large range of non-expert subjects. However, the high levels of semantic information required should not be disregarded simply because they are shared by a large proportion of the population.

Thirdly, the evaluation stage of a design study can vary greatly in complexity. This may not only affect the final checking of a solution but also may affect the design process in complex designs (where design strategies may be based on the results of a series of evaluation stages). For example, in some tasks the solution may be a single number which can be simply checked for accuracy. Alternatively, the solution may be a description of a complex engineering system which will require detailed analysis to estimate its efficiency; in this case partial evaluation may be an integral process determining subsequent design strategy.

Finally, as well as differences in the complexity of design and evaluation stages there are also differences in the levels of fuzziness in both the original problem and in the evaluation. The concept of fuzziness relates to uncertainties brought about by vagueness of definition and can be seen as closely related to the need for expertise (i.e. semantic richness) in a problem, in that an expert can often fill in gaps in a fuzzy definition by referring to previous experience. However this need to search for additional information complicates the design process and can create uncertainties beyond the scope of the designers' semantic knowledge. Thus fuzziness is an important category in its own right and cannot simply be subsumed under the semantically rich/poor category.

Thus this categorisation of design studies involves evaluation on each of five relatively independent scales. Table 1.1 lists these five categories which are also separated in terms of the design stage in which they have the most influence. In this table each of the categories represents a design task variable presented with the simplest extreme value on the left and the

most complex extreme value on the right.

Table 1.1

Categorisation scheme for design tasks

| | | |
|---|---|---|
| Design stage | 1. **Single attribute** ..... | **Multi-attribute** |
| | 2. **Defined problem** ..... | **Fuzzy problem** |
| General | 3. **Semantically poor** ..... | **Semantically rich** |
| Evaluation stage | 4. **Simple evaluation** ..... | **Complex evaluation** |
| | 5. **Clear evaluation** ..... | **Fuzzy evaluation** |

Using the above template studies could be classified by assigning a value on a scale for each of the five task variables. Noticeably, by considering general classes of design research study, patterns of categorisation readily emerge. For example, studies falling at the right hand end of all categories would generally be real life design studies; such as architectural or engineering design projects. This class of study typically necessitates analysis by verbal protocol. Studies utilising tasks which fall at the left hand end of these categories are typically artificial or logical tasks; such as the goal oriented problem solving tasks mentioned earlier. These studies are usually designed for formal experimental analysis. Thus, particular groups of studies can be identified by this classification method.

It should be noted that there is a further category not included here. Design is often studied as a group process and the number of designers involved is another important variable in categorising design studies. In very complex real life design tasks it is usual for a team of designers to be responsible for the final design. For example, Stauffer and Ullman (1988) quote a study of engineering design involving thirty seven individuals over a three year period. As group decision making is beyond the range of this review only studies

relating to individual design processes or those in which individual methods of design can be determined will be included. As Stauffer and Ullman point out "Organisations and the individuals in those organisations are a part of the behavioural aspect of design" (page 113). They also note that designers working alone will often investigate solutions one at a time (serially) whereas in group design solutions are generally investigated in parallel. Thus group design can fundamentally change the design strategy for a given problem. It should be remembered that this will limit the generalisation of results from studies of individuals.

## 1.4 RELEVANT FINDINGS FROM PROBLEM SOLVING RESEARCH

By applying the above categorisation scheme to problem solving research it becomes immediately apparent that great care should be taken in generalising findings from this area to the area of realistic design tasks. Laboratory problems such as the 'towers of Hanoi' fall at the left hand end of all the categories (see Table 1.1). Firstly, they have no true design stage, instead there a simple statement of the initial and goal states. The problem is rigidly defined and usually involves the manipulation of a single attribute for which clear rules are set. This is in sharp contrast to the complex determining of multiple inter-related attributes fuzzily defined in real life design tasks.

Secondly, as the problem state and rules are rigidly defined there is no requirement for semantic knowledge. Nevertheless, skill related transfer effects can be seen with homomorphic (Reed et al. 1974) and isomorphic problems (Luger and Bauer, 1978) these effects depend on similarities of underlying rule structure. Transfer can also be negative; where related real-world knowledge is in opposition to the defined problem rules this serves to make the problem more difficult (Kahney, 1993). This situation is analogous to the biasing and de-biasing effects of knowledge known as "belief bias' in deductive reasoning (See Evans, 1989 for a complete review).

Finally, the evaluation stage in problem solving tasks is so simple and clearly defined as to be practically non-existent. It is only necessary to determine if the state you have reached matches the desired goal state defined in the initial problem.

Despite these clear differences some very general conclusions can be drawn from this research which should be relevant to experimental design tasks. Although design and evaluation stages are qualitatively and quantitatively different to those in experimental design, there are similarities of problem structure. These involve the moving from an initial state to a goal state often through a series of sub-goals. A similar sub-goal strategy is often reported in the design literature. Notably the positive transfer effects reported by Reed et al. (1974) and Luger and Bauer (1978) are specifically facilitated where problems have a similar sub-goal structure and if subjects are aware that the problems are similar. Thus in experimental design we could expect the formation of a positive expertise where a designer has experience of a particular format of design problem within a particular area where sub-goal strategies are common. This 'expertise' would however become a negative influence when attempting any new design which was not amenable to the same sub-goal strategy. In addition the effects of analogical transfer reported by Gick and Holyoak (1980) may also be of some relevance. It seems reasonable to assume that where a valid analogy could be made with the skills required for successful experimental design this analogy would facilitate subsequent design.

There is also no reason to suppose that the findings related to the limitations of working memory capacity in problem solving would not have similar influences in the field of experimental design. The work of Simon (1978) on the limits of working memory capacity clearly predicts performance limits in complex problems, although in practice these may be to some extent alleviated by taking and referring to notes. As all experimental design tasks can be expected to go beyond the limits of working memory, the work of Kotovsky et al. (1985) on the interactions between working memory and long-term memory may be of more practical importance. This research shows that in complex problems expertise in terms of learning and practising the application of rules is crucial in developing the ability to plan sequences of moves. In terms of experimental design this research strengthens the earlier predictions of the importance of specific task related expertise.

## 1.5 RELEVANT FINDINGS FROM DESIGN RESEARCH

It was possible with the experimental problem solving tasks outlined above to make a general and all inclusive categorisation. This sort of generalisation is impossible with the tasks used in design research. Design research has focused on 'real world' problems in a number of quite specific and different areas; e.g. engineering, architecture, software design, etc. These areas produce radically differing tasks and even within one specific area the range of tasks can be immense. Thus in order to make any valid conclusions which may be relevant to experimental design tasks research will be analysed in terms of its problem structure based on the categorisation scheme proposed in section 1.3.2. Comparisons will be drawn within these categories where a number of studies have common values in that category. As a result of these limitations comparisons are necessarily of a very general nature.

### 1.5.1 Methodological problems in design research

In addition to the problems of comparison across different design tasks outlined above there are also some methodological issues which must be addressed. Studies of the process of design have typically relied on 'think aloud' verbal protocols recorded whilst subjects complete a design task. This method has the advantage of very good external validity, however, the subsequent analysis of these protocols is constrained by the level of self-knowledge that can be attributed to subjects. Whilst most researchers appear to be theoretically aware of these constraints; in practice the conclusions drawn from protocol analysis often overstep the bounds of this methodology. Therefore it is necessary to be acutely aware of the limitations of this methodology before assessing the results of this research.

Think aloud verbal protocols have become an important tool in tracing cognitive processes. The increased use of verbal protocols has led to increasing debate on their validity. Ericsson and Simon (1984) prescribe a theoretical basis for the use of verbal protocols as data and this has become the standard reference for all work in this field.

Within this theoretical framework is the assumption that given the correct instructions and using the right kind of task subjects will reliably report the contents of their short term memory (STM). Protocols must be recorded concurrently with instructions to avoid theorising and interpretation of thoughts. In terms of task selection they suggest that only tasks which lead to the contents of the subjects' STM being coded in verbal form should be used. Tasks that involve re-coding from a non-verbal representation in STM or automated tasks which leave little trace in STM will lead to invalid protocols.

This raises two questions. How do we know that a particular task will lead to verbal coding in STM? Also, if a task generally does lead to verbal coding will this be the case for all subjects? The answers to these questions are not clear; whilst it is easy to avoid tasks which are clearly pictorial and are likely to lead to visual representations in STM this does not guarantee that the remaining tasks will necessarily be coded verbally.

In addition to this problem is the question of self-knowledge. The self-knowledge debate is concerned with how much awareness people have of their own thought processes. This question is of crucial importance to the study of real life design processes. Within this review the details of self-knowledge research will not be presented (for a review of the area see Evans, 1983, Chapter 5). Nevertheless, despite different theoretical standpoints and continuing debate there are few commentators who would not agree that people have a limited knowledge of their own cognitive processes.

Within the self-knowledge debate it is postulated that there are two forms of knowledge; explicit and implicit. Explicit (declarative) knowledge is consciously accessible and thus can be verbalised. For example knowledge which has been formally learned and committed to memory such as the names of the kings and queens of England or factual knowledge such as what you ate for tea yesterday. Implicit (procedural) knowledge is not consciously accessible in a verbal form. This includes such knowledge as how to ride a bicycle or how to play a musical instrument. It should be noted that implicit knowledge cannot be learned by rote but must involve practice and experience. For example in learning to play a musical instrument there are associated explicit rules such as which key produces which note and which notes are in which musical key, nevertheless the actual playing must be learned through practical experience.

Expertise in any particular area can be seen to be based to a large extent on detailed implicit knowledge. Firstly expertise is based on extensive practical experience and as we have seen this is clearly related to the learning of implicit knowledge. Secondly it can be argued that even where the learning of a task involves explicit rules (such as how to drive a car) the achievement of expertise necessitates such a familiarity with these rules that they become internalised. At this stage the originally explicit consciously remembered rules seem to have been assimilated into an implicit procedure. For example it is a common experience that when learning to drive at first it is necessary to consciously remember to put the clutch in before changing gear. As one becomes more expert this rule becomes "automatic" i.e. subconscious. Research into bias in human reasoning suggests that biases should not be seen as errors of reasoning but rather as the direct result of the use of goal directed implicit cognitive processes in areas where explicit rule based processes would be more appropriate. Evans and Over (1996) define these processes as Type 1 and Type 2 and argue that both lead to different forms of rationality in a two-factor theory of reasoning. This approach will be discussed in detail in Chapter 6.

This situation of limited self-knowledge in the cognitive processes related to design leads to a number of practical implications for knowledge elicitation from designers. Early studies of design and studies related to the creation of expert systems often gave experts a task and simply asked them directly to report the strategies used to solve the task. This method took no account of the limits of self-knowledge and demonstrated the allied problem that even if a person has no awareness of their cognitive processes they will still produce a reasonable (but often completely false) account of the strategies they have used. These accounts will be based on post-hoc rationalisations rather than on any actual strategy used.

This situation of subjects readily generating false strategy reports was clearly demonstrated in several experiments reported by Evans (1983). Wason and Evans' (1975) study of matching bias in the Wason selection task shows that whilst subjects' reports of what cards they were attending to during the task were reliable reports of <u>why</u> they were attending to these cards were based on a rationalisation deduced from the consequences of this choice. This was further supported by an experiment in which subjects were shown to readily justify any of several solutions to the Wason selection task presented as 'correct' by

the experimenters, (Evans and Wason 1976). Nisbett and Wilson (1977) in a review of verbal reports conclude "little or no direct introspective access to higher order cognitive processes" to the extent that "subjects are sometimes (a) unaware of the existence of a stimulus that importantly influenced a response, (b) unaware of the existence of the response, and (c) unaware that the stimulus has affected the response" (Page 231). This is not to say that verbal reports are of no use as data (indeed none of the above authors suggest this). However, it does appear to be necessary constantly to remind researchers of the danger of believing verbal reports to be a description of underlying cognitive processes. These limitations are reiterated by Ericsson and Simon (1980) who present a general model of verbalisation containing variables related to; recognition, long-term memory, short term-memory, control of attention, fixation and automation. All these factors contain unspecified variables which have consequences for the verbalisation process. Despite their general support for the use of verbal reports as data Ericsson and Simon limit their definitive statements to "the information verbalised will ...be *some portion of* the information currently attended to" (page 225, italics mine). Thus there is no guarantee either of completeness or of any particular links to underlying cognitive processes.

The general failure of most early expert systems (Ullman 1991) and the failure of design research to reach any firm conclusions on the cognitive processes of design may well be due to the failure by many design researchers to take into account the limitations of this method of knowledge elicitation.

At the other extreme the use of strictly experimental methods necessitates the use of simple and easily repeated tasks. Unfortunately these tasks do little to test the abilities of experts; the nature of expertise being precisely in dealing with the complexity and confusion created by real life tasks. Thus in attempting to create tasks which are as close to real life as possible studies of design have avoided the constraints of experimental control giving subjects freedom in their approach to design. This realism is necessary as designing is a complex multi-attribute decision making process which cannot be fully understood when separated into a series of simpler tasks. For example evidence of parallel processing in complex design (Stauffer and Ullman, 1988) would be absent in a series of simpler problems. Unfortunately this need for realism precludes the use of experimental method and

limited self knowledge precludes the use of direct questioning.

### 1.5.2 Interpretation of findings from the literature

The inherent methodological limitations outlined above necessitate great caution in the interpretation of findings from the literature. The information in verbal protocols tells us only where a designer's attention was focused at a particular time in the design process. Given these limitations we can expect only very general conclusions based on commonalities in verbal report data. Therefore we can expect to discover common stages in design and perhaps a typical order in which these stages occur.

Looking at design from a theoretical point of view it is possible to infer some general processes which must be present in any successful design. Starting with the simplest possible theoretical model of a design process there are three factors which are necessary for success. First, it is necessary to understand the nature of the design problem. Thus we would expect design to begin with an evaluation of the task itself, leading to a clear definition of the final requirements of a completed design. Second, it is necessary to create a putative design. It is this stage that has been the focus of most research and yet to a great extent has remained a mystery. Third, it is necessary to evaluate this putative design solution; at least in terms of the problem definition achieved in the first stage. Further evaluation of factors such as efficiency and cost of design are preferable but not entirely necessary in this stage.

Initially then we are expecting evidence of design which can be grouped into these stages. It should also be possible to discover from the research some of the common factors to which a designer pays attention in these stages. We can also expect that conclusions about underlying cognitive processes drawn from verbal protocol data will be at best misleading and contradictory.

## 1.6 THE DESIGN STAGE

### 1.6.1 General findings from the literature

In some design areas it is possible to define the important variables in the initial design stage in terms of the number of attributes to be considered and the relationship between these attributes. For example, in the design of psychology experiments there is a need to balance a number of attributes to maximise the efficiency of the design. The choices determining the final design are, between or within subjects or mixed design, the subject numbers, the subject population, the levels of independent and dependent variables and their measurement. These attributes are inter-related, each affecting the evaluation factors of internal and external validity, ethics, power and cost. The effect of altering one design attribute can influence the values of the other design attributes, as well as all the evaluation attributes. Often these changes occur in a complex non-linear manner.

Many architectural and engineering studies have areas of similarity in the design stage. A number of related design attributes must be determined in order to maximise subsequent design quality (subsequent quality being based on a number of evaluation attributes). Here quality related factors such as strength, speed of operation, etc. must be balanced against cost.

The following conclusions were based on a number of studies covering over 100 subjects and 30 different design tasks. These were: Adelman and Bresnick (1992), Bhaskar and Simon (1977), Byrne (1977), Chan (1990), Eckersley (1988), Eekels and Roozenburg (1991), Evans and Bradshaw (1986), Gill (1987), Hubka and Eder (1987), Reagan (1989), Roozenburg and Cross (1991), Stauffer and Ullman (1988), Stone and Schkade (1991), Tempczyk (1986), Ullman (1991) and Ward (1989). The initial overwhelming impression from this review is that findings from these studies show more disagreement than agreement, even within studies using the same design problem and subjects with similar training backgrounds. For example Eckersley (1988), in a study of interior designers, concludes that there are "remarkable differences in the design problem solving behaviour of five individuals" (page 93). Because of these numerous differences no specific general model of

design processes can be determined from the literature and there is continuing disagreement over models which have been proposed (Roosenburg and Cross, 1991). This was not unexpected for the reasons outlined above and as predicted agreement between studies is limited to a few common strategies of a very general nature in the design stage.

At the beginning of the design process, many studies report task clarification to determine clear objectives of the design. This phase is particularly evident in engineering design and is defined as the first stage in the consensus model of engineering design, (Roosenburg and Cross, 1991). However engineering problems are generally extremely clearly defined, as their final solution has to meet very specific criteria, thus in this case initial clarification is clearly important. In designs relying on more aesthetic evaluations such as architecture or interior design (and thus having more fuzzy task instructions) this stage is less often reported and may be completely absent (Eckersley, 1988).

Also relatively early in the design process the formation of a hierarchy of design is often reported. This can involve final goal setting; the formation of a goal plan involving a series of sub-goals which focus on the critical areas, or a sequential ordering of these sub-goals. This practice is also more apparent in clearly defined tasks such as engineering design. This may be due to the inherent nature of mechanical or electronic systems which are traditionally seen as a series of inter-connected functional sub-systems rather than holistically, and certainly engineers have been trained to understand complex systems in this way. Eekels and Roozenburg (1991) argue that it is only in clearly defined areas such as engineering, that this hierarchical structure and essentially linear process of design stages can be determined. They argue for a more cyclic structure with less clearly defined stages in areas of design where problem definition and evaluation are fuzzy.

Thus both the initial task clarification and the subsequent task structuring are seen to be dependent on the levels of fuzziness of the problem. Clearly the complexity of the problem will also influence these potential first stages in terms of the number of attributes to be considered and the relationships between these attributes. With increasing problem complexity the limits of working memory will necessitate the breaking down of a problem into stages and the structuring of these stages into a goal plan for a successful solution to be achieved.

Also both parallel and serial approaches to problem solving have been reported in the literature. Stauffer and Ullman (1988) in a review of design studies argue that the use of parallel or serial development of solutions is largely dependant on the complexity of the original task; "where a problem is difficult or there is a lack of manpower, solutions are examined in series" (page 111).

Variations in the initial methods of approaching a design task may not only depend on the differing varieties of task as seen above, there is also evidence demonstrating the importance of the specific presentation of a task. In the design stage we can expect multi-attribute decision making to be largely influenced by task characteristics (Stone and Schkade, 1991), and even the sequence in which information is presented (Adelman and Bresnick, 1992).

One area of agreement across different design studies is the stage of creating putative design proposals or "conceptual design". In engineering studies this stage follows task clarification and task structuring. Often in architectural studies it is seen as the first stage of design and leads to subsequent refining of goal plans (Ward, 1989). Nowhere in the literature is there any attempt to further define this stage beyond naming it and reporting that it is a creative process. This inability of the research to gain insight into the creative processes of design is clearly predicted by the limitations of the methodology described above. Finally there is also general agreement that once a putative design has been created the next stage is an evaluation of the quality of this design.

### 1.6.2 Problems in studying the design stage

The results of previous research relating to the design stage of various design tasks are clearly confused and often contradictory. When the design task is broken down into design and evaluation stages (as has been done here) rather than looked at as a whole these results are not surprising. It becomes clear that these two stages may to a large extent involve different types of cognitive process.

The design stage in its purest sense involves at most the creation of new and original ideas or at least the balancing of the levels of a complex set of variables. Berry and Broadbent (1988) report that it is exactly this sort of complex and possibly creative task

which is subject to implicit learning and thus is not associated with verbalisable knowledge. It may be argued that this design stage relies on the use of heuristics rather than being truly creative. If this is the case then it is the choice of a particular heuristic which is the determining factor of successful design. Evans (1989) presents evidence that in reasoning tasks the choice of a heuristic is pre-attentional and that this leads directly to biases and errors. If this is also the case in design tasks then this choice will be affected by the form of the original design specification. This would lead to the same bias and errors in judgement which have been shown with varying task presentations in decision making studies and the same lack of verbalisable knowledge and false strategy reports shown by Evans (1989) (See Section 1.3.3). It seems likely therefore that a large number of previous studies have reached no firm replicable conclusions on the nature of design because they have employed a methodology almost totally ineffective in charting implicit reasoning.

In fact when we look at the few consistent results from the design stage we find such things as task clarification and in some cases the formation of sub-goals. However task clarification is actually a form of evaluation (where the description of the task itself is evaluated) and has only been included in descriptions of the initial design stages because it inevitably happens early in the total design process. It can be seen as a logical and formal process of understanding and is clearly not a creative process involved directly in the formation of a conceptual design even though the resulting evaluation may affect the nature of the subsequent design process and may play a large role in defining subsequent sub-goal strategy.

## 1.7 THE EVALUATION STAGE

### 1.7.1 The importance of evaluation

Given that some basic aptitude for design is present, it could be argued that contrary to the prevailing view the evaluation stages (task clarification and evaluation of quality of the putative solution) are the most important stages in design. If no errors are made in either of these evaluation stages any putative solution not meeting the requirements of the task will be rejected. If the task requirements are not met the solution must then be changed and re-evaluated. By an iterative process of change and re-evaluation finally a correct solution will be found or the task abandoned. Looked at in this way it becomes apparent that an inefficient design stage will only serve to make the whole design process longer. However a failure to correctly clarify the task or an inefficient evaluation stage will result in errors in design solutions. That is not to say that good creative designing is not important. A good designer will arrive at a quick and efficient solution. However, accurate evaluation is crucial if design errors are to be avoided.

Apart from the obvious importance of studying design evaluation there is the added advantage that evaluation processes are much more likely to be amenable to investigation using verbal protocol methodology. Unlike the creative (or heuristic) design stage which is likely to be largely a result of implicit processing; the evaluation stages involve logical processes of checking defined factors, these are more likely to be explicit processes. It is also possible to study evaluation of designs experimentally by asking subjects to judge the quality of designs which vary systematically across a number of factors. This method is particularly appropriate for tasks such as engineering or experimental design where it is possible to clearly define the quality of a given solution.

### 1.7.2 Evaluation factors reported in the literature

Previous design studies particularly in the fields of engineering and architecture demonstrate many similarities in terms of evaluation. The consideration of cost (in terms of time, as well as financial) is universal and it is the consideration of cost versus quality which makes the design process difficult. By removing cost as a consideration a simple design heuristic of maximising the values of all the other attributes could be applied. However, this would not necessarily lead to parsimony of design but would be more likely to create inefficient designs which would safely satisfy evaluation requirements.

Cost can be seen as always reciprocally related to the other attributes of design quality, for example, power and validity in psychology experiments, accuracy, longevity or technical quality in engineering and numerous measures of quality in architecture including longevity, strength and aesthetics.

Thus in different types of design study there are similar factors involved in evaluation and findings from these studies are in general agreement. Evaluation is seen as a final stage where if evaluation criteria are met the potential design becomes the solution. If evaluation criteria are not met a feedback system operates and the designer returns to an earlier stage. Changes specific to the evaluation failure are made and the new design is re-evaluated. Although this process may be common to all design activity there are large differences in the actual procedures involved. In tasks where the evaluation criteria are specific and well defined, (e.g., engineering) a failure to meet one or more of the criteria is generally clear from a simulation of a potential design and often leads to quite small and specific changes in the potential design. With more fuzzy evaluation criteria the situation is not so clear, designs may have to be tested in the real-world. Failure at this stage is expensive and may lead to a complete re-design rather than minor adjustments. To take an example from architecture; socially disastrous tower blocks are routinely demolished and replaced with low level housing estates.

This situation is even more complex in experimental design where acceptance of a bad design may never be noticed without substantial replication of the experiment. An experiment which demonstrates the hypothesised effect may be showing a true effect or may

be subject to Type 1 error. Conversely an experiment may find no effect either because no effect exists or because of Type 2 error. Thus assessing the probability of Type 1 and Type 2 errors in a putative experimental design is crucial to the evaluation stage and yet errors made in this stage will only become apparent if an experiment has been replicated a number of times.

## 1.8 EXPERIMENTAL DESIGN

In order to draw any conclusions from the previous research which will remain valid for experimental design tasks, it is necessary to determine the level of similarity between the tasks used in this previous research and the task of designing an experiment. Using the categorisation scheme proposed earlier (Section 1.3.2) it is possible to generate a description of the factors involved in experimental design tasks in order to enable comparison with other areas of design.

In experimental design the initial design stage is inevitably multi-attribute; it involves determining appropriate values for at least six factors some of which are inter-dependent. Problem definition in the design stage is generally fuzzy. Levels of semantic knowledge required depend to some extent on the specific task, however, in most areas of psychology tasks are semantically rich. In the evaluation stage assessment of the quality of putative designs is both complex and fuzzy. The following sections present a detailed analysis of each of these specific attributes as related to the design of psychology experiments.

### 1.8.1 Design attributes

The important attributes in the design of psychology experiments are, ethics, external validity, internal validity, experimental power and cost. The question of ethics may be seen as different to the other attributes as it holds a veto over the whole design. A design with bad ethics will not be run whereas a design with bad validity or power may be run anyway, it will just be less efficient. Despite this difference ethics must still be considered (at least as an evaluation attribute) as there is always a need to check any prospective design for possible

ethical considerations. Thus there are at least four main attributes to be considered. These are also inter-related in a complex manner and each one is dependent on a number of factors.

External validity is the extent to which the results of an experiment can be generalised to the real-world. This is dependent on a number of factors relating to choice of hypotheses, experimental method and subject selection. External validity is to a large extent a fuzzy concept. Validity is enhanced by making the experimental situation as close to real life experience as possible. This is, however, inevitably balanced against the need for experimental control in order to reduce extraneous and possibly confounding variables. In terms of evaluation there is no definitive measure and the external validity of a given experimental finding may only become apparent after a considerable number of replications under differing conditions. Under these circumstances the designer can only rely on heuristics to avoid obvious problems.

Internal validity is the extent to which an experiment tests a given hypothesis. It is enhanced by avoiding confounds and the possibility of alternative hypotheses and by constraining the probability of Type 1 error. The avoidance of confounds and alternative hypotheses involves the control of all extraneous factors which may have an influence on the dependent variable. In practice the designer can only hope to have recognised any confounding variables amongst the infinite number of variables present in any given situation. This process involves a mental search of the problem space (Klahr et al. 1993) and there is no prescriptive method for evaluating the completeness of this search. Again the presence of confounds may only be discovered through subsequent replications under differing experimental conditions.

Unlike the previous factors related to internal validity, evaluation of the probabilities of Type 1 and Type 2 errors is to some extent calculable. The probability of Type 1 error is pre-set by the experimenter as the level below which an effect will be regarded as significant. This significance level is generally standardised in the behavioural sciences at less than $p = 0.05$. The probability of Type 2 error (the probability of failing to discover an effect where one exists) is much more difficult to determine. The power of an experiment is the probability (usually expressed as a percentage) of avoiding Type 2 error. Experimental power is influenced by a number of factors and estimating power can be quite complex.

Given a specific size of effect (and predetermined significance level) the power of an experiment fundamentally depends on the sample size; with power increasing as sample size increases in a complex non-linear relationship known as the power curve. Thus it is possible to calculate the exact number of subjects an experiment requires for a given power. However, this requires the experimenter to know in advance the expected effect size (a large effect requiring lower subject numbers for equivalent power). (For examples and mathematical analysis of these relationships see Lipsey, 1990) Appendix 7 describes the definition of effect sizes used throughout this thesis.

Whilst the need to know expected effect sizes introduces an element of fuzziness into power calculations this can be dealt with by adopting standard guidelines for small, medium and large effects such as those suggested by Cohen (1988). The adoption of such guidelines and the quoting of effect sizes in the literature would prevent two common errors of experimental interpretation. Those due to excessive rates of Type 2 error and those due to the reporting of significant results which relate to effects so small as to be of no practical importance (See Cohen 1990). However, as yet categorising and reporting of effect sizes is far from common practice despite these obvious advantages.

An additional factor influencing experimental power is the accuracy of measurements of both the dependent and independent variables. The fact that any form of random measurement error (or indeed anything increasing the variability of observations) will reduce power, introduces more uncertainty into experimental power analysis. Thus an estimate of experimental power should be seen as the maximum available power given perfect experimental conditions and not a definitive probability of Type 2 error.

### 1.8.2 Problem definition

The original task definition is an area where the design of psychology experiments differs greatly from other fields of design. The starting point of design of a psychology experiment is a hypothesis which defines the design and evaluation attributes to a greater or lesser extent. The overall design and evaluation goal of any experimental design is that it should test the related null hypothesis. However, it is rare in psychology that a hypothesis is sufficiently detailed for this evaluation to be specific.

In engineering design it is usually the case that task instruction consists of explicit requirements of the final design. In these and similar studies it is commonly reported that the early stages of design involve clarification of specific design requirements (Stauffer and Ullman 1988). In experimental design this stage may be present but it would be qualitatively different. The hypothesis is unlikely to contain detailed information on requirements for subject populations or for independent and dependent variables and their measurement; decisions on these factors are usually seen as part of the design process itself. For example, it is possible to start a design with a clearly specified hypothesis such as 'Manual reaction time to the stimulus of a red light increases in a background noise condition consisting of a continuous tone of 200Hz at 40db, for right handed boys aged between the ages of 16 and 18'. In practice a much more likely hypothesis would be 'Reaction time increases in background noise' the specific details of the variables and subjects being left to the designer (who is often the person who created the original hypothesis anyway). This complicates the design and evaluation processes and although experimental design is a scientific pursuit based on specific statistical rules and procedures its related design processes may have a lot in common with the more fuzzily defined areas of design.

Smith (1989) explores some effects of unstructured decision problems. Where initial problem definitions are incomplete there is often a failure to identify key problem elements. This constrains the scope of problem solving activities and harms overall performance.

### 1.8.3 Semantic knowledge

Required levels of semantic knowledge associated with experimental design are high. The minimum level of domain knowledge required to successfully design an experiment is an understanding of statistical research methods. This involves a working knowledge of all the design attributes and the relations between these attributes described in Section 1.7.1. Within psychological research there are also domain specific research methodologies which may lead to additional effects of expertise which are limited to quite specific domains.

There is a large body of research confirming the facilitation due to expert (domain specific) knowledge in semantically rich domains (Abdolmohammadi and Shanteau, 1992, Bhaskar and Simon, 1977, Chan, 1990, Hammond et al., 1987, Sanbonmatsu et al., 1992,

Shanteau, 1992). It is of course self-evident that a problem requiring semantic knowledge will be easier for a person with expertise in this domain; i.e. with the required knowledge readily at hand. However, the overwhelming impression from this research is that the mere availability of domain knowledge is not as important as the expertise in applying this knowledge to the problem in hand. Schraagen (1993) supports this view by demonstrating that when experts are confronted with a novel problem in experimental design "their form of reasoning remains intact , but the content of their reasoning suffers due to a lack of domain knowledge" (page 285).

These selected effects of expertise are briefly mentioned here for the sake of completeness. It is beyond the scope of this thesis to form a comprehensive review of what is in itself a huge research area.

### 1.8.4 Evaluation attributes

Evaluation of experimental designs involves checking the appropriateness of decisions made in the design stage. The subsequent judgement of experimental quality should be based on some form of cost-benefit analysis. At a basic level the experiment should satisfy minimum requirements for ethics, cost (financial and practical constraints), internal and external validity. Beyond this level optimum quality involves obtaining the highest possible levels of internal and external validity for given cost constraints. Estimating these levels requires exactly the same processes as described in the design stage (see section 1.8.1) and these will not be repeated here.

It can be argued that evaluation in not a separate stage; rather that the design stage is one continuous process which is only complete when evaluation criteria are met. However, in the study of experimental design there are some good reasons for continuing to make this conceptual division into design and evaluation stages. Firstly, because of the ill-defined nature of experimental design evaluation criteria are never explicitly stated in the task. Determining what criteria the final design should meet must therefore be included in the overall design process. This may be a complex process as these factors are inter-dependent and optimum levels will of course vary depending on the hypothesis to be tested. For example, if a drug company is testing for possible side effects in a new drug it would be of

critical importance to constrain the possibility of Type 2 error. In this case failing to find an effect where one exists would lead to very damaging repercussions. On the other hand if the drug were being tested for a *desired* effect it might be more important to constrain Type 1 error in order to avoid introducing an ineffective drug.

It is also useful to view evaluation as a separate stage because in a wider view it is an important skill for researchers to be able to evaluate existing designs. When interpreting the importance of previous research studies and particularly in refereeing new research it is necessary to judge the quality of experimental designs. It is this evaluation which determines whether a particular interpretation of an experimental result enters the literature.

### 1.8.5 Findings from related research

Research into experimental design suggests the presence of a number of specific and intractable problems relating to the evaluation of experimental design. Foremost amongst these is the widespread failure of experimenters and reviewers to consider experimental power. Cohen (1962) has demonstrated that psychologists often severely overestimate the power of a prospective experimental design. In a meta-analysis of the 1960 volume of the Journal of Abnormal and Social Psychology, Cohen found the median power to detect a medium effect (Pearson r of 0.4) to be .46 (46%). Thus in these experiments there was a better than 50% probability of failing to achieve a significant result where a medium sized effect existed (Type 2 error). Since this original study there has been great interest in experimental power in the literature and a number of definitive works on power analysis have been published (i.e.Cohen, 1988). Despite this general interest the average power of experiments has not improved. Sedlmeier and Gigerenzer (1989) replicated Cohen's meta-analysis using the 1984 volume of the same journal only to find that twenty four years later median experimental power had fallen to .37 (37%). In addition only two out of sixty four experiments reviewed in this study mentioned power at all and it was never estimated. To make matters worse the authors report "Nonsignificance was generally interpreted as confirmation of the null hypothesis (if this was the research hypothesis) although the median power was as low as .25 in these cases". Clearly these findings demonstrate a failure to evaluate power both by the original experimenters and the reviewers of the subsequent

articles. It is hard to believe that all these researchers are either unaware of the basic need for sufficient experimental power or are incapable of formally analysing experimental power. It seems more likely that they are relying on an intuitive statistical judgement or utilising some sort of heuristic to judge power. In either case the subsequent power evaluation is clearly inadequate.

The results of these meta-analyses are entirely consistent with the results of a study by Tversky and Kahneman (1971) showing that experimental psychologists consistently overestimate the statistical power of experiments with small sample sizes. In addition Kahneman and Tversky (1972) have shown subjects' judgements to be unduly influenced by sample proportion to the detriment of considerations of sample size in a study based on the construction of sampling distributions. They suggest that this may be due to the use of a 'representativeness' heuristic rather than a logical statistical analysis of the problem. Evans and Dusoir (1977) have subsequently shown these errors to be mediated by problem complexity. However, in this study insight into the role of sample size was present only in extremely simple problem formats. Well, Pollatsek and Boyce (1990) report a series of studies specifically aimed at measuring subjects' understanding of the effects of sample size on the variability of the mean in which aspects of problems relating to sampling distributions were systematically varied. They conclude that although some subjects appreciated some aspects of the law of large numbers this "does not seem to result from an in-depth understanding".

The general failing of published psychology research to take account of power requirements and the research findings cited above demonstrate lack of intuitive statistical judgement to be a pervasive problem even amongst experts.

### 1.8.6 Bias in evaluation of psychology experiments

Having demonstrated pervasive and persistent failures of evaluation in the previous section, this section speculates on some possible sources of bias which might account for these problems. Some of these speculations will be tested in subsequent chapters.

As we have seen Cohen (1990) reports numerous errors of design and interpretation of results throughout published psychological experiments; these errors and resistance to

change are entirely consistent with designs biased by confirmation and positivity and based on overconfidence. Allied with Confirmation and Positivity biases overconfidence of judgement is likely to be detrimental to accurate evaluation of experimental design. Designers of experiments can be expected to ignore factors which may subsequently dis-confirm their results, ignore factors which negatively affect their design and yet remain confident that their design is much better than it actually is.

Overconfidence in intuitive judgements is a strong and well replicated effect in which there is a tendency for subjects to maintain a much greater confidence in the correctness of their judgements than is warranted. This research has been reviewed by Lichtenstein, Fischhoff and Phillips (1982) who conclude that people consistently overestimate what they know. There is evidence that experts making numerous similar judgements with clear feedback of success rates, (weather forecasters) do not suffer from overconfidence. However, clear feedback is almost non-existent in experimental design. In fact the immediate result of an experiment (either a significant or a non-significant result) should not be used as evidence of experimental quality as there is no way of knowing whether either of these results reflects the true state of the world or the presence of Type 1 or a Type 2 error (see section 1.7.2).

This failure of the outcome of an experiment to give useful feedback relating to the quality of the design is unusual. In the majority of design and judgement tasks the final outcome is at least a guarantee of the design having achieved sufficient quality. Under these circumstances it seems quite likely that the evaluation of experimental designs is influenced by the subsequent experimental results; indeed it would be difficult to ignore them. If this is the case then it is not surprising that problems arise in judgements of experimental quality based on incorrect feedback. Under these conditions these judgements can be regarded as a form of 'hindsight bias'.

Hindsight bias is also a form of overconfidence in judgement, first shown by Fischhoff (1975). It is the tendency for subjects to overrate the probability of an event when given the knowledge of its outcome (and similarly to underestimate the likelihood of an event given the knowledge that it did not occur). This effect includes a bias towards overestimating the relevance of factors supporting the known outcome. Within experimental

design hindsight bias could be expected to influence choice of experimental factors. Experience of previous experiments which have produced a significant result would lead to designers having an overconfident view of the relevance of the factors used in these successful experiments (particularly when experiments with negative results which may have used similar factors are rarely published). Thus whether through schemas, mental models or preconscious heuristics created on the basis of this past experience any new experimental design is more likely to be based on previous designs than on an objective analysis of required power. This situation seems on the surface to be a reasonably logical approach to design; based as it is on previous successes. However, when we consider that these apparently successful experimental results may have been due to consistent errors this approach can equally lead to persistent repetition of the kind of experimental shortcomings described above.

## 1.9 SUMMARY OF CHAPTER ONE

Chapter 1 reviewed a wide range of studies into the process of design and explored the difficulty of determining general conclusions from this highly disparate research field. As there were no previous studies into the design of experiments, a design task categorisation scheme was proposed in order to distinguish those findings which might generalise to the process of experimental design. Utilising this categorisation scheme relevant findings from both problem solving research and design research were discussed as were methodological problems in research into design processes in general. Subsequent discussion of these findings was presented in terms of a design stage and an evaluation stage.

Tentative conclusions for the design stage of experimental design predicted the presence of a task clarification phase early in the design process and the possibility of the subsequent creation of a hierarchical system of sub-goals leading to the formation of a putative design solution. Despite these tentative conclusions, the overwhelming view from previous research was of inconsistent and often contradictory results. It was proposed that this confusion of results might be based on methodological problems associated with

determining the nature of the largely implicit cognitive processes in the creative design stage.

It was possible to form more reliable conclusions in terms of the evaluation of experimental design. It was predicted that as in other design areas there would be some form of evaluation and re-design loop only terminating when a putative design was deemed acceptable. The clear differences between the rather fuzzy evaluation of putative experimental designs and the straightforward evaluation in areas such as engineering design were also noted. In the light of these differences the specific problems associated with evaluating the quality of experimental designs were discussed.

Previous psychology research was also reviewed. This research has reported consistent failure to consider the effects of experimental power and sample size in experimental design. Possible reasons for the persistence of these failings were discussed in terms of overconfidence and hindsight in the process of evaluation of experimental designs

## Table of contents - Chapter two

# CHAPTER TWO

## Two initial studies

## 2.1 CHAPTER TWO INTRODUCTION

This chapter comprises the first two experimental studies. These studies were designed to explore evaluation in the design of psychology experiments and the factors involved in this evaluation.

The first study explored the extent to which evaluation of experimental quality played a part in the design of psychology experiments. This study utilised concurrent verbal protocols of research students designing psychology experiments. In this study subjects were presented with experimental hypotheses and asked to design experiments to test these hypotheses. In addition to concurrent verbal protocols subjects were required to complete an answer sheet listing the details of their final designs. This was a prospective study designed only to elicit evidence of the extent of evaluation processes in experimental design. Due to the methodological problems outlined in Chapter 1 (Sections 1.5.1 and 1.6.2) it was decided not to attempt further analysis of these verbal protocols beyond a gross measure of the extent of subjects' considerations of evaluation factors. From the answer sheets it was also possible to gain some measure of the quality of the resulting experimental designs.

The second study focused directly on the evaluation of completed designs. This study utilised a more formal task structure in which a number of sampling procedures were presented to subjects who were asked to judge their quality. These sampling procedures varied in terms of sample size, pre-set significance level and experimental outcome. The experiment was designed to assess the relative influence of these factors on subjects' judgements of experimental quality. As experimental outcome information was available to subjects, it was predicted that the results of this study would demonstrate the presence of significant outcome bias in the evaluation process.

## 2.2 EXPERIMENT ONE

**A prospective study of evaluation stages in the design of Psychology experiments using concurrent verbal protocols**

### 2.2.1 Introduction

This study explored the nature of design processes used by psychologists in designing studies to test given hypotheses. As there were no previous studies in the field of psychology involving realistic experimental design tasks it would have been premature to generalise results from other design areas. In Chapter 1 it was argued that the immense variation in design tasks precludes direct comparison between different design processes without reference to their specific attributes; as described in the suggested classification scheme (Section 1.3.2). In order to place experimental design in the general field of design research it was therefore necessary to complete an exploratory study. This study had the intention of discovering common cognitive strategies and approaches and delineating differences to other design tasks in the specific area of design evaluation.

The present study used verbal protocols of experimental design tasks completed by experienced subjects proficient in the design of psychology experiments. Although it has been shown that the use of protocol analysis may be inappropriate for the study of the creative processes of design (see Sections 1.5.1 and 1.6.2) it remains the only practical method of determining general strategies in complex real life decision making tasks. As mentioned above (Section 1.3.1) three general stages can be assumed to be present in any design process; a task definition (evaluation) stage, a creative design stage and a design evaluation stage. Bearing in mind the limitations of this methodology, this prospective study was designed only to explore the extent and nature of the evaluation stage in the experimental design process. In addition it would be possible to check the resulting verbal protocols for the presence of an early task clarification stage. Thus a full analysis of the verbal protocols was not intended and no conclusions would be drawn on the nature of the creation of putative experimental designs. Analysis of the protocols was thus limited to a basic categorisation of verbalisations relating to evaluation either of the initial task or of the factors

relating to putative designs.

The four tasks used in this study were presented in the form of experimental hypotheses (See Section 2.2.2 below). In each case the subject was required to design an appropriate experiment to test the hypothesis. The four hypotheses (A,B,C andD) were designed to elicit specific design factors. Hypotheses A and D varied in terms of expected effect size. In hypothesis A it could be predicted that differences in table height would only have a small effect on writing speed. In hypothesis D the difference in recognition times between pictures of close friends and people only seen very rarely could be predicted to be very large. If subjects were properly considering experimental power requirements it would be appropriate for them to use a greater number of subjects in an experiment to test hypothesis A than in an experiment to test hypothesis D. It was predicted that subjects would fail to make appropriate considerations of experimental power; in line with the general failures to consider power reported by Cohen (1962). Hypothesis B presented physical pain as an independent variable in order to explicitly cue the need for ethical considerations. It was predicted that ethics would only be considered as a factor of experimental design where it had been specifically cued.

The above three questions required relatively simple one-way designs. Question C introduced greater difficulty in terms of two independent variables requiring a two-way design (background noise and mathematical ability). This makes the overall task more difficult and particularly makes any initial task clarification stage more complex.

### 2.2.2 Method

#### Subjects

Subjects were ten psychology postgraduate research students studying at the University of Plymouth . The subjects were experienced in the kind of task presented (design of psychology experiments) but were not necessarily expert in the particular areas of experimental design required. Two subjects were used in the pilot study and the remaining eight in the experiment proper.

## Materials and Procedure

Pilot study

It was the intention of this study to be as close to real life experimental design as possible. This meant presenting the design task with the minimum of instruction to prevent cueing of the subjects in the categories of answer required for a successful design. The initial intention was only to present subjects with very basic instructions such as:

> **Design an experiment to test the hypothesis - 'Differences in table height affect writing speed'**

However, it was recognised that a completely open task devoid of any instruction might serve to confuse subjects. A complete lack of task constraints would be extremely rare in a real life experimental design, where during the process of generating an experimental hypothesis task constraints would normally become apparent as the hypothesis was refined. Given this lack of direction subjects might use an initial task clarification phase to generate a more specific hypothesis including constraints of their own choosing. If this was the case it would result in subjects effectively attempting different tasks thus precluding any between subjects analysis of results. It was also recognised at a purely practical level that some time constraints might be necessary. For these reasons a pilot study was carried out on two of the experimental subjects.

Each subject was presented with the following four problems:

A. Design a study to test the hypothesis:
**Differences in table height affect writing speed.**

B. Design a study to test the hypothesis:
**Physical pain affects reaction time to non-painful stimuli.**

C. Design a study to test the hypothesis:
**Background noise will increase solution times for quadratic equations for more mathematically able subjects, less able subjects will not be affected.**

D. Design a study to test the hypothesis:
**Faces of close friends in photographs are more readily recognisable than faces of people only seen very rarely.**

Subjects were also given an instruction sheet containing the following guidance to

thinking aloud at the outset of the experiment. This sheet remained on the table in front of them throughout the experiment as did the question sheet on which they were allowed to jot down notes if required.

> **Thinking aloud;**
>
> During this experiment it is important to talk aloud, you should try simply to report whatever thoughts are in your head at the time. You will not be judged on what you say so please try to relax and be spontaneous, the focus of the experiment is on what you are thinking about at a particular time, rather than whether or not these thoughts make sense. It is not necessary to explain or justify your thoughts to the experimenter.
>
> Experimental recordings will be kept strictly confidential and will be erased once the verbal protocols have been transcribed. The resulting written transcriptions will not be traceable to individual subjects.

Subjects were given an hour to complete the tasks. No other instructions were given.

Results from the pilot showed the necessity of a more well defined problem structure particularly in terms of guidance on time constraints. One subject spent the whole hour on the first task without ever achieving a completed design whilst the other subject felt they had completed all four tasks within the first twenty minutes and could not think of any more details which might be relevant. Neither subject presented enough detail for a complete experimental design in terms of subject population, required subject numbers, within or between subjects analysis, etc.

Main study

It was decided to define the task requirements and time constraints more rigidly for the main study. Firstly, subjects were instructed that they had fifteen minutes for each question. Secondly, an answer sheet requesting specific details of the completed design was presented to the subjects after the first ten minutes on each task. This had the disadvantage of cueing the subjects to the requirements of a completed design on all but the first task, but ensured the availability of explicit details of designs.

The following task instructions were added to the thinking aloud instructions:

> **Task Instructions**
>
> You will be presented with four problems, in the form of experimental hypotheses. Your task, in each case, is to design a study to test the hypothesis. Your design should attempt to find the best possible method of testing the given hypothesis.
>
> During the task it is important that you talk aloud describing what you are thinking about at that moment. If you fall silent you will be prompted to continue talking. You may jot down any notes you need on the question sheet, you will have fifteen minutes to consider each question. After the first ten minutes you will be given an answer sheet requesting specific information about the study.

The above four questions were then presented one at a time. After ten minutes subjects were presented with an answer sheet requesting details of their experimental design. (See appendix one for complete instructions and answer sheet). The order of question presentation across subjects was balanced using a Latin square. Verbal protocols were recorded and later transcribed from audio tape. Video tape recordings were also made in order to record the timings of written notes; these were not used in the present analysis.

### 2.2.3 Results and Discussion

The following results were based on the coding of utterances from the experimental verbal protocols into sixteen categories. (See appendix one for coding scheme and coded results) Results for subjects' estimates of required sample sizes were taken from written answer sheets. As there was no formal verbal protocol analysis these prospective results will be discussed as they arise.

Categorisation of utterances was performed by the experimenter. In order to check the reliability of this categorisation two randomly chosen protocols were categorised by a blind judge. These categorisations were scored and correlated with the originals. The reliability coefficient was 0.88.

Consideration of independent and dependent variables

By far the greatest proportion of utterances were related to considerations of independent and dependent variables. These factors accounted for 30.8% of all utterances

and were generally related to setting levels of the independent variable and appropriate measurement of the dependent variable.

Confounds and practicality

The second greatest proportion of utterances were related to considerations of possible confounding variables and to the practicality of the physical running of putative experiments. These factors accounted for 25% of all utterances. Confounds and the physical practicality of experimental designs have been assessed together as there was a great deal of overlap between these categories and they were often both contained in a single utterance. This is not surprising when you consider that the discovery of a potential confound will nearly always lead to a change in the physical way in which an experiment is run in order to avoid this possibility.

Evidence of a task definition stage

There was clear evidence of an initial task definition stage. In 24 of the 32 trials (75%) clarification of the task was included in the first three utterances. There were only 4 cases where task definition was not referred to at all. Overall, 9.6% of utterances were concerned with task definition. However, this figure may be artificially low as an attempt to define either the independent or dependent variables stated in the question would be classified as a consideration of those variables rather than an attempt to clarify the task .

Ethical considerations

Explicit references to ethical considerations were only made in response to question B which included physical pain as the independent variable. Five of the eight subjects referred to ethics in this question. Overall 13.7% of all responses to question B were concerned with ethics. No considerations of ethics were made in any of the other questions. This result was in line with the prediction that explicit consideration of ethics would only be present where ethical problems were cued by the question content. (See section 2.2.1)

Evidence of evaluation

Overall there was surprisingly little evidence of explicit evaluation of experimental designs. Any statement relating to considerations of experimental power, potential effect sizes or the probabilities of either Type 1 or Type 2 error was considered an instance of evaluation. This category also included any return to and reconsideration of an experimental variable already determined or any reference to the overall appropriateness of the experimental design. Of the eight subjects in the study only four made any reference to evaluation of their putative designs. One of these four made no evaluations on the first two questions (Questions A and B) and only made one reference to evaluation on each of the remaining two questions. The most complex question (question C) attracted the largest amount of evaluation. Evaluation accounted for 5.8% of all utterances. However, this overall figure was subject to considerable individual differences. See Table 2.1 for details.

Table 2.1

Overall number of utterances related to experimental evaluation

| | Question A (Table height) | Question B (Physical pain) | Question C (Noise and maths ability) | Question D (Face recognition) | **Total evaluation** |
|---|---|---|---|---|---|
| Subject 3 | 0 | 0 | 0 | 0 | **0** |
| Subject 4 | 0 | 0 | 0 | 0 | **0** |
| Subject 5 | 0 | 0 | 0 | 0 | **0** |
| Subject 6 | 0 | 0 | 0 | 0 | **0** |
| Subject 7 | 0 | 0 | 1 | 1 | **2** |
| Subject 8 | 3 | 0 | 5 | 1 | **9** |
| Subject 9 | 3 | 3 | 5 | 3 | **14** |
| Subject 10 | 1 | 8 | 3 | 2 | **14** |
| **Total** | **7** | **11** | **14** | **7** | |

Considerations of subject numbers and experimental power

There was no evidence whatsoever of any explicit consideration of experimental power. None of the utterances from any subject on any question were categorised as referring to experimental power. However, it is possible that subjects were making implicit judgements of appropriate power based on knowledge of expected effect sizes and the determination of requirements for significance levels and subject numbers.

There was also no evidence of any utterances referring to significance levels. This

was not surprising given the almost universal practice of using the $p < 0.05$ level as a standard for significance in psychology experiments. Although it is theoretically reasonable to determine and pre-set an appropriate significance level for any given experiment, the use of any other level has become extremely rare (in any area other than medical research where the more stringent level of $p < 0.01$ may be used as a standard). Thus it is likely that subjects automatically assume a level of $p < 0.05$ and may utilise a simpler intuitive estimate of experimental power based only on knowledge of expected effect sizes and the subsequent determination of subject numbers. However, as any mention of expected effect size would have been categorised as a consideration of experimental power there was also no evidence of any subject making any explicit reference to effect size.

This leaves the setting of subject numbers as the only possible determinant of appropriate experimental power. There is also very little evidence of subjects determining required numbers of subjects during the design process. In this experiment subjects were cued to the need to determine subject numbers on all but the first question by the presentation of an answer sheet requesting this information at the end of each design phase. Despite this clear cue only four of the eight subjects made any determination of required subject numbers before reaching the point on the sheet where they had to fill in a specific value. In addition none of these four subjects determined subject numbers for all of the four questions they attempted. (See Table 2.2 for details) In general this meant that subject numbers were only decided upon as an after-thought when the experimental design was apparently complete. Even then figures were only supplied when specifically requested. It would appear that subjects did not consider this information to be part of the design process.

Table 2.2

Instances of considerations of subject numbers and decisions made during the design process

| | Question A (Table height) | Question B (Physical pain) | Question C (Noise and maths ability) | Question D (Face recognition) | **Total decisions** |
|---|---|---|---|---|---|
| Subject 3 | 1 | 1 * | 0 | 2 | **2** |
| Subject 4 | 0 | 0 | 0 | 0 | **0** |
| Subject 5 | 0 | 0 | 0 | 0 | **0** |
| Subject 6 | 1 | 1 | 0 | 0 | **2** |
| Subject 7 | 0 | 2 | 1 * | 0 | **1** |
| Subject 8 | 0 | 0 | 0 | 0 | **0** |
| Subject 9 | 0 | 0 | 0 | 0 | **0** |
| Subject 10 | 1 | 2 | 0 | 2 | **3** |
| **Total decisions** | **3** | **3** | **0** | **2** | |

N.B. * denotes a final decision was not made

Each subject was compelled to give a figure for required numbers of subjects in the subsequent answer sheet stage of the experiment. Thus it was also possible to analyse these forced decisions which are presented in Table 2.3.

Table 2.3

Forced decisions on subject numbers

| | Question A (Table height) | Question B (Physical pain) | Question C (Noise and maths ability) | Question D (Face recognition) |
|---|---|---|---|---|
| Subject 3 | 30 | 20 b | 16 b | 30 |
| Subject 4 | Not specified | 20 b | 50 b | 20 |
| Subject 5 | 20 b | Not specified | 10 b | 30 |
| Subject 6 | 40 b | 50 b | 15 b | 100 |
| Subject 7 | Not specified | 10 | Not specified | 5 b |
| Subject 8 | 10 | Not specified | 20 b | Not specified |
| Subject 9 | 60 | Not specified | Not specified | Not specified |
| Subject 10 | 40 | 10 | 40 | 20 |
| **Mean subject numbers** | **33.3** | **22** | **25.2** | **34.2** |
| **Mean experimental power** | **20%** | **50%** | **54%** | **95%** |

N.B. The letter b after a number denotes a between subjects design in which case subject numbers are for each subject group. The figure for mean subject numbers is therefore per subject group in any given design.

Despite the absolute requirement of the answer sheet for subjects to provide specific figures for sample sizes there was a general unwillingness to comply. In nine cases values were not specified. From the figures that were provided there was no evidence for considerations of experimental power. Two questions were specifically designed to elicit power considerations; question A had a potentially very small effect and thus required a large sample whilst question D had a potentially very large effect and thus only needed a small sample. The other two questions required a sample somewhere between these two extremes. As a guide to the sort of sample sizes required the following estimates were calculated on the basis of the power of a t-test at a significance level of $p < 0.05$. These calculations used Cohen's (1988) definitions of small and large effect sizes ($d = 0.2$ and $d = 0.8$ respectively) to achieve a relatively conservative experimental power of 75%. (i.e. 25% chance of Type 2 error) Values given are for within subject designs and should be doubled for between subject designs.

For question A (small effect) this would require 275 subjects. The mean value of subjects' estimates of the sample required for this question (33.3) reflects an experimental power of 20%. For question D (large effect) only 18 subjects would be required to achieve a power of 75%. The mean value of subjects' estimates of the sample required for this question (34.2) reflects an experimental power of 95%. For a medium effect size ($d = 0.5$ ) 44 subjects would be required to achieve an experimental power of 75%. The mean values of subjects' estimates of the sample required for questions B and C reflect experimental powers of 50% and 54% respectively. Calculations of mean experimental power have been included in Table 2.3.

From these results it is clear that subjects had a strong tendency to ignore sample size in the design process. When forced to provide a value they seriously underestimated sample size requirements and generated designs with very high probabilities of Type 2 error. This reluctance to use large sample sizes may have been due to considerations of experimental cost. However, there was little evidence of explicit references to cost in the verbal protocols. Cost was only mentioned by four of the subjects in a total of five of the 32 completed experimental designs. Considerations of cost amounted to less than one percent of all responses.

### 2.2.4 Conclusions

The results of this experiment showed a strong concern by the designers for the practicalities and mechanics of the potential experiment and virtually no consideration of the probabilities of Type 1 and Type 2 errors in any subsequent results this experiment might generate. In general there was evidence for an initial clarification stage. This stage was followed by considerations relating to manipulation and measurement of experimental variables and consideration of possible confounds and the physical practicality of the potential experiment. There was little evidence of any form of explicit evaluation of putative designs. Evidence of any form of power analysis (explicit or implicit) was completely absent with estimates of sample sizes reflecting completely inadequate experimental power for designs testing small or medium effects.

This failure to consider power had been predicted (see section 2.2.1). However the lack of any form of explicit evaluation of putative designs was surprising and had not been reported in previous design research. This may be due to task differences in different areas of design research as discussed in chapter one (section 1.7.2). The task used in this case was in the form of a minimal hypothesis and specified no particular limitation criteria. This is quite unlike the clearly defined tasks typical of engineering design studies which are often in the form of a list of criteria to be met by the final design and thus cue the need for specific evaluations.

It is possible, therefore, that psychologists are quite capable of efficient experimental evaluation but will only make this evaluation when specifically cued to do so by question content. Experiment two attempts to assess this possible capability more directly by eliminating the design stage and asking subjects to evaluate the quality of existing experiments.

## 2.3 EXPERIMENT TWO

### A study of evaluation of the quality of sampling procedures

### 2.3.1 Introduction

The results of Experiment 1 demonstrated little evidence of formal experimental evaluation during the design of psychology experiments. Nevertheless these formal evaluations are made by psychologists on a regular basis when reviewing the experiments of others. These evaluations are particularly important in such areas as peer review of potential journal articles and the examination of PhD theses. The present experiment was designed to explore this type of evaluation by presenting subjects with descriptions of sampling procedures which they were required to judge for quality.

The quality of a given experimental design is largely dependent on the magnitude of chosen experimental factors. Assessing these factors is normally a complex multi-attribute task including considerations of confounding variables, practical restraints, measurement issues, ethical considerations, choice of statistical analysis, etc. Experiment 1 showed that most of these factors were considered in the design process, but evidence of consideration of factors specifically relating to experimental power was not apparent in this study. For this reason Experiment 2 focused on a simplified task which included only the most basic factors of sample size, chosen significance level and cost of sampling.

In this case rational choice of appropriate subject size and significance level in a given experiment depends on a cost-benefit analysis balancing the probability and cost of Type 1 and Type 2 errors against the cost of sampling. Thus a design of high quality would retain sufficient power and low probability of error at the minimum cost.

As described in Chapter 1 (Section 1.8.5) there is little evidence in previous research to suggest that subjects would approach these choices rationally, findings generally show a lack of understanding of statistical principles in both students and researchers. The results of this previous research were supported by the results of Experiment 1 in which experienced research students failed to take due consideration of experimental power requirements when designing experiments. In this second experiment it was hoped that systematic variation of

sample sizes and pre-set significance levels in experimental descriptions would permit a more direct assessment of the nature and extent of subjects' understanding of the principles of statistical sampling. Quality judgements should be based on a clear understanding of the statistical implications of varying these factors. However, previous studies strongly suggest that they will not reflect this understanding.

As this experiment was designed in part to reflect the evaluation task facing the reviewer of an experimental study it was decided to include the results of each sampling procedure. This information is available in the normal course of evaluation of any previous study and is likely to significantly influence a reviewer's judgement of quality in hindsight. (See Chapter 1, Section 1.8.6) The given result of an experiment should in fact have little or no relevance to the quality of the original design. A significant result may be a true finding (demonstrating at least sufficient experimental power) or may be a Type 1 error. A non-significant result may be a true finding or a Type 2 error; representing no additional information either way. Thus without additional long-term outcome information on replication, initial experimental results are no help in judging the original quality of an experimental design. Even with this information there are only minor implications for the quality of the original experimental design. Nevertheless, it was decided to include an experimental condition in which subjects were presented with long-term outcome information on the replication or otherwise of the initial experimental results given in each description. The interaction of long-term outcome information with the initial experimental result created four different overall possible outcomes: a significant result subsequently shown to be correct, a significant result subsequently shown to be false, a non significant result subsequently shown to be correct and a non significant result subsequently shown to be false. Each of these four long-term outcomes had specific implications in the experimental scenario presented. (See below for details)

### 2.3.2 Method

#### Subjects

Subjects were 72 first year psychology undergraduates at the University of Plymouth. The experiment took place at the end of their first year when they had all completed a research methods and statistics course. Subjects participated on a voluntary basis and were allocated to experimental conditions by random distribution of question booklets; ensuring only that equal numbers of each type of question booklet were distributed. All responses were anonymous.

An analysis of experimental power based on the the calculations of Lipsey (1990) and assuming a medium effect size estimated that a minimum of 65 subjects would be required to achieve experimental power of 80% in the overall analysis of this experiment.

#### Materials and Procedure

Materials comprised an instruction sheet and a question booklet. The instructions contained an introduction followed by a scenario description. Having read the instructions subjects were presented with a booklet containing sixteen short descriptions of experiments. For half the subjects these descriptions were followed by long term outcome information.

Subjects were required to make a judgement of the quality of the experimental design of the described experiment in each case.

Instructions and scenario were as follows;

**Introduction**

The following description and series of questions are designed to study understanding of the concept of quality in experimental design. Your co-operation in this experiment is greatly appreciated, and the results should be of use to you as they will demonstrate general levels of understanding of experimental design principles, results and explanations will be made available as soon as possible. You do not need to put your name on this booklet.

**Scenario**

Over the last five years all the major drug companies have been competing to licence synthesised plant extracts. A large number of these new drugs have been created to treat

hypertension, all these drugs are designed to reduce blood pressure. They have all been experimentally tested on human subjects to see if they do significantly reduce blood pressure. These initial clinical trials can be expensive to run and the cost of developing these drugs is high.

In initial trials, experimental methods vary between companies. A brief description of the experimental methods used in a selection of comparable drug trials from four different companies is outlined below (the companies are not named). All these experiments used between subjects designs comparing experimental (drug) conditions with placebo conditions they were all analysed using t - tests.

**Task**

Your task is to read each description and give an estimate of the quality of the experimental design of the initial experiment by writing a percentage mark (from 0 to 100) in the space provided. Please try to use the whole range of marks.

A typical question (with outcome information) was as follows;

1.

In an experiment to test the effects of the drug Largacil 15 adult subjects were each given the standard therapeutic dose. Resulting blood pressure levels were compared with a placebo group of 15 subjects, using a t - test. Significance levels were set at 0.01. A significant reduction in blood pressure was found.

**Outcome :** Subsequent clinical trials confirmed this effect on blood pressure. The drug was licenced for clinical use and has made profits for this company.

| Quality score | % |
|---|---|

For details of all sixteen questions see Appendix 2.

## Design

Three experimental factors (each with two levels) were varied on a within subjects basis. These factors were: subject numbers (100 or 30), pre-set significance level (.05 or .01) and experimental result (significant result or non-significant result).

As exposure to long-term outcome information would cue subjects to consider the validity of initial results, the presence of long-term outcome information was manipulated as a between subjects factor (36 subjects receiving long term outcome information and 36 receiving only the initial result).

The manipulation of these four factors within a standard experimental description created the sixteen different sampling procedures presented to the subjects (see example

above and Appendix 2). In these experimental descriptions all other factors were held constant. All experiments were based on the testing of a drug designed to reduce blood pressure. All experiments were between subjects designs comparing a drug condition to a placebo condition. The dependent variables in each case were measures of blood pressure. Wording of experimental descriptions were varied to prevent cueing of subjects to the relevant design factors.

For those subjects receiving long-term outcome information this also had two levels (experimental result subsequently shown to be correct or subsequently shown to be false). These were manipulated on a within subjects basis for this subject group. For these subjects the interaction of long-term outcome information with the initial experimental result created four different possible overall outcomes: a significant result subsequently shown to be correct, a significant result subsequently shown to be false, a non significant result subsequently shown to be correct and a non significant result subsequently shown to be false. (See Appendix 2 for details of presentation of outcome information)

Although this factor created an unbalanced design it was still possible to utilise an overall five factor analysis in which the presence of long-term outcome information was a between subjects factor. This meant that the nature of long-term outcome (experimental result subsequently shown to be correct or subsequently shown to be false) was analysed as a within subjects factor but was a mock factor for the group of subjects not receiving this information.

The dependent variable was the subjects' intuitive judgement of the quality of each experimental design. Subjects were instructed to use a range from 0 to 100 for quality scores.

A normative analysis of the power of these sampling procedures to discover a medium sized effect (using the stated t-test) gives the following results:

30 subjects at $p < 0.05$ = 47% power

100 subjects at $p < 0.05$ = 94% power

30 subjects at $p < 0.01$ = 24% power

100 subjects at $p < 0.01$ = 82% power

### 2.3.3 Results

Influences of all factors

A five factor Anova was performed with four within subjects factors (subject numbers, pre-set significance level, initial result and long term outcome) the fifth factor the presence or absence of long term outcome information being between subjects.

Effect sizes for main effects were calculated following the recommendations of Cohen (1988). Values given are for product moment correlation coefficients (r) and are quoted in square brackets. Using this convention definitions of small, medium and large effect sizes are 0.1, 0.3 and 0.5 respectively. (Cohen, 1988; Page 532)

In this overall analysis both subject numbers ($F = 59.0$, $p = 0.001$) [$r = 0.74$] and pre-set significance level ($F = 16.1$, $p = 0.0001$) [$r = 0.32$] had significant influences on quality ratings. Experimental descriptions with higher subject numbers were rated as having higher quality; as were those with smaller pre-set significance levels.

In addition both the experimental result ($F = 6.0$, $p = 0.017$) [$r = 0.2$] and long-term outcome information ($F = 15.9$, $p = 0.002$) had significant influences on quality ratings. Overall significant experimental results attracted higher quality ratings than non-significant results and replicated results attracted higher quality ratings than non-replicated results. However, the precise nature of these effects was masked by the fact that in this analysis the nature of long-term outcome information was a mock factor for one subject group. That this factor influenced these main effects was demonstrated by the presence of a significant three-way interaction between the experimental result, the presence of outcome information and the nature of this information (experimental result replicated or not replicated), $F = 5.4$, $p = 0.02$. In order to gain a clearer picture of the influences of these factors it was necessary to perform separate analyses of the two different subject groups (those receiving and those not receiving long-term outcome information).

Table 2.4 shows the nature of the influences of all factors in this initial overall analysis.

Table 2.4

Means for experimental factors from overall analysis
(Standard deviations in brackets)

| | | | 30 Subjects | | 100 Subjects | |
|---|---|---|---|---|---|---|
| | | | Significance level 0.01 | Significance level 0.05 | Significance level 0.01 | Significance level 0.05 |
| No outcome information given | True Result (Mock factor) | Significant | **56.2** (17.1) | **49.6** (18.9) | **66.9** (21.6) | **62.1** (21.3) |
| | | Non-significant | **51.1** (17.8) | **50.1** (20.1) | **64.0** (21.5) | **60.9** (19.6) |
| | False Result (Mock factor) | Significant | **56.1** (17.6) | **50.4** (17.7) | **71.8** (19.0) | **61.4** (20.3) |
| | | Non-significant | **49.5** (20.1) | **49.5** (22.0) | **62.6** (21.7) | **58.7** (21.6) |
| Outcome information given | True Result | Significant | **56.4** (24.9) | **54.6** (23.0) | **72** (17.2) | **66.5** (18.8) |
| | | Non-significant | **54.5** (22.1) | **51.1** (23.6) | **64.5** (22.1) | **56.4** (22.4) |
| | False Result | Significant | **47.3** (17.7) | **41.2** (17.9) | **59.2** (24.4) | **50.7** (21.5) |
| | | Non-significant | **49.6** (19.5) | **42.7** (21.7) | **55.3** (21.7) | **54.0** (18.9) |

## Influences of experimental results and long-term outcome information

A. Analysis for subject group receiving experimental result only

An Anova was performed on the data for those subjects who did not receive long-term outcome information, this had three within subjects factors (subject numbers, pre-set significance level and initial result).

In this analysis (as in the previous analysis) both subject numbers ($F = 40.7$, $p = 0.001$) and pre-set significance level ($F = 6.3$, $p = 0.017$) had significant influences on quality ratings. Experimental descriptions with higher subject numbers were rated as having higher quality, as were those with smaller pre-set significance levels.

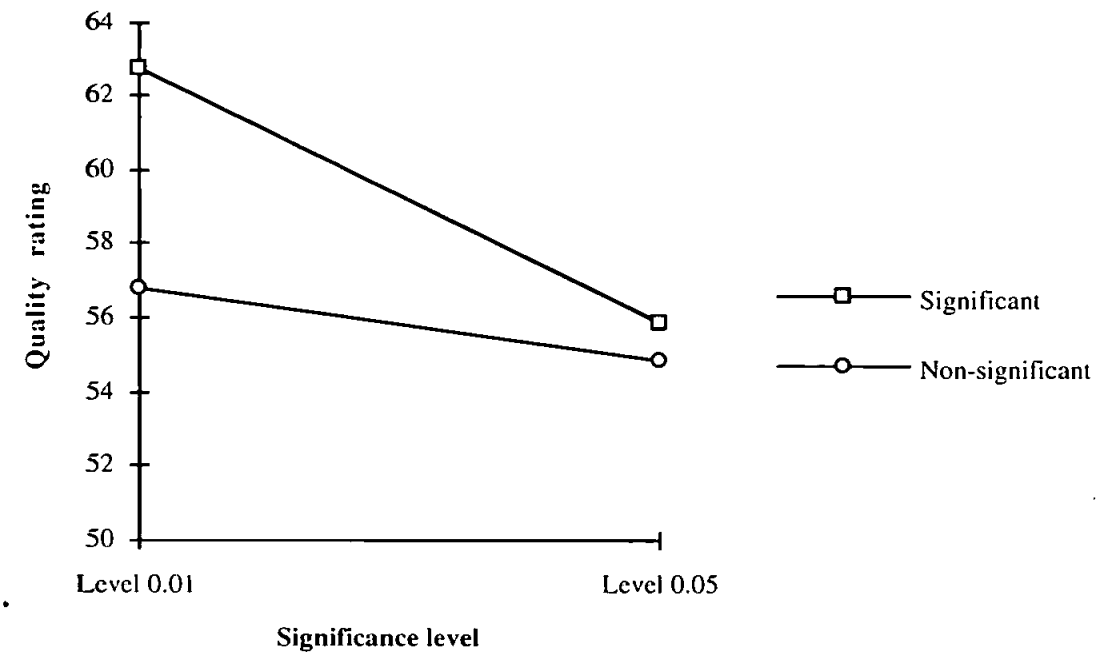
However, for these subjects ($n = 35$) the effect of experimental result failed to achieve significance ($F = 3.57$ $p = 0.067$) [$r = 0.25$]. Neverthless, the difference was in the expected direction; higher quality ratings were given to those experiments achieving a significant result and the effect was larger than the effect in the overall analysis. In this analysis this effect was mediated by a significant interaction between the experimental result and the significance

level of that result (F = 6.07, p = 0.019) such that an experimental result of high significance was rated as having higher quality than an experimental result of low significance. These results can be seen graphically in Figure 2.1. In this interaction an experimental description which had a significant result at the $p < 0.01$ level attracted significantly higher quality ratings than all other combinations. There were no other significant differences. (See appendix two)

Figure 2.1

Interaction between experimental result and significance level

for subjects not receiving long-term outcome information




B. Analysis for subject group receiving experimental result and long-term outcome information

An Anova was performed on the data for those subjects who did receive long-term outcome (replication) information; this had four within subjects factors (subject numbers, pre-set significance level, initial result and long-term outcome).

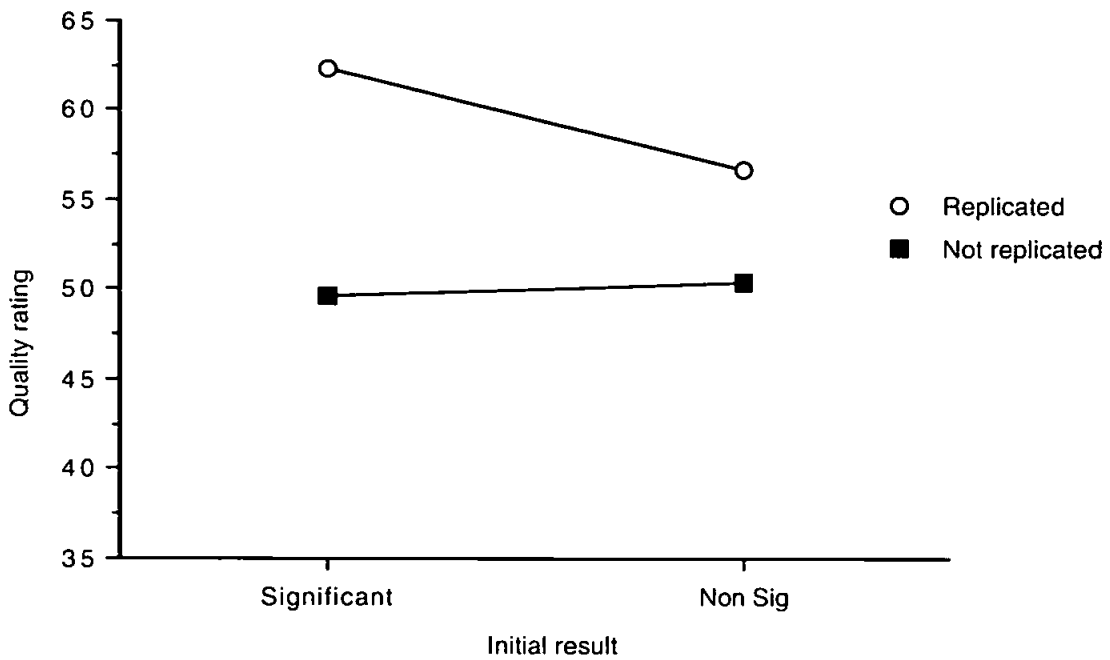
In this analysis (as in both previous analyses) both subject numbers (F = 21.6, p = 0.001) and pre-set significance level (F = 10.1, p = 0.003) had significant influences on

quality ratings. As before experimental descriptions with higher subject numbers were rated as having higher quality; as were those with smaller pre-set significance levels.

For those subjects given long term outcome information (n = 35) experimental result alone did not significantly influence quality ratings (F = 2.4, p = 0.13). However there was a significant main effect of long-term outcome (F = 18.7, p = 0.0001) [r = 0.60] with experimental results which had been replicated attracting higher quality ratings than those which were not replicated. The interaction between experimental result and long term outcome failed to achieve significance (F = 3.73 p = 0.061). These results can be seen graphically in Figure 2.2.

Figure 2.2

Interaction between experimental result and long-term outcome




For complete Anova tables and tables of means see Appendix 2.

### 2.3.4 Discussion

Overall the results showed a clear preference for designs with larger sample sizes; this was the largest effect of any of the variables. This effect appeared not to have been mediated by considerations of experimental cost. The only factor influencing cost in these scenarios was the number of subjects used and across all conditions the highest number of subjects attracted the highest quality ratings. In addition experimental descriptions with higher pre-set significance levels attracted higher quality ratings than those with lower pre-set significance levels in all conditions. This preference was not consistent with judgements based solely on maximising experimental power (see Section 2.3.2 for normative power analysis). However, this preference for higher significance levels may have reflected a desire to reduce the possibility of a false positive result which in this scenario would lead a drug company to waste a great deal of money on unnecessary clinical trials.

Both these findings may be seen as reasonable judgements. Firstly, no information on desired effect sizes was given; although it would be appropriate to assume that a drug would need to have a reasonably large effect on blood pressure to be clinically useful. Secondly, no information on the costs of running subjects was given and in this scenario it is reasonable to assume the potential financial benefits (or losses) to a drug company developing a new drug may be immense. In comparison the cost of running subjects may be seen as negligible.

Another possible interpretation of the effect of pre-set significance level is that there was little understanding of statistical principles and that quality judgements were based on a feeling that a result at the level of 0.01 was somehow better than one at 0.05. There is evidence for this latter view from the interaction between significance level and experimental result shown in Figure 2.1. This interaction demonstrated that a higher pre-set significance level only had a positive influence on quality ratings when it was known in hindsight that the experiment found a significant result. Thus this would seem to be more a hindsight bias effect (not only a significant result but a very significant result) than one due to considerations of statistical implications.

The prediction that the presence of outcome information would produce hindsight bias

(See Section 2.3.1) was upheld by the results. Overall, subjects were biased towards higher ratings of quality for those experiments finding a significant result. In the condition where subjects were also given long-term outcome information this initial experimental result failed to have a significant effect. In this condition this effect was replaced by an effect of long-term outcome information where replication of the experimental result was seen to significantly increase quality ratings. Notably, this was a much larger effect than that for initial experimental result (or pre-set significance level). This is consistent with the prediction that the presence of long-term outcome information would cue subjects to consider the possibility that the initial experimental result might be an error; and thus they would be less influenced by this initial information (See Section 2.3.1). However this does not explain the size of the subsequent bias towards sampling procedures which have been replicated. Before accepting an explanation of these results based purely on hindsight bias the possibility of subjects utilising outcome information as a valid predictor of original experimental power must be considered. There is some justification in the view that an experiment finding a significant result has at least a high probability of having been powerful enough. As the probability of Type 1 error is equal to the pre-set significance level then the probability of an experiment with a significant result having had sufficient power is inversely proportional to this value (i.e. the probability that the significant result is not a Type 1 error). Similarly, a non-significant result would imply a probability that the experiment lacked power which was inversely proportional to the probability of Type 2 error. If this were the case and subjects were treating experimental results as probabilistic evidence of sufficient power, then there should be evidence of an increasing tendency to favour significant results in those scenarios where sufficient experimental power is in question; as compared to those known to have high power. There was no evidence for this trend in the results. In addition the tendency to favour significant results remained present in the outcome condition; specifically the quality of a significant result subsequently proved to be true was rated much higher than that of a non-significant result subsequently proved to be true. In this case there is no logical reason to differentiate between these results.

It is worth noting that the replication of an experimental result increases the probability that this result reflects the true state of the world. This does not imply that any sampling

procedure finding the same result must have been a good procedure. For example, an experiment with very low power will have a high probability of producing a non-significant result (Type 2 error) which will always be replicated if no effect exists. Conversely, any experiment producing a Type 1 error will subsequently be replicated if a effect does exist. Therefore, this preference for significant results overall and towards replication in the long-term outcome condition would appear to be clear evidence of bias. This bias involves the consideration of evidence (experimental outcome) which would not have been available to the designer when the experiment was designed and to this extent it is a form of hindsight bias. In addition the presence and strength of this bias is all the more remarkable as the evidence on which it was based is also not relevant to the judgement being made.

Although these results appear to demonstrate failures in statistical reasoning and clear influences of hindsight bias, alternative hypotheses may exist. Firstly the definition of what the experimenters had meant by "quality of the experimental design" may not have been sufficiently clear to prevent subjects from including the quality of the outcome in their judgements. Secondly, it was possible that these effects were specific to the statistical nature of the task; either subjects may have attempted (and failed) to make specific statistical calculations, rather than intuitive judgements, or subjects may have been confused by the apparent statistical complexity of the task.

Despite these possible experimental problems it was apparent that at least in this case the evaluation of experimental designs was to a large extent biased by the presence of outcome information. In addition the size of this bias was dependent on the type of outcome information presented. Consequently, it was decided to run further experiments to explore the nature of these biases. In order to address the problems apparent in the present study these experiments would involve more detailed instructions clearly defining quality and using an equivalent task based on a betting scenario. In the proposed scenario for the next experiment all explicit references to statistical concepts will be removed as will any direct references to statistical significance levels in sampling descriptions. In addition the sixteen descriptions in the question booklet will be made shorter and more concise in order to make the task less confusing.

## 2.4 SUMMARY OF CHAPTER TWO

Chapter 2 presented two experiments which developed out of an interest in the cognitive processes involved in the evaluation of experimental designs. Evaluation of the quality of an experimental design is a necessary stage in the general process of experimental design. Following the creation of a putative experimental design some form of assessment must take place to ensure that at least the experiment is testing the desired hypothesis. Ideally this evaluation should include specific determination of the probabilities of Type 1 and Type 2 errors (in terms of choice of pre-set significance level and experimental power) and the balancing of these attributes against limitations of cost. Outside the design process it is also a commonly practised and necessary skill for students and researchers to evaluate the quality of the designs of others.

Experiment 1 demonstrated that during the design of psychology experiments only the minimum levels of evaluation were made. Subjects evaluated designs in terms of their appropriateness to test the hypothesis in question and checks were made for possible confounding variables and for the practicality of putative designs. Beyond this there was little evidence of formal evaluation of potential statistical validity. In general this led to the use of standardised significance levels and inappropriate choices of sample size (usually far too low for sufficient experimental power).

Given this failure of Experiment 1 to elicit formal evaluation of the statistical validity of experimental designs, it might be assumed that this form of evaluation rarely takes place. This view is supported by numerous reports in the literature of the failure of psychological research studies in general to utilise sufficient experimental power. (See Chapter 1, Section 1.8.5) Nevertheless, the ability to judge the quality of an experimental design accurately is clearly an important skill for both researchers and students. If this skill is not evident in experimental design then it should at least to some extent be present in the assessment of published studies. Experiment 2 was therefore designed to reflect the evaluation processes necessary in peer review or Ph.D. examination where assessment is made of the quality of descriptions of existing sampling procedures. The results of Experiment 2 again

demonstrated little understanding of statistical validity and showed judgements of the quality of an experimental design to be largely influenced in hindsight by outcome information. Given this large influence of hindsight information it was necessary to make a complete review of current research into hindsight bias. This review is the subject of Chapter 3.

## Table of contents - Chapter three

# CHAPTER THREE
## A Review of hindsight bias

## 3.1 INTRODUCTION

Experiments 1 and 2 (reported in Chapter 2) showed that subjects were reluctant to make formal evaluations of the statistical validity of sampling procedures; either in the design of their own research or after the event in assessing existing studies. In Experiment 1 verbal protocols recorded during the design process showed very little evidence of any formal evaluation of the statistical validity of putative designs. In Experiment 2 subjects were forced to make a judgement of the quality of an experimental description. In this case subjects generally failed to make appropriate use of sample size and significance level information. Their final judgements relied to a large extent on information given relating to the outcome of the experiment that they were judging.

Given this finding the present chapter reviews research on hindsight bias and includes a wider range of studies with findings related to the influences of hindsight information. Subsequent experiments presented in this thesis will be designed to extend the findings of these first two experiments in terms of hindsight bias in experimental evaluation. In addition these experiments will seek to extend knowledge of the influences of hindsight information in a new domain (judgements related to experimental designs) and using a new paradigm where outcome information has no direct relevance to the decision being made.

## 3.2 HINDSIGHT BIAS STUDIES

Hindsight bias was first reported by Fischhoff (1975). This seminal study is remembered for its description of the "knew-it-all-along" effect which has become synonymous with hindsight bias. However, it is worth looking at this study in some detail as it described a number of effects related to the presence of outcome information which have not gained the same popular attention as the "knew-it-all-along" effect. This early study also used a paradigm somewhat different from that which has now become the standard paradigm.

Fischhoff (1975) reported three experiments. The first experiment presented subjects with descriptions of real life events. These descriptions either contained no information on their outcome or they presented information describing one of four mutually exclusive outcomes as the true outcome. Subjects were asked to rate the probability of the occurrence of each of the four possible outcomes. Results showed higher likelihood ratings where the outcome rated was the one which had been presented as true, compared to the rated likelihood where no outcome was presented. This is the familiar hindsight bias effect where if subjects know a true outcome they perceive this outcome to have been more likely all along. Subjects in this study were also asked to rate the relevance of specific sentences in the event description. These relevance scores were also influenced by the stated outcome in a way which may suggest a post hoc rationalisation of the biased probability scores. Generally sentences which supported the given outcome were perceived as being more relevant to the required probability judgement.

Fischhoff's' second experiment was a replication of the first except that subjects given outcomes were asked to respond "as they would if they had not known the outcome" (Fischhoff, 1975 page 293). This instruction had no effect on hindsight bias; thus subjects were either unaware of the bias or, if aware they were unable to alter its effects. In a third experiment subjects were told that previous subjects had not been given outcome information, they were then required to respond as had these previous judges. This manipulation also had no effect on levels of hindsight bias. Overall these studies

demonstrated a number of effects associated with the presence of outcome information. Firstly, given outcome information judges will overestimate how likely this 'true' outcome would have appeared, either to themselves if they had not had this information or to others without this information. Secondly, judges are unaware of this effect or they are unable to control it. Finally, information supporting the given outcome will be seen as more relevant to the required judgement.

These findings were extended by Fischhoff and Beyth (1975) in a study where subjects were asked to predict in advance the probability of a number of possible outcomes of a real life event. After the event had taken place subjects were asked to recall their original judgements; subjects remembered having given higher probabilities for those outcomes they believed to have occurred. Thus the effect of subjects overestimating what they would have known without outcome information can be extended to include an overestimation of what they did know before the outcome. Notably this study also demonstrated that hindsight bias also works in the opposite direction; lower probabilities were remembered for those outcomes believed not to have occurred. Task definition is another important factor in this study. Whereas in the previous studies it was possible that subjects had interpreted the instructions to include current estimates of outcome likelihood, in this study subjects were clearly presented with a memory task in which they were required to recall their own original judgements. It is hard to imagine that they could have misunderstood this task.

Since these early studies there have been a whole range of studies demonstrating consistently replicable effects of hindsight over a varied range of tasks and conditions. The 'memory' paradigm used by Fischhoff and Beyth (1975) has now largely become the standard paradigm: here subjects rate the probabilities of an outcome in foresight and at a later time attempt to reproduce these probabilities in hindsight. Studies have utilised subjects' judgements of the probability of the occurrence of a wide range of real life and artificial events. The presentation of hindsight information and subjects' subsequent attempts to reproduce their earlier judgements have followed after intervals varying from one day to several months. Hawkins and Hastie (1990) present a review of these studies. Tasks utilising this memory paradigm include the use of real life events such as Pennington's (1981) study before and after the Fireman's strike of 1977, Leary's (1982) study of

hindsight distortion in the 1980 U.S. presidential election and Powell's (1988) study using the 1984 U.S. presidential elections, to name but a few. In all these cases subjects had predicted the probability of various outcomes before the event. After the event subjects consistently overestimated the probability they had assigned to the actual outcome.

Another common task involves the use of almanac information questions in the memory paradigm . Here subjects rate the probability that their answer is correct before and after receiving the actual correct answer. With numerous minor task variations this paradigm demonstrates a robust hindsight bias (Fischhoff, 1977. Wood, 1978. Hasher et al. 1981. Hoch and Lowenstein, 1989.). In addition to these common tasks hindsight bias in memory of previous judgements has also been demonstrated in a number of more unusual areas such as; personal history (Ross and Conway, 1986) and the outcome of pregnancy tests (Pennington et al. 1980).

The findings listed above are based on two different paradigms. In the original (on-line) paradigm, subjects were presented with a 'true' outcome as a section of the original text and asked to estimate the likelihood of this outcome as if they had not received this information. Subjects consistently rated the given outcome as more likely when compared to control groups with no outcome information. In the second (memory) paradigm subjects are required to judge the likelihood of an outcome with no knowledge of what happened. At some point in the future they are then told the 'true' outcome and asked to recall their original judgement. It should be noted that although both these paradigms lead to what is commonly termed hindsight bias these are essentially two different effects. One is a bias of likelihood judgement the other is a bias of the memory of a likelihood judgement. The original paradigm which leads to a direct bias of judgement has been described as an 'on-line' judgement task by Hastie and Park (1986) who argue for a distinction to be made between these tasks and memory based tasks as there is no clear evidence for any commonality in the causal relations underlying memory and judgement.

Whilst the memory paradigm has the advantage of clarifying the task instructions and thus avoiding the need to ask subjects to ignore information already presented it is quite possible that the resulting bias results from different causal processes. Certainly it would be difficult to argue for the same motivational basis for these biases. In the case of on-line

judgements the effect could be due to a motivation for subjects to include outcome knowledge in an attempt to appear intelligent or knowledgeable. If a similar motivation was present in the memory paradigm it would be mediated by a motivation to accurately remember the original judgement. This would at least reduce the apparent size of any bias. In addition, in the memory paradigm the effect of hindsight bias may be significantly reduced by the subjects' ability to remember the original judgement. Both these mediating effects have been demonstrated by Hell et al. (1988) who explored the influences of motivation and memorability over time in memory based almanac type hindsight tasks. Although the results of this study ruled out purely motivational effects [as did an earlier study by Fischhoff (1977)] they did demonstrate that motivation to recall correctly reduced hindsight bias. By varying the factors influencing memorability of the original response they demonstrated a reduction of hindsight bias from 22% in a weak memory condition to only 3% where the memory trace is strong and well motivated.

Hindsight bias has important and potentially very damaging effects on judgement depending upon the situation in which it arises. The focus on studies using the memory paradigm has led a number of researchers to underestimate both the size of the bias and its potentially damaging effects. In this paradigm the bias only affects the memory of an earlier judgement. This serves to make subjects overconfident in their own judgements and may limit the amount they learn from previous experience. However, in the on-line paradigm it is the judgement itself that is influenced by the bias and this may have serious consequences in a number of real life situations where outcome information is present. In addition to being of more practical importance the on-line paradigm would be expected to lead to larger biases as it is not subject to the mediating effects of memory and motivation to remember outlined above. The remaining sections of this chapter are concerned with hindsight based studies in areas which have considerable practical importance.

### 3.2.1 The inability to ignore relevant information

In Fischhoff's (1975) study subjects' perceived likelihood of an outcome was biased by the knowledge of the 'true' outcome presented on-line as part of the task description. This effect can be seen as an inability to ignore relevant information once it has been presented and includes the fact that subjects are unaware of having been influenced by this information. As suggested by Fischhoff (1977), this effect has real life implications particularly in terms of dealing with inadmissible evidence in the courtroom and is not reduced by telling subjects about the bias or exhorting them to try harder. A large number of studies have explored this inability to ignore inadmissible evidence; Wyer and Budesheim (1987), Wyer and Unversagt (1985), Schul and Burnstein (1985), Carette and Moreland (1983), Werner et al. (1982), Tanford and Penrod (1982), Thompson et al. (1981), Horowitz et al (1980), Wolf and Montgomery (1977), Hans and Doob (1976), Hoiberg and Stires (1973), Sue et al. (1973), Doob and Kirshenbaum (1972). All these studies have demonstrated consistent influences of information which subjects have been asked to ignore. It should be noted that these effects relate to biases in judgements of the guilt of a third party rather than biases in likelihood judgements so far reported in the hindsight literature. In this respect these studies have more in common with 'outcome' bias studies and will be discussed in section 3.3 on outcome bias.

### 3.2.2 Reverse hindsight effects

A number of studies have demonstrated a reduction or possibly a reversal of the hindsight effect under conditions where the given outcome is perceived as very unlikely or surprising. Slovic and Fischhoff (1977) introduced the notion of surprisingness in hindsight using a number of tasks based on descriptions of experiments each of which had two possible outcomes. Subjects were asked to rate the probability of each of the two possible outcomes being replicated. Hindsight bias was found in all conditions but hindsight effects were smaller for those results which subjects rated as more surprising. However, none of the outcome results in this series of experiments was particularly unlikely leading to only small differences in subjects' ratings of surprisingness.

In a more extreme real life case Verplanken and Pieters (1988) in a study of attitudes to nuclear power plants found that after the Chernobyl disaster subjects recalled their earlier judgements of the probability of a disaster as lower than they had been before the disaster. This result is in direct opposition to the usual hindsight effect. However, it should be noted that the result was in line with the official viewpoint consistently repeated in the media that the disaster was an extremely unlikely event resulting from a series of improbable coincidences and that "nobody could have known it was going to happen".

A study by Mazursky and Ofir (1990) also claims a reversal of hindsight bias using a somewhat different procedure to previous studies. The experiments in this study used expected and post-exposure quality judgements as the dependent measure and manipulated the quality of various stimulus materials. Materials used were good and bad educational films, plastic suction hooks and graphics software packages. In each case the materials were rated for quality after subjects were given practical experience of them. The results showed that where subjects expected low quality materials and subsequently found high quality materials their estimate of what their original quality judgement might have been was divergent from their actual quality judgement. However, the results from this study are confounded by a number of factors. Mark and Mellor (1990) point to a possible contrast effect and also suggest effects may be due to the desire to rate one product as superior to another. It is difficult to compare this study with normal hinsdsight studies as subjects in these experiments did not make an original estimate of quality before they were exposed to the products. The comparison made was between an actual quality rating and an estimate of what an initial quality rating might have been. Both these ratings were made post-exposure.

Both the previous studies claiming reversals of hindsight bias have introduced a new factor in terms of positive or negative outcome information. In earlier studies outcome information simply referred to whether a particular result had happened or a particular answer to a question was right and the required judgements were based on likelihood. These judgements have little or no personal importance to the subjects and the questions or scenarios they were based on were generally neither particularly negative or positive from the point of view of the subjects. In exception to this were the studies related to political elections where presumably some of the subjects had strong preferences for one party or the

other. This would lead to individual beliefs of particular outcomes being highly negative or positive. Surprisingly this factor of personal preference does not appear to have been included in the analysis of any of these studies. Considering this factor it seems reasonable that a value laden outcome would influence the resulting hindsight bias. For example, in the Chernobyl scenario the event itself has very strong negative connotations. A subject following the normal hindsight pattern would suggest they knew the disaster was going to happen all along. In this case this would leave them in the uncomfortable position of having known and done nothing to prevent the disaster or warn others. Also a judgement that this event could have been foreseen by others is tantamount to an accusation of negligence of all parties concerned. This is clearly not comparable to the judgement facing subjects in the standard hindsight paradigm.

The Mazursky and Ofir (1990) study described earlier in this section also included value laden outcomes although of a somewhat different form. In this case products were shown to be either good or bad in practical demonstrations. It is important to note that the results showed differential biases for these positive and negative outcomes although the authors do not discuss these. In fact the 'reverse' hindsight effect was only demonstrated where initial expectations were negative and the subsequent outcome was unexpectedly positive. The methodology of this study has a great deal in common with what have come to be known as 'outcome bias' studies and these results are not unusual when looked at from an outcome bias point of view where the subjective value of the outcome is the major determinant of the resulting bias.

These effects also have a great deal in common with belief bias effects. Possible theoretical links between outcome information and belief bias will be discussed at length in Chapter 6.

## 3.3 OUTCOME BIAS STUDIES

It is difficult to make a clear distinction between hindsight bias studies and outcome bias studies and there is considerable confusion and overlap within the literature. Outcome bias studies generally involve the use of on-line paradigms where outcome information is presented along with information on which the subsequent judgement is to be made. In addition, outcome bias studies utilise judgements other than outcome probabilities. Nevertheless, these two factors alone do not prove that an outcome bias effect is fundamentally different to the hindsight bias effect in terms of underlying processes. In a number of cases (e.g. Mazursky and Ofir (1990) cited earlier) studies describing effects as hindsight bias fall within these criteria. Within this section the terms outcome bias and hindsight bias follow the relevant authors' usage, possible definitions and differences between these two biases are discussed later.

There are methodological problems within both the hindsight bias paradigms discussed earlier. In the on-line paradigm there are a number of problems which relate to the fact that hindsight information presented as the outcome of a set of events or questions is relevant and useful information. As Hawkins and Hastie (1990) point out, where outcome information is relevant to the original judgement it is a rational response to change beliefs in the light of this new information. In addition it would not be surprising if subjects were confused in paradigms where relevant outcome information is presented by the experimenter along with the instruction to ignore this same information. If this information is to be ignored there is little justification for the experimenter presenting it. Outcome bias studies eliminate these methodological problems. They differ in that there are generally much clearer logical reasons for ignoring the biasing information or alternatively this information is not directly relevant to the decision required.

There are a number of studies where there are clear reasons for subjects to ignore information which has been presented to them, for example, the courtroom based inadmissible evidence scenarios described earlier. The results of these studies are generally described in terms of hindsight bias, despite the fact that the scenarios involve on-line

presentation of information to be disregarded and the required judgement is not one of likelihood. In these respects these studies have a great deal in common with outcome studies. The inadmissible evidence in these cases may be true or false and has been purposely introduced in an attempt to bias the jurors (subjects) who are aware that they should try to avoid this bias. In these tasks subjects have clear logical reasons for ignoring this information and yet this information still influences judgements of the guilt of a third party. Whilst these studies show the inability of subjects to ignore any information deemed as relevant to the required decision, the bias in these cases does not specifically relate to outcome information.

In a study relating to actual outcome information Lipshitz (1989) has shown that judgements of the correctness of the actions of others are strongly biased by the outcomes of these actions. This study presented subjects with descriptions of the actions of officers of the Israeli defence forces in which the officer made a decision either to obey or disobey orders. The subsequent description of events included outcome information in terms of either success or failure of the mission in question. Subjects evaluating the decision made by the officer were influenced more by the eventual success or failure of actions based on the decision than by the decision itself. In this case it is clear that the decision maker could not know the final outcome at the time the decision had to be made and thus in order to be impartial subjects should ignore this information. In judging the quality of the decision by its outcome subjects overestimate the knowledge of the decision maker. In this study there was some possible justification for the bias shown by subjects as it was clear from the experimental descriptions of events that the officer in question had a great deal more information at his disposal than was presented to the subjects. Under these conditions it was possible that the officer was in a better position to judge the outcome of his decision and that subjects' judgements reflect this belief. Therefore, it could still be argued that hindsight bias is simply an inability to ignore relevant information even though the effect shows a gross exaggeration of the importance of this information relative to any other relevant factors.

There are however a small number of studies which have demonstrated that outcome information which has no logical relevance to the judgement task will still influence judgements. Mitchell and Kalb (1981) have shown that a supervisors' evaluation of a

nurses' performance is influenced by the outcome of the nurses' actions. In this study exactly the same action by a nurse was presented as either having a benign outcome (no harm to the patient) or a negative outcome in which the patient was harmed. Results showed that subjects with outcome knowledge rated given outcomes as more probable in all cases. In the case of negative outcomes supervisors not only evaluated the nurses' performance as worse but also held the nurse more responsible for the behaviour. Baron and Hershey (1988) reported five studies in which subjects rated the quality of decisions made by others. The rated decisions were based on either medical matters or monetary gambles. In each case subjects rated the decision maker as more competent and their thinking as better when the outcome of the decision was favourable, despite the fact that when asked, subjects felt that they should not consider the outcome when making these evaluations.

In both the above studies the outcome information was not relevant to the judgement being made yet still had a significant influence on that judgement. Baron and Hershey (1988) refer to this direct effect of outcome information on the evaluation of decisions as "outcome bias" (page 569) and see it as a separate effect working in addition to hindsight bias; they suggest that this effect "does not operate on a judges assessed probabilities of outcomes" (page 570) and as such it is distinct from hindsight bias. However, as we have seen from Fischhoff's (1975) studies, hindsight effects include more than a simple exaggeration of assessed likelihood. In these studies outcome information also affected the perceived relevance of factors supporting the given outcome; thus even from this early stage of the research there is evidence of a more pervasive effect in which subjects bias more than simple estimates of likelihood.

In all probability "outcome bias" is simply hindsight bias seen from another point of view. In outcome bias the subjects' biased viewpoint is attributed as being available to the decision maker before the outcome is known. Thus the issue is whether the decision maker could have reasonably been expected to predict the given outcome beforehand. Surely this is exactly the effect seen in standard hindsight paradigms; an overestimation, given the outcome, of the prior predictability of that outcome. This overestimation of prior predictability remains the same whether it is reflected in the belief that your own prior estimate *was* different, or the belief that another persons' prior estimate *should have been*

different; an "I knew it all along and they should have known it all along too' effect. A very similar effect has already been described by Fischhoff (1975) in the condition where subjects were asked to respond in the way they expected previous subjects who had not been given outcome information would have responded. In this condition the hindsight bias had the effect of overestimating the knowledge attributed to others in exactly the same way as is seen in these outcome bias studies.

The differential influence of positive and negative outcomes in these studies is also mirrored in early hindsight studies. As described earlier Fischhoff and Beyth (1975) demonstrated that hindsight bias also works in the opposite direction; lower probabilities were remembered for those outcomes believed not to have occurred.

### 3.3.1 Outcome information as feedback

The concern that any form of outcome information may supply useful feedback on some aspect of the given problem was explored by Hoch and Lowenstein (1989) who analysed the extent to which outcome feedback is information in its own right. They suggest "people can extract diagnostic information from feedback despite overestimating what they would have known in foresight" (page 605). In these studies feedback was shown to facilitate accurate performance on a series of difficult trivia questions (no effect was found for medium or easy questions) by enabling subjects to assess general item difficulty. However this facilitation of the accuracy of judged probabilities of correct or incorrect answers did not preclude the additional presence of hindsight bias. Also this facilitation was not present in experiments using insight problems where the strength of hindsight bias effects overwhelmed any effect of information feedback.

It should be noted that this feedback facilitation effect can only occur where outcomes are given for a consistent series of repeated questions of a particular type and difficulty. In outcome bias studies where a single judgement is made on the basis of the outcome of one set of events this feedback effect would not be possible. Nevertheless, it is interesting to note the possibility that gross overestimation of the diagnostic value of a single outcome could be a contributory factor in outcome bias effects. A judgement which is unreasonable when based on a single outcome would be perfectly reasonable if based on a large number of

similar outcomes; this could be seen as an extreme case of subjects failing to apply the law of large numbers.

## 3.4 PRACTICAL RELEVANCE OF HINDSIGHT BIAS EFFECTS

Recently Christensen-Szalanski and Willham (1991) have called into question the practical significance of hindsight bias. In a meta-analysis of hindsight studies, mostly using the standard paradigms, they report a consistent but small (r = .17) effect of hindsight bias. (See Appendix 7 for a definition of the measure of effect sizes used in this study and also used in the present experimental results.) They also argue that 'the use of "almanac" questions can generate an unusually large hindsight effect' (page 147). This leads to the impression that in real life only a small proportion of decisions are influenced by hindsight bias. However, this meta-analysis does not consider the more ecologically valid outcome bias studies as a separate case from studies using the standard hindsight paradigm, although in these cases outcome bias persists in evaluations of others' decisions even when the outcome information is not relevant to the original decision. These effects may be a great deal more pervasive and of a much greater practical significance than the effects seen in the standard hindsight paradigm and may have been lost in the process of averaging across a large number of different studies. Unfortunately, there are no direct measures of effect sizes in any of these reported outcome bias studies and thus they cannot be directly compared with the effects related to the standard hindsight paradigm. However, there is evidence that motivation and memorability severely limit the size of hindsight effects in the memory paradigm (Hell et al., 1988, as reported in section 3.2), therefore it is likely that effect sizes would be significantly larger in outcome bias paradigms. Indeed in all the outcome bias studies described above, outcome information had a large enough effect for it to be the major determinant of subjects' final judgements.

At a practical level the nature of the outcome alone made the difference between promotion or disciplinary action for officers in the Lipshitz (1989) study. In addition, it seems probable that the perceived importance or utility of a particular outcome may increase

the size of any resulting bias. This is suggested in the Baron and Hershey (1988) study where subjects strongly defended the importance of outcomes in a condition where a doctors' decision may result in the death of the patient. However, analysis of outcome salience has not been reported in the literature so far.

## 3.5 RELATED EFFECTS

Effects relating to the presence of outcome information have a number of parallels in other research areas; these will be briefly enumerated here but only discussed in full at a later stage in the theoretical discussion of the results of the subsequent experiments.

Research relating to overconfidence where subjects consistently rate their own judgements to be better than they actually are has obvious similarities to hindsight effects where subjects consistently rate their own judgements to be better than they actually were. In both these cases the biases serve to support a view of the world in which the subject performs better than is actually the case. This view is also supported by reconstructive remembering where subjects alter their memories of past events to fit in with their current beliefs. This can be seen as a form of post-hoc rationalisation in the same way as subjects attribute higher relevance to those factors supporting known outcomes in hindsight. Another effect which serves to support a biased view of the world is the Halo effect. In this effect subjects attribute greater skill or intelligence to people they rate as more physically attractive. The Halo effect can also be seen as an inability to ignore any form of irrelevant information which parallels the inability to ignore any form of inadmissible evidence seen in courtroom based hindsight studies.

## 3.6 SUMMARY OF CHAPTER THREE

Chapter 3 reviewed literature related to hindsight and outcome biases. This review showed that effects associated with the presence of hindsight information go beyond the commonly recognised 'knew it all along' effect. This well known effect has most often been demonstrated in a memory paradigm. In this paradigm subjects' memories of likelihood judgements made in foresight are shown to be biased in hindsight after the actual outcome is known. This hindsight bias leads subjects to overestimate the likelihood they originally assigned to what they now know to be the true outcome (and to underestimate the likelihood they originally assigned to any false outcomes).

Other related effects which have been discussed include overestimation (in hindsight) of judgements of the relevance of information supporting a given outcome. The persistence of these effects is shown by subjects' inability to ignore relevant outcome information, despite direct instruction and strong motivation. Reverse hindsight effects have also been discussed. These seemingly paradoxical effects occur in certain specific cases where, in hindsight, subjects underestimate their earlier likelihood estimates. Comparisons were drawn between these effects and the effects seen in outcome bias studies and an explanation of these effects in terms of subjects' motivation was suggested.

In the second part of Chapter 3 outcome bias studies were reviewed. Effects described as outcome bias were shown to generally relate to value judgements of the actions of others, made after the outcome of these actions was known. A number of direct parallels were drawn between these effects and hindsight bias effects and it was suggested that hindsight bias might be a specific case of a more general and all inclusive outcome bias. Some important differences between the paradigms used in hindsight and outcome bias studies were also noted. It was shown that the more ecologically valid paradigms used in outcome bias studies generally resulted in larger biases than those reported in hindsight paradigms. In addition, the size of these effects and whether they were positive or negative was shown to depend to some degree on the specific nature of the outcome information provided.

## Table of contents - Chapter four

# CHAPTER FOUR

## Studies showing the effects of outcome information on judgements of the quality of sampling procedures

### 4.1 INTRODUCTION

The two initial studies reported in Chapter 2 demonstrated a number of practical problems associated with designing Psychology experiments. In the first study a pervasive lack of critical evaluation of putative experimental designs was demonstrated. The second study confirmed that failures of statistical evaluation were not limited to the design process but were also present when subjects were specifically directed to evaluate existing designs. This second finding ruled out the possibility that deficiencies in the design process were due to some form of satisficing or errors of oversight and implied a general inability to accurately evaluate as the causal factor. In addition, Experiment 2 demonstrated that judgements of the quality of a sampling procedure are strongly biased towards favourable outcome information.

In the light of these results Chapter 3 described hindsight bias and focused on a number of effects specifically associated with the presence of outcome information. This research implied the possibility that using salient outcome information in an on-line paradigm would result in biases consistently larger than the average reported hindsight bias. The practical importance of this potential bias in the evaluation of existing experimental designs lies in such areas as the peer review of journal articles or the examination of Ph.D. theses which may be largely influenced by the presence of positive or negative experimental outcomes.

The experiments presented in this chapter were designed to explore the extent to which different experimental outcomes influence judgements of the quality of experimental design. In addition to the practical importance of making reviewers and examiners aware of potential sources of bias, the use of experimental results as outcome information also creates a particularly useful paradigm. In previous outcome bias studies subjects judged the actions or decisions of a third party after the event when the outcome was known. In these previous studies subjects should ignore the outcome information as it was not available to the third

party in question at the time the action or decision was made. Thus in these cases biased judgements based on outcome information imply *'they* should have known it all along'. However, in these experiments the nature of the outcome was inevitably relevant to the judgement being made and this relevance weakens the argument that the inclusion of outcome information is a true bias. In all these experiments subjects were given a brief description of a previous event clearly containing less information than would be available to the third party acting in the real world. Thus it could be argued that it was reasonable for subjects to assimilate this information in any judgement of the actions of a third party where they could assume that the third party was privy to more information. Thus the effect becomes 'they should have known it all along because they had more information than I have'.

The use of an experimental design paradigm where the outcome information is the subsequent experimental result precludes this more reasonable explanation of outcome bias for two reasons. Firstly, in an experiment, the result (initial outcome) may be significant or non-significant and either of these results may truly reflect the state of the world. Thus the statistical result alone has no relevance to the quality of the experimental design; a well designed experiment might have either result depending on whether an effect exists in the world or not. The truth or otherwise of this initial result can never be guaranteed but the probability of it reflecting the true state of the world can be refined by replication. Thus replication suggests four possible "long-term outcomes"; true significant or non-significant results or false significant or non-significant results. Secondly, the probability of each of the false results (Type 1 and Type 2 errors) is defined by the design factors within the experiment. Thus a subject who is aware of these factors can be certain that they have exactly the same information as that available to the original designer of the experiment. Therefore the present paradigm utilises outcome information irrelevant to the judgement being made and in addition ensures all possible relevant information is available to the person making the judgement.

Where long term outcomes (replication information) are available the judge has more information than the original designer and thus it cannot be argued that the designer should have 'known it all along'. However, this information is also not directly relevant to the

initial design quality. Having accepted predetermined probabilities of Type 1 and Type 2 errors the designer has specified the likelihood of a false outcome; it would therefore be unreasonable to treat the chance occurrence of a false outcome as a design fault.

The following studies presented subjects with descriptions of sampling procedures including either initial outcome information alone or both initial outcome and long term outcome information. The inclusion of long-term outcome (replication) information enabled an exploration of the relative influences of positive and negative outcomes which was unavailable to previous outcome bias studies as the initial outcome can either be confirmed or disconfirmed by subsequent replication information.

## 4.2 EXPERIMENT THREE

### 4.2.1 Introduction

This experiment was designed to further explore the effects of outcome information on judgements of the quality of experimental sampling procedures first demonstrated in Experiment 2.

In the design of this experiment a number of potential problems first identified in the discussion of Experiment 2 (Section 2.3.4) were addressed. Firstly the definition of what the experimenters had meant by "quality of the experimental design" had not been sufficiently clear to prevent subjects from including the quality of the outcome in their judgements. Secondly it was possible that these effects were specific to the statistical nature of the task; either subjects may have attempted (and failed) to make specific statistical calculations rather than intuitive judgements, or subjects may have been put off by the apparent statistical complexity of the task. Thirdly it was possible that the specific research methods or statistics teaching which these students had received had in some way influenced the results and that they may have chosen a specific strategy of error minimisation regardless of cost.

It was decided therefore to run an experiment with more detailed instructions, clearly defining quality. The definition presented to the subjects emphasised that a sampling

procedure of high quality was one which had a high probability of determining a true result without incurring excessive cost (see appendix three). In order to prevent the potential problems associated with the statistical nature of the task this experiment utilised a task based on a betting scenario in which all explicit references to statistical concepts were removed. Nevertheless, the sampling procedure described in this betting scenario was a direct equivalent of the experiments described in the previous study with the exception of the presentation of percentage differences instead of pre-set significance levels. Direct reference to statistical significance levels was therefore removed from sampling descriptions. The effects of this variation will be addressed in the discussion section. In addition the sixteen descriptions in the question booklet were made shorter and more concise in order to make the task less confusing and more emphasis was put on the cost of large samples to cue the need for cost-benefit judgements. The need for consideration of cost was mentioned in both the scenario and formally stated in the definition of quality.

### 4.2.2 Method

Subjects and task

Subjects were 90 first year psychology undergraduates at the university of Plymouth. The experiment took place at the beginning of their first year before any research methods or statistics lectures had been completed.

Materials and design

Materials comprised an instruction sheet and a question booklet. The instructions contained an introduction followed by a scenario description (See Appendix 3). Having read the instructions subjects were presented with a booklet containing sixteen short descriptions of experiments. For half the subjects these descriptions were followed by long term outcome information. Subjects were required to make a judgement of the quality of the sampling procedure described in each case.

Scenario and instructions were as follows;

**Introduction**

The following description and series of questions are designed to study understanding of the concept of quality in experimental sampling.

Your co-operation in this experiment is greatly appreciated, and the results should be of use to you as they will demonstrate general levels of understanding of sampling principles. Results and explanations will be made available as soon as possible.

You do not need to put your name on the answer sheets but please print your name on this instruction sheet and hand it back separately at the end.

**Scenario**

A worker in a roulette wheel factory had been bribed to produce a number of wheels in which the slots for even numbers were slightly larger than those for odd numbers.

It was the intention of a well financed team of fraudsters to bet heavily on even numbers on the biased wheels. Half the casinos ordering new wheels had received the biased wheels. However due to security measures it was impossible for the team to find out which of these casinos received true or biased wheels.

It was decided that each member of the team would go to a different casino and make a number of small bets whilst counting the number of even wins to see whether the new wheel was one of the biased ones. Simply standing at the table without betting was regarded as too suspicious as casino security consistently video and study behaviour at roulette tables. If it was decided that the wheel was biased towards even numbers then the rest of the team would return to make a great number of large bets on even numbers.

Tests involving extensive counting of even wins can be very expensive to run and dangerous in terms of alerting the casino to a specific interest in the roulette wheel. The potential cost of the larger bets is high.

**Test bets**

In initial tests sampling methods varied between individuals. Different numbers of spins were counted and there were also differences in the percentage of even number wins which were regarded as sufficient evidence of a biased table. On an unbiased table even numbers have a 50% probability of winning if zero wins are disregarded; biased tables have a greater probability of even wins. Below are a series of descriptions of the tests used by different team members.

**Task**

Your task is to read each description and give an intuitive estimate of the quality of the test for a biased roulette wheel by writing a percentage mark (from 0 to 100) in the space provided.

**A test of high quality is one which has a high probability of differentiating whether a table is truly biased or not without incurring excessive cost.**

Please try to use the whole range of marks.

A typical example of a sampling procedure description for the group not given long term outcome information is shown below.

> In the Monte Carlo Grand casino 300 spins of the roulette wheel were counted. The team member involved decided to regard anything higher than 56% even wins as sufficient evidence of a biased table. A bias to even numbers was found.

Descriptions of sampling procedures remained constant across the different questions with only the casino name, the number of spins (300 or 700) and the percentage even wins (54% or 56%) varying in each case. For subjects receiving long term outcome information each description was followed by a description of an outcome.

For complete details of all questions see Appendix 3.

Design and procedures

The experimental design and procedures were roughly equivalent to the previous experiment. Two levels of sample size were presented with either 300 or 700 spins of the roulette wheel being counted. Two levels of even number wins regarded as significant evidence of bias were presented (56% or 54%). The use of percentage values in this case meant that the calculation of equivalent significance levels was now interactive with sample size.

As in experiment two subjects were presented with either one or two levels of outcome information. In this case the initial result was whether a bias to even numbers was found or not. The long term outcome information again reflected whether the initial sampling result was subsequently shown to be correct or subsequently shown to be false. As before the dependent variable was the subjects' intuitive judgement of the quality of each test; however this was now defined clearly in the instructions.

Thus three experimental factors (each with two levels) were varied on a within subjects basis. These factors were: number of spins (700 or 300), percentage even wins regarded as significant evidence (54 or 56) and initial result (bias found or no bias found). The

counterbalancing of these four factors within a standard experimental description created sixteen different sampling procedures presented to the subjects.

As in experiment two the presence of long-term outcome information was manipulated as a between subjects factor (45 subjects receiving long term outcome information and 45 receiving only the initial result). For those subjects receiving long-term outcome the interaction of long term outcome information with the initial experimental result created four different possible overall outcomes. These long-term outcomes were presented as follows:

> **Outcome :** Subsequent large bets confirmed this wheel to be biased. The wheel has made large profits for the team.
>
> **Outcome :** Subsequent betting by customers confirmed this finding. The table was not biased.
>
> **Outcome :** Subsequent large bets by the team however showed no evidence of bias. A significant amount of money was lost.
>
> **Outcome :** The team did not go ahead with large scale bets. However subsequent betting by customers did show a bias towards even numbers and the wheel was eventually replaced.

As in experiment 2 it was possible to utilise an overall five factor analysis in which the presence of long-term outcome information was a between subjects factor. This meant that the nature of long-term outcome (experimental result subsequently shown to be correct or subsequently shown to be false) was analysed as a within subjects factor but was a mock factor for the group of subjects not receiving this information.

A normative analysis of the probability of the results of these sampling procedures to be due to Type one error gave the following results:

300 spins with 54% even wins - $p = 0.836$

300 spins with 56% even wins - $p = 0.019$

700 spins with 54% even wins - $p = 0.017$

700 spins with 56% even wins - $p = 0.0007$

The sampling procedures used in the experimental tasks were designed specifically to produce this range of normative results. The sampling procedure with 300 spins and 54% wins has an unacceptably large probability of Type 1 error; thus this finding should not be relied on as evidence of a biased wheel. Both the sampling procedures with 300 spins and

56% wins, and those with 700 spins and 54% wins, have acceptable probabilities of Type 1 error; either of these results constitutes acceptable evidence of a biased wheel. The sampling procedure with 700 spins and 56% wins has a very low probability of Type 1 error and thus is acceptable evidence of a biased wheel. However, it should also be remembered that procedures using 700 spins have higher sampling costs than those using 300 spins. In addition procedures with lower sample sizes and/or higher percentage differences will result in higher probabilities of Type 2 error. As the size of the bias in biased wheels (the effect size) is not defined in this model it is not possible to calculate exact values for the probability of Type 2 errors.

### 4.2.3 Results

A number of subjects handed in incomplete question booklets and these could not be included in the analysis; this left 73 subjects overall, 29 in the long term outcome condition and 44 in the no long term outcome condition. As in the previous experiment an overall five factor Anova was performed.

#### 4.2.3.1 Influences of all factors

A five factor Anova was performed with four within subjects factors (subject numbers, pre-set significance level, initial result and long term outcome) the fifth factor the presence or absence of long-term outcome information being between subjects. As in the previous experiment effect sizes for main effects were calculated following the recommendations of Cohen (1988) and are quoted in square brackets.

In this overall analysis both number of spins ($F = 24.4$, $p = 0.0001$) [$r = 0.22$] and percentage even wins ($F = 29.4$, $p = 0.0001$) [$r = 0.22$] had significant influences on quality ratings. Sampling procedures with higher sample sizes were rated as having higher quality; as were those with higher percentage wins.

In this overall analysis the experimental result had no significant effect on quality ratings ($F = 0.6$, $p = 0.8$). However, there was a strong effect of long-term outcome information ($F = 25.5$, $p = 0.0001$) and a significant interaction between the initial result and long-term outcome ($F = 12.5$, $p = 0.0007$). As in Experiment 2 the influence of these

effects was masked by the fact that in this overall analysis the nature of long-term outcome information was a mock factor for one subject group. In order to gain a clearer picture of the influences of these factors it was necessary to perform separate analyses of the two different subject groups (those receiving and those not receiving long-term outcome information). Means for all the factors in the overall analysis are shown in Table 4.1. The complete Anova table can be seen in Appendix 3.

Table 4.1

Means table for overall analysis - Experiment 3

(Standard deviations in brackets)

| | | | 300 Spins | | 700 Spins | |
|---|---|---|---|---|---|---|
| | | | 56% difference | 54% difference | 56% difference | 54% difference |
| No outcome information given | True Result (Mock factor) | Bias found | **46.9** (20.7) | **43.4** (21.4) | **57.0** (22.8) | **53.9** (23.4) |
| | | No bias found | **43.3** (16.9) | **40.1** (16.5) | **52.4** (21.6) | **49.7** (21.2) |
| | False Result (Mock factor) | Bias found | **46.0** (19.7) | **42.7** (21.4) | **59.3** (23.8) | **54.2** (22.1) |
| | | No bias found | **47.1** (17.4) | **39.3** (17.6) | **53.0** (22.0) | **51.9** (21.0) |
| Outcome information given | True Result | Bias found | **67.2** (16.4) | **57.5** (21.0) | **69.1** (17.0) | **65.3** (18.5) |
| | | No bias found | **62.4** (16.0) | **53.1** (15.5) | **70.1** (15.1) | **61.9** (20.7) |
| | False Result | Bias found | **47.0** (21.0) | **40.1** (23.7) | **50.8** (26.1) | **41.0** (22.1) |
| | | No bias found | **55.1** (18.1) | **45.2** (18.5) | **60.5** (16.0) | **51.2** (19.2) |

### 4.2.3.2 Analysis for subject group receiving experimental result only

An Anova was performed on the data for those subjects who did not receive long-term outcome information (N = 44) this had three within subjects factors (number of spins, percentage even wins and initial result).

In this analysis (as in the overall analysis) both number of spins ($F = 23.3$, $p = 0.0001$) and percentage even wins ($F = 12.0$, $p = 0.001$) had significant influences on quality ratings. Sampling procedures with higher sample sizes were rated as having higher quality; as were those with higher percentage wins.

The effect of experimental result failed to achieve significance (F = 3.0 p = 0.08). However it should be noted that the factor "experimental result" could only be analysed for the subject group not receiving outcome information (n = 44). Thus, this analysis had limited experimental power and given that the potential effect of this factor might be small there is a strong possibility that this result is a Type 2 error. There were no other significant differences. (See Appendix 3 for full Anova table)

#### 4.2.3.3 Analysis for subject group receiving experimental result and long-term outcome information

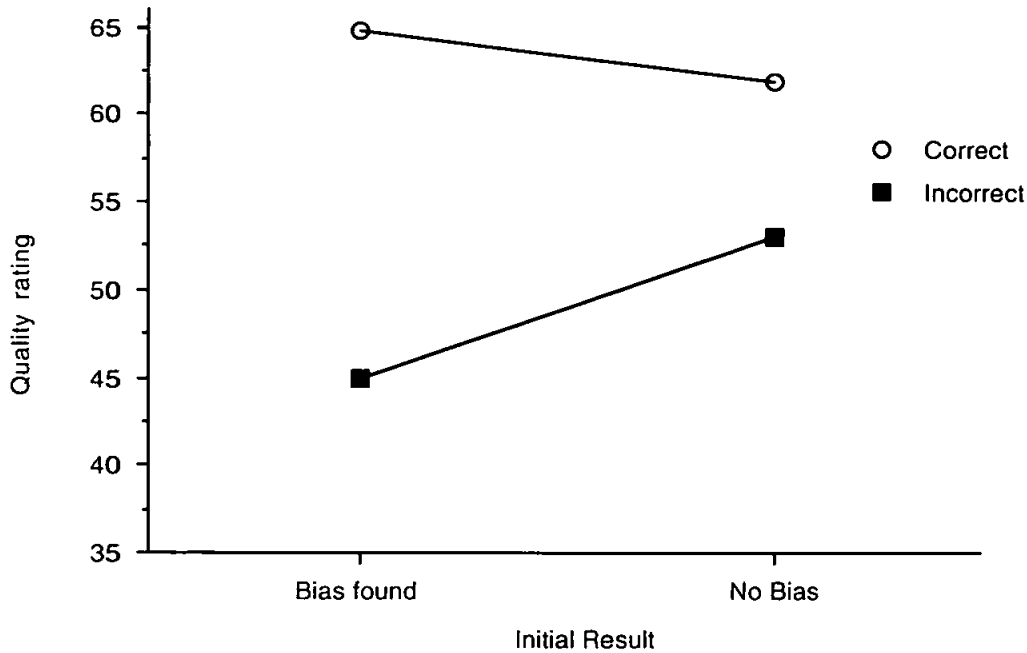
An Anova was performed on the data for those subjects who did receive long-term outcome (replication) information (n = 29); this had four within subjects factors (number of spins, percentage even wins, initial result and long-term outcome).

In this analysis (as in both previous analyses) both number of spins (F = 6.0, p = 0.02) and percentage even wins (F = 14.2, p = 0.0008) had significant influences on quality ratings. As before sampling procedures with higher sample sizes were rated as having higher quality; as were those with higher percentage even wins.

For those subjects given long term outcome information experimental result alone did not significantly influence quality ratings (F = 1.6, p = 0.21). However there was a significant main effect of long-term outcome (F = 21.1, p = 0.0001) [r = 0.37] with experimental results which had been replicated attracting higher quality ratings than those which were not replicated. The interaction between experimental result and long term outcome was also significant (F = 8.5 p = 0.006). This interaction can be seen graphically in Figure 4.1. The interaction shows that the highest quality ratings were associated with those cases where a bias was found and this result was subsequently shown to have been correct. The same initial result subsequently shown to have been false attracted the lowest quality ratings. Where no bias was found initially the long-term outcome had considerably less differential effect.

Figure 4.1

Interaction between experimental result and long-term outcome (Experiment 3)




N.B. The terms correct and incorrect refer to the nature of the long-term outcome information (either confirming or discomfirming the initial result)

### 4.2.4 Discussion of Experiment Three

Experiment 3 showed parallel results to those of Experiment 2 despite the changes in instructions and task and despite the statistical naiveté of the subjects. In both cases sample size was a strong influence on quality ratings. There was an effect in Experiment 2 of subjects giving higher ratings to experiments with smaller significance levels. Although this factor was not present in Experiment 3, a similar strategy can be seen in relation to subjects' higher ratings of experiments with larger effect sizes. In both these cases these choices are consistent with a desire to minimise probabilities of error without reference to potential costs. This strategy has persisted in Experiment 3 despite clear cues to the importance of the cost of sampling in the scenario and direct instructions to consider cost in the definition of quality. The details of this apparent strategy and its implications will be discussed further in the overall discussion at the end of this chapter.

The effects of initial and long-term outcome information in Experiment 3 are directly comparable to those in Experiment 2 and appear largely unchanged despite the task differences.

## 4.3 EXPERIMENT FOUR

### 4.3.1 Introduction

Experiments 2 and 3 had used subjects drawn from subsequent first year psychology courses at Plymouth University. Although these subjects had different levels of statistical training the effects of outcome information were the same for both groups. These results were therefore open to the criticism that they were due to either statistical naivety or specific failures in teaching in the psychology department. It was decided therefore to perform a partial replication of Experiment 3 using experienced second and third year statistics students as subjects. These subjects had received entirely separate training in research methods (taught by different staff) to either of the previous subject groups and were considerably more expert in statistics. As there was a small number of available subjects for this experiment it was decided to run them all in the long-term outcome condition. This enabled results to be comparable with the previous long-term outcome groups and lost very little other information.

### 4.3.2 Method

Subjects and task

Subjects were 40 second and third year mathematics and statistics undergraduates at the University of Plymouth. The experiment took place at the start of a combined statistics lecture. All subjects were volunteers.

Materials and design

The experimental materials, design and procedures were exactly the same as those for the outcome group in the previous experiment. (See Section 4.2.2)

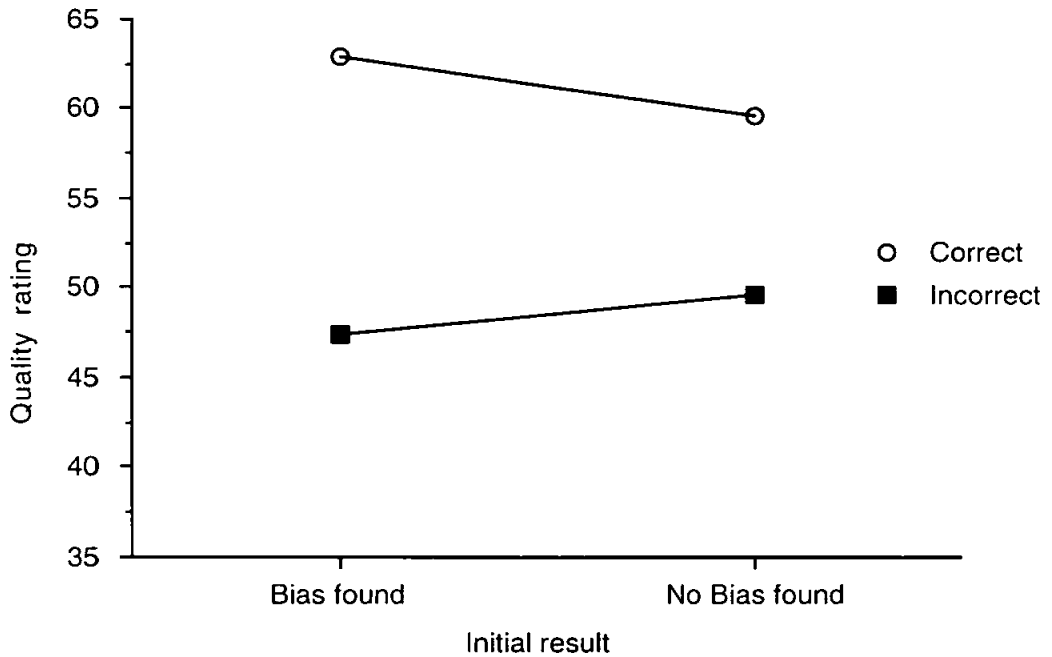
### 4.3.3 Results

All subjects were given long term outcome information. An Anova was performed on the data. There were four within subjects factors (number of spins, percentage even wins, initial result and long-term outcome). In this analysis number of spins ($F = 22.2$, $p = 0.0001$)[$r = 0.28$] had a significant influence on quality ratings. As before, sampling procedures with higher sample sizes were rated as having higher quality. However, in this case the effect of percentage even wins failed to achieve significance ($F = 0.4$, $p = 0.55$)[$r = 0.09$] .

As in the long-term outcome group of Experiment 3; experimental result alone did not significantly influence quality ratings ($F = 10.1$, $p = 0.76$). Again there was a significant main effect of long-term outcome ($F = 23.3$, $p = 0.0001$) [$r = 0.32$] with experimental results which had been replicated attracting higher quality ratings than those which were not replicated. The interaction between experimental result and long-term outcome was also significant ($F = 4.7$, $p = 0.03$). This interaction can be seen graphically in Figure 4.2 and is directly comparable to the interaction of these factors in the previous experiment, shown in Figure 4.1.

Figure 4.2

Interaction between experimental result and long-term outcome (Experiment 4)




N.B. The terms correct and incorrect refer to the nature of the long-term outcome information (either confirming or discomfirming the initial result)

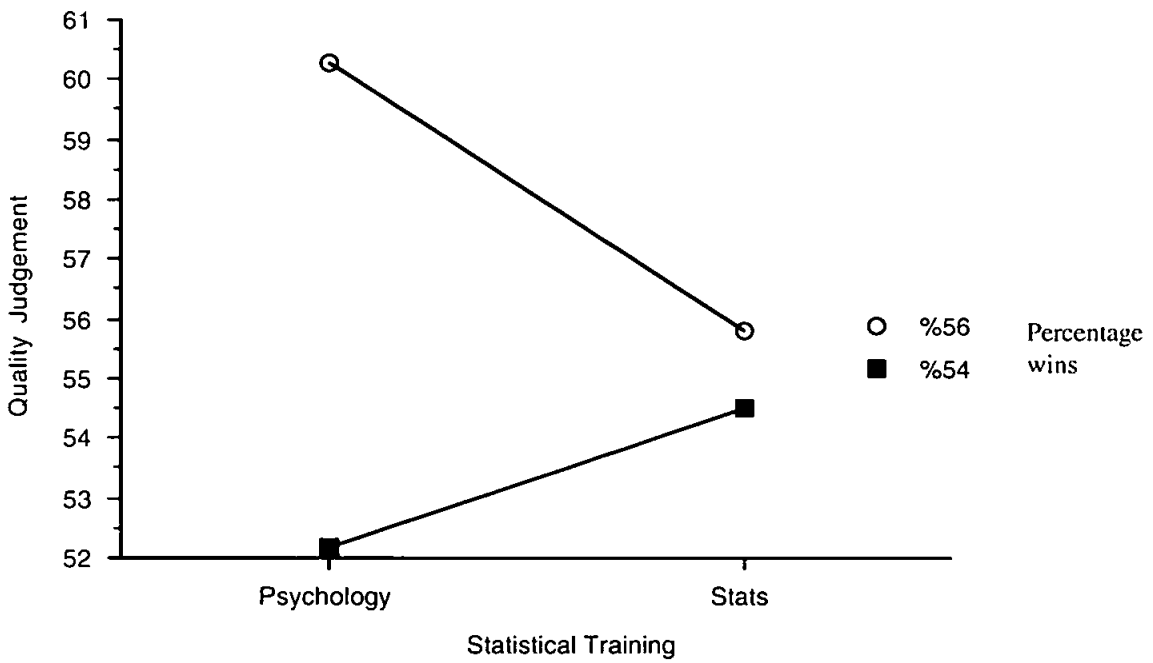
#### 4.3.3.1 Additional comparative analysis

As Experiments 3 and 4 had exactly the same instructions and tasks (in the outcome condition) it was possible to compare their results on a between subjects basis. An Anova was performed on the data from both these groups using the different subject groups as a between subjects factor. This analysis showed no significant differences between the statistically naive subjects (Experiment 3) and the statistically expert subjects (Experiment 4) for the effects of number of spins, or for the initial result / long term outcome interaction. It was not the intention to draw the inference that these groups were the same from a failure to achieve significance in this analysis. However, it was necessary to check for the possibility that the effects of outcome information would be significantly reduced by statistical expertise. There was one difference in the way in which these two groups used the information presented in the tasks, this was demonstrated by a significant interaction

between the percentage differences and the subject groups, F = 4.88, p = 0.03. The graph of this interaction is shown in Figure 4.3. (For complete Anova table see Appendix 3)

Figure 4.3

Interaction between percentage difference and subject group (Experiments 3 and 4)




From this graph it can be seen that the statistically naive psychology subjects utilise the percentage of even wins as a strong indicator of the quality of a sampling procedure. The more statistically expert subjects are considerably less influenced by this factor. It is difficult to find a logical reason for this difference in strategy. The normative analysis shows the probability of error in these procedures to be based on the interaction of sample size (number of spins) and the percentage even wins. (See section 4.2.2) This would suggest that the statistically naive subjects might be using a more appropriate strategy as they at least take this important factor into account. However, as the results of Experiment 3 have shown their overall strategy is simply additive with both higher sample sizes and higher percentage wins attracting higher quality ratings. Thus, although this strategy favours minimum probability of error, it takes no account of the cost of sampling.

It should be remembered, however, that despite this apparent difference in strategy the effects of long term outcome information are still the overriding influence in both cases.

### 4.3.4 Discussion

Experiments 2, 3 and 4 showed clear and consistent influences of both levels of outcome information. Where only the initial result was given, subjects were biased in favour of positive results and against negative results. When subjects were also given long term outcome information this effect was replaced by a main effect of long term outcome, subjects were biased in favour of results which were subsequently proved correct.

A comparison of the graphs of these results for the first three experiments (Figure 2.1, Figure 4.1 and Figure 4.2) shows these effects to be consistent across all three experiments and thus unaffected by changes in task, increasingly explicit task instructions, clear definition of quality and differing levels of statistical expertise. This bias towards significant results in the no-outcome condition and positive outcomes in the long term outcome condition would appear to be clear evidence of hindsight bias. Subjects appear to believe that the designer of the experiment should have 'known it all along' and should have done something to avoid these results.

However it is not possible to immediately assume a purely hindsight explanation. Firstly there is some justification in the view that an experiment finding a significant result can at least be expected to have been powerful enough, whilst a non-significant result may mean either that there is no effect or that the experiment lacked power. However if this was the case then there should be evidence of an increasing tendency to favour significant results in those scenarios where sufficient experimental power is in question; as compared to those with high power. This would lead to an interaction between experimental result and sample size, there was no evidence for this trend in the results.

Secondly, and more importantly, all three experiments showed a similar interaction between the two levels of outcome information. The tendency to favour significant initial results remained present in the long term outcome condition; a significant result subsequently proved to be true was consistently rated higher than a non-significant result subsequently proved to be true. In this case both these results have been shown in hindsight to be correct. However, this situation was reversed when the long term outcome showed the original result to have been false. In this case a positive original result subsequently shown to be false

consistently attracted *lower* quality ratings than a negative original result shown to be false. This is the opposite to what would be expected from a logical analysis of outcome. A false positive result (Type 1 error) can be seen as an unlucky event, the probability of which has been determined by the pre-set significance level and is known beforehand. A false negative result (Type 2 error) does, however, imply an experimental design which lacks power; therefore the logical response would be to rate a false negative as having less quality than a false positive all other things being equal. Thus, this pattern of results cannot be explained by a logical use of outcome information to imply experimental quality. Similarly, this consistent pattern of interaction between initial results and long-term outcomes cannot be explained by the presence of hindsight bias along with an additional bias to positive results. In this case, as in the logical analysis, a false negative would be rated as having less quality than a false positive

This pattern of results can, however, be explained in terms of hindsight bias when the financial implications of different outcomes are taken into account. The true positive result which attracted the highest quality ratings reflected the most financially rewarding outcome in each scenario. Conversely the false positive result which attracted the lowest quality ratings reflected the most financially punitive outcome. Where the initial result was negative there is no great financial gain or loss with either long-term outcome. Therefore one would expect the hindsight effects to be generally smaller as is the case. Within these smaller effects if an initial negative result was subsequently proved to be false this implies a failure to realise a potential financial gain and thus this condition would be rated lower than an initial negative result subsequently proved true which implies a negligible financial loss. This hypothesis implies that subjects utilising hindsight information see the financial outcome to be more indicative of quality of an experiment than either the relevant statistical design factors or the nature of the result. An interesting feature of this hypothesis is that the financial salience of a given outcome not only influences the direction of hindsight effects (either increasing quality ratings where money is made or reducing quality ratings where money is lost) but also influences the size of the bias which appears to be related to the amount of money involved.

## 4.4 EXPERIMENT FIVE

### 4.4.1 Introduction

Experiment 5 was designed to test the hypothesis that both the size and direction (positive or negative) of outcome bias seen in the previous experiments was dependent on the financial implications of the differing experimental outcomes. This hypothesis was based on observation of the specific patterns found in the interactions between initial and long-term outcomes in the previous experiments. In these interactions both the size and direction of bias varied with particular outcomes. This consistent pattern of results could not be explained by the nature of these outcomes alone but was consistent with their respective financial implications.

This potential factor "outcome salience" can be defined as the relative importance of the implications of an outcome from the point of view of the subject. Thus outcome salience relates to the subjective utility of a given outcome and can vary in size in both positive and negative directions. In terms of outcomes with financial implications the situation is relatively clear. Positive or negative outcome salience would be related to gains or losses of money and their sizes would vary with the amounts of money in question. In terms of more general outcomes it is harder to assign specific values to outcome salience; nevertheless it is generally clear whether a given outcome is at least positive or negative from the point of view of the subject and some gross estimate of relative importance is usually possible. That the size of a given bias might be dependent on outcome salience has been tentatively suggested in the review of previous outcome bias studies (Chapter three, Section 3.3.1). Certainly the direction of a bias can be influenced by the subjects' perception of an outcome; this effect can be seen in reverse hindsight studies (Section 3.2.2) and in some outcome bias studies where a particularly bad outcome leads to punitive judgements ( Section 3.3, Mitchell and Kalb, 1981).

In order to test for the predicted effects of outcome salience this experiment sought to vary the financial implications of a given outcome whilst keeping the task and the actual nature of the outcome constant. In order to achieve this it was decided to utilise the same

task presented in the previous experiment but to change the perspective for half the subjects. Thus, half the subjects in this experiment would receive the same scenario, tests and outcomes used in Experiments 3 and 4 (the 'fraudsters' perspective). As before these subjects would receive a scenario in which a team of fraudsters were attempting to win large amounts of money from identifying biased roulette wheels. For the remaining subjects there would be a minor change in the scenario description which would cause subjects to consider the results of the same sampling procedures from the perspective of manufacturers of roulette tables (the 'manufacturers' perspective). In this case the manufacturers of roulette wheels would be attempting to identify ones which are biased in order to prevent losing large amounts of money.

Both these groups have the task of using sampling procedures to accurately determine whether a given wheel is biased or not. These tests lead to the same pattern of initial results and long-term outcomes, however, the financial implications of these results are different from each perspective. For example, for a test initially finding a biased wheel where this result subsequently proves to be true; the fraudsters perspective implies large profits and the manufacturers perspective implies a small loss (the replacement cost of the table). Thus it was predicted that this simple change in perspective would have the effect of reversing quality judgements if these judgements were truly based on outcome salience.

Contingency tables showing the financial implications of each possible outcome from both points of view are shown below in Table 4.2. This table includes estimated values of outcome salience for each possible outcome (in brackets). These estimates roughly follow the financial implications of each outcome although small variations have been made to include other factors thought to be relevant as described below. Values for outcome salience are on a scale from minus one to plus one; where a positive value indicates a good result and a negative value indicates a bad result. Larger values in either direction reflect better or worse outcomes from the point of view of the subject. Outcome salience may also be seen as a form of goal relevance and similarities between these factors will be discussed later.

Whilst these estimates of outcome salience were largely based on the financial implications of each outcome it was necessary to differentiate between financially equal outcomes which had other less well defined positive or negative attributes. For example,

from the fraudsters' point of view an initial result finding no bias leads to no significant monetary gains or losses whether this result is subsequently proved to be true or false. However, if this result is subsequently shown to have been false the fraudsters have missed the opportunity to make a large profit and this situation could clearly be construed as a worse outcome than if their initial result had been correct. In experiments 3 and 4 this implication of missed opportunity was presented to subjects as part of the experimental outcome description and the results of these experiments have shown small yet consistent differences between mean quality ratings for these two outcomes in line with this difference in subjective utility. From the manufacturers' perspective a similar situation occurs where a false positive result implies unnecessary replacement costs as compared to a true positive result where the same replacement cost is warranted. Thus estimates of outcome salience were adjusted to take these factors into account. As the whole point of the manufacturers running tests was to discover biased wheels and replace them; the outcome salience of the true positive result was given a small positive value despite its overall negative financial implications. The outcome salience of the false positive result was given a small negative value due to the associated unnecessary cost outlined above.

These estimated values will allow for subsequent specific analysis of the experimental hypothesis and have been arranged so that their overall totals sum to zero. In the case of the fraudsters' perspective these values were consistent with the pattern of results found in the previous experiments. Indeed they were created from an analysis of these results in terms of the experimental hypothesis that the size and direction of quality judgements arising from outcome bias is mediated by outcome salience. From the manufacturers' perspective the estimated values of outcome salience predict a quite different pattern of outcome bias which reflects the specific predictions of the experimental hypothesis.

Table 4.2

Financial implications and outcome salience estimates (in brackets) for outcomes of sampling procedures from each perspective

Fraudsters' point of view

| | True result | False result |
|---|---|---|
| **Bias found** | Large profit (+ 0.9) | Significant loss (- 0.7) |
| **No bias found** | No gain or loss (+ 0.1) | Missed opportunity of large profit (- 0.3) |

Manufacturers' point of view

| | True result | False result |
|---|---|---|
| **Bias found** | Necessary replacement cost of wheel (+ 0.3) | Money wasted on unnecessary replacement of wheel (- 0.1) |
| **No bias found** | No losses (+ 0.7) | Large losses (- 0.9) |

### 4.4.2 Method

Subjects and task

Subjects were 61 third year psychology undergraduates at the University of Plymouth. The experiment took place during third year option group lectures. All subjects were volunteers.

Design and procedures

As in the previous experiment all subjects were given long term outcome information. For half the subjects (31) the experimental design and procedures were exactly the same as those for the outcome group in the previous experiment. In this condition roulette wheels are tested for bias from the perspective of a team of fraudsters who hope to make large profits on biased wheels.

The remainder of the subjects (30) were given the same experimental description except for a small change in the wording of the scenario section. This change shifted the perspective to that of the roulette wheel manufacturers who fear large losses from biased

wheels. The instructions and scenario for the manufacturers was as follows;

**Introduction**

The following description and series of questions are designed to study understanding of the concept of quality in experimental sampling.

Your co-operation in this experiment is greatly appreciated, and the results should be of use to you as they will demonstrate general levels of understanding of sampling principles. Results and explanations will be made available as soon as possible.

**Scenario**

A worker in a roulette wheel factory had been bribed to produce a number of wheels in which the slots for even numbers were slightly larger than those for odd numbers. It was the intention of a well financed team of fraudsters to bet heavily on even numbers on the biased wheels. Half the casinos ordering new wheels had unknowingly received the biased wheels.

Acting on a rumour the roulette wheel manufacturers employed a team to go to individual casinos and test each wheel, hoping to quietly replace the biased wheels without damaging their reputation or being sued for damages by casinos losing large amounts of money.

It was decided that each member of the team would go to a different casino and make a number of small bets whilst counting the number of even wins to see whether the new wheel was one of the biased ones. Simply standing at the table without betting was regarded as too suspicious as casino security consistently video and study behaviour at roulette tables and the company did not want to alert casinos to possible problems.

Tests involving extensive counting of even wins can be very expensive to run and dangerous in terms of alerting the casino to a specific interest in the roulette wheel. The potential cost of missing a biased wheel is high.

**Test bets**

In initial tests sampling methods varied between individuals. Different numbers of spins were counted and there were also differences in the percentage of even number wins which were regarded as sufficient evidence of a biased table. On an unbiased table even numbers have a 50% probability of winning if zero wins are disregarded; biased tables have a greater probability of even wins. Below are a series of descriptions of the tests used by different team members.

**Task**

Your task is to read each description and give an intuitive estimate of the quality of the test for a biased roulette wheel by writing a percentage mark (from 0 to 100) in the space provided.

**A test of high quality is one which has a high probability of differentiating whether a table is truly biased or not without incurring excessive cost.**

Please try to use the whole range of marks. Thank you for your co-operation.

This subject group received the same task questions with the exception that the section entitled 'outcome' was described from the manufacturers' perspective. This manipulation had the effect of varying the outcome salience of each of the four possible long-term outcomes depending on which perspective was presented. All four possible outcomes for the manufacturers were as follows:

**Outcome :** Subsequent large bets confirmed this wheel to be biased. The wheel was quietly replaced at some cost.

**Outcome :** Subsequent betting by customers confirmed this finding. The table was not biased.

**Outcome :** The wheel was quietly replaced at some cost. However, subsequent factory tests showed no evidence of bias. The wheel had been replaced unnecessarily.

**Outcome :** The team did not replace the wheel. However subsequent betting by customers did show a bias towards even numbers and the company paid considerable damages to the casino.

### 4.4.3 Results

An Anova was performed on the data. There were four within subjects factors (number of spins, percentage even wins, initial result and long-term outcome) and one between subjects factor (perspective).

Effects of experimental factors

Overall, as in previous experiments number of spins had a significant influence on quality ratings ($F = 14.8$, $p = 0.0003$) [$r = 0.28$]. The effect of percentage difference accepted was also significant ($F = 11.1$, $p = 0.001$) [$r = 0.15$]. As in experiments three and four larger sample sizes and larger percentage even wins attracted higher quality ratings.

Effects of outcome information

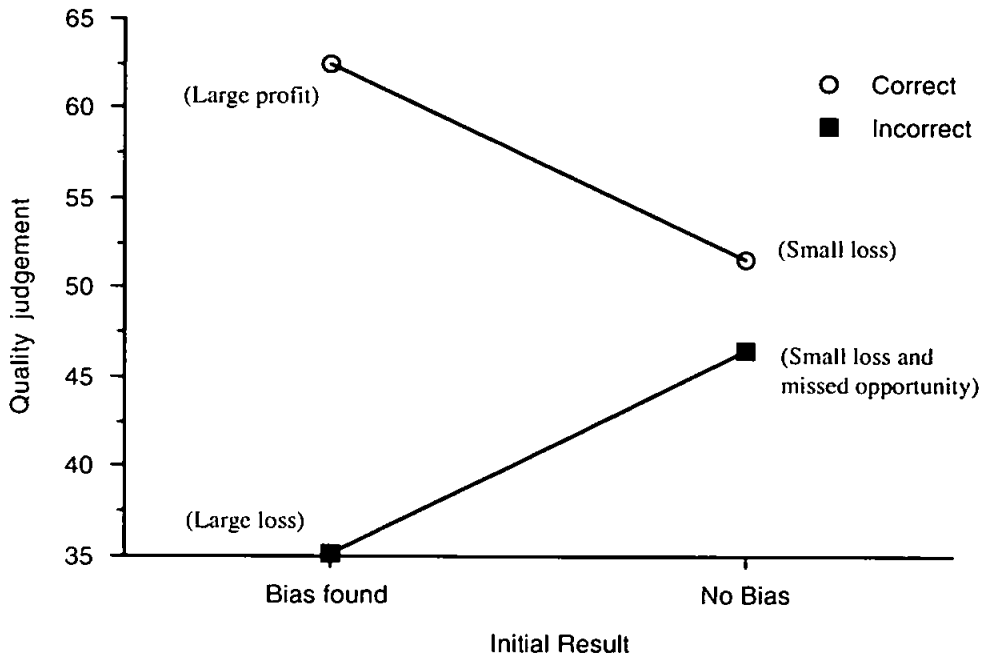
Overall there was no significant effect of initial outcome ($F = 0.03$, $p = 0.9$). For all subjects long-term outcome information did significantly influence quality ratings ($F = 36.8$, $p = 0.0001$) [$r = 0.31$]. The three-way interaction between initial result, long term outcome and subjects' perspective was significant ($F = 28.9$, $p = 0.0001$). The nature of this three-

way interaction can be best explained and compared with previous results by analysing the different effects of outcome information on the different subject groups.

Separate analyses were calculated for each subject group. Subjects given the "fraudster" perspective (the same as Experiments 2 and 3) demonstrated a significant main effect of long term outcome information ($F = 19.8$, $p = 0.0001$) [$r = 0.33$] and a significant interaction between initial result and long term outcome ($F = 24.9$, $P = 0.0001$). Figure 4.4 shows the interaction between initial and long-term outcome information for those subjects with the 'fraudster' perspective. These effects of outcome are directly comparable to those of Experiments 3 and 4 shown in Figures 4.1 and 4.2 respectively and once again replicate these results.

Figure 4.4

The interaction between initial and long-term outcome information for those subjects with the 'fraudster' perspective. Including financial implications of each outcome (in brackets).

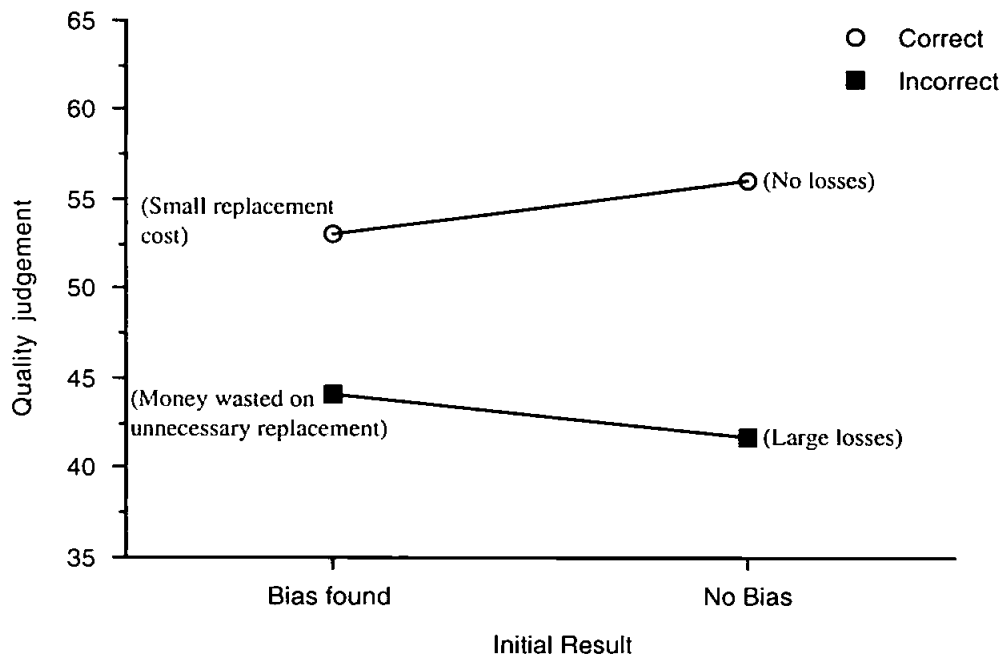



N.B. The terms correct and incorrect refer to the nature of the long-term outcome information (either confirming or discomfirming the initial result)

Subjects given the "manufacturer" perspective also demonstrated a significant main effect of long term outcome information ($F = 17.8$, $p = 0.0002$) [$r = 0.28$] and a significant interaction between initial result and long term outcome ($F = 4.8$, $P = 0.04$). Figure 4.5 shows the interaction between initial and long-term outcome information for those subjects with the 'manufacturer' perspective. When compared with Figure 4.4 it can be seen that the change in perspective has had the hypothesised effect of changing the quality judgements for each particular outcome. In both these graphs the relative financial outcomes have been included and it can be seen that the relative quality judgements follow these outcomes in every case. (See Appendix 4 for complete Anova tables.)

Figure 4.5

The interaction between initial and long-term outcome information for those subjects with the 'manufacturer' perspective. Including financial implications of each outcome (in brackets)

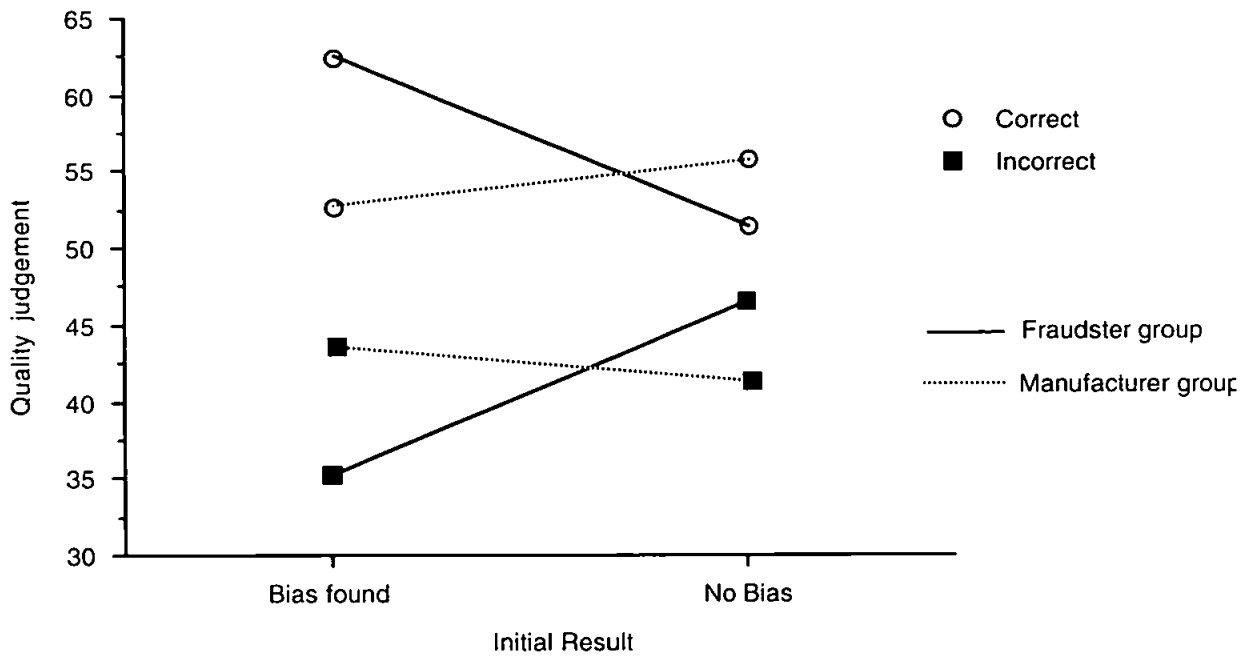



N.B. The terms correct and incorrect refer to the nature of the long-term outcome information (either confirming or discomfirming the initial result)

When the above interaction graphs demonstrating the effect of outcomes on quality ratings for each of the subject groups are compared it can be seen that the overall three-way interaction is due to differences in the effects of specific outcomes for subjects with different perspectives. To make this comparison easier this effect can be seen in Figure 4.6 which demonstrates the nature of the overall three-way interaction between initial result, long-term outcome and perspective. This graph was created by overlaying the graphs of the two-way interactions between initial result and long-term outcome for each perspective (Figures 4.4 and 4.5) and clearly shows that the change of perspective has had the effect of reversing the direction of these two-way interactions.

Figure 4.6

Overlay graph of the interaction between initial result and long-term outcome for both perspective groups.




N.B. The terms correct and incorrect refer to the nature of the long-term outcome information (either confirming or discomfirming the initial result)

The specific predictions of the influences of outcome salience were tested in two planned comparisons. These were based on the estimates presented in the introduction (Section 4.4.1, Table 4.2). For subjects given the fraudster perspective there was a significant effect of outcome salience ($F = 74.4$, $p = 0.0001$) such that quality ratings differed from the mean in the pattern predicted. For subjects given the manufacturer perspective there was also a significant effect of outcome salience ($F = 83.0$, $p = 0.0001$) again quality ratings differed from the mean in the pattern predicted from this perspective. Correlations between predicted values of outcome salience and mean quality ratings for each outcome were 0.98 for the fraudster perspective and 0.93 for the manufacturer perspective. (0.94 overall) For complete details of this analysis see Appendix 4.

#### 4.4.3.1 Comparative results of Experiments two, three, four and five

Experiments 2, 3, 4 and 5 demonstrated consistent influences of experimental and outcome factors on subjects' judgements of the quality of sampling procedures. In each case the relative influences of each factor were similar; the individual sizes of these effects and the overall mean effect sizes are shown in Table 4.3. This table shows long-term outcome to have the largest influence on quality ratings in every case. Next largest are the influences of sample size followed by those of effect size. Two of these experiments presented subject groups with an initial result alone; in these cases there was a relatively small effect in which subjects preferred significant results. That these relative influences are of a consistent size across different experiments is shown by the small standard deviations related to each of the mean values.

It should be noted that these measures of effect size only relate to main effects and do not reflect the effects of the significant interactions between initial results and long-term outcomes. The differences between cell means in these interactions reflect the influences of specific outcomes and were larger than the differences between main effect means in every case. Thus the conclusion that outcome information has the largest effect on subsequent quality ratings is a conservative one; the overall effect is greater than that suggested by these figures.

Table 4.3

Effect sizes for each factor across Experiments 2, 3, 4 and 5.

| Experiment / Factor | Expt. 2 | Expt. 3 | Expt. 4 | Expt. 5 | **Mean (S.D.)** |
|---|---|---|---|---|---|
| Effect size (Significance level in Expt.2) | 0.18 | 0.22 | 0.09 (Non Sig.) | 0.15 | **0.16** (0.05) |
| Sample size | 0.34 | 0.22 | 0.28 | 0.28 | **0.28** (0.05) |
| Initial result (For groups without long-term outcome) | 0.11 | 0.11 | ---NA--- | ---NA--- | **0.11** (0) |
| Long-term outcome | 0.27 | 0.37 | 0.32 | 0.31 | **0.32** (0.04) |

### 4.4.4 Discussion

The results of Experiment 5 were entirely consistent with the results of Experiments 2, 3 and 4 in terms of the influences of the relevant experimental factors on judgements of the quality of sampling procedures. As in these previous experiments, both larger sample sizes and larger effect sizes attracted higher quality ratings. Large influences of outcome information were also demonstrated in Experiment 5. Again long-term outcome was the most influential factor in determining subsequent quality ratings. (For relative influences of all factors see Table 4.3 above)

In this experiment it was possible to differentiate between the nature of varying outcomes and their outcome salience from a given perspective. When subjects' points of view were taken into account the results strongly supported the view that subjects' quality judgements were largely biased in favour of positive financial outcomes and against negative ones. This basic result was shown by the significant interaction between outcome information and subjects' perspective (See section 4.4.3). Graphical representations of the different pattern of results from each perspective can be seen in Figures 4.4 and 4.5 and these differing results can be compared in the overlay graph; Figure 4.6. These figures show the effects of outcome information on mean quality ratings to be roughly consistent with differences in the financial implications of these outcomes from the subjects' perspective.

Further detailed analysis utilising prior estimates of outcome salience showed this hypothetical factor to be an excellent predictor of both the direction and the size of outcome biases. As described in the introduction (Section 4.3.1), estimates of outcome salience were largely based on differing financial implications. Yet, as monetary values alone did not completely reflect differences in subjective utility for every outcome, these estimates were adjusted to include the negative influence of missed opportunities or unnecessary costs. The resulting estimated values for this factor showed a significant fit with the specific pattern of results and very high correlations with mean quality ratings for each possible outcome from both perspectives. This is not particularly surprising in the case of the fraudster perspective as the estimates of outcome salience used for this analysis were consistent with the patterns of results seen in Experiments 2, 3 and 4 and this perspective replicated those experiments. However, for the manufacturer perspective there was a new pattern of outcome bias results and it is the very accurate prediction of these results which gives strong support to the experimental hypothesis that the subjective utility of each outcome as defined in outcome salience was the defining factor in outcome bias.

## 4.5 OVERALL DISCUSSION

This discussion will focus on the results of Experiments 2, 3, 4 and 5, where subjects were asked intuitively to estimate the quality of various sampling procedures. In all cases four basic factors were manipulated. Descriptions of sampling procedures varied in terms of two relevant experimental factors; sample sizes and either significance levels or effect sizes. Secondly two levels of outcome information were included in these descriptions; the initial result of any sampling procedure and the long-term outcome (i.e. whether the initial result was subsequently replicated). The first three experiments have shown that subjects' judgements of experimental quality were consistently biased by outcome information. Experiment 4 has shown that the size and direction of this bias was dependent not on the particular outcome itself, but rather on the salience of that outcome from the perspective of the subject.

The focus of these experiments was on the effects of the presence of irrelevant outcome information on subjects' subsequent quality judgements. Even with little statistical knowledge it should have been possible for subjects to recognise the need to ignore this information and focus solely on the relevant experimental factors. There may be some defence of subjects' inclusion of outcome information in the view that a replicated experimental result implies a better experiment than one which failed to replicate. However, there is little that is defensible in the view that an experiment which happened to lead to the subjects' preferred outcome regardless of its actual result is a better experiment.

Although these experiments have shown outcome information to be the largest influence on quality judgements in every case, subjects' judgements were also influenced to a large extent by the relevant experimental factors. These influences followed a pattern which appeared to be unchanged across subject groups with differing statistical knowledge. Thus this overall discussion will first address subjects' use of relevant experimental factors and then go on to address their use of outcome information.

### 4.5.1 Influences of experimental factors

These tasks are difficult. It would be unreasonable to expect subjects to be able to intuitively calculate a normative analysis of the probabilities of Type 1 and 2 error for these descriptions of sampling procedures. To make a rational judgement of the quality of a given sampling procedure the probability of these errors would then have to be balanced against the potential costs of these errors and the potential costs of sampling in each case. Given that some of the subjects in these experiments had little statistical training, comparisons of their performance with a normative model were neither necessary nor warranted. Despite this fact some basic expectations of subjects' performance on these tasks seems reasonable. For example, it would not be difficult for subjects to adopt a strategy which at least relies exclusively on changes in the relevant factors.

In Experiment 1, given equal outcomes, quality judgements were based on an additive combination of the largest sample size and the most stringent significance level. In terms of sample size subjects favoured more powerful designs apparently completely ignoring considerations of experimental cost. The choice of more stringent significance levels was not

consistent with judgements based on maximising experimental power and thus reducing Type 2 error. However, this strategy may have reflected a desire to reduce the possibility of a false positive result (Type 1 error). These results may be seen as reasonable judgements in a scenario which was based on drug companies testing new products. In this scenario it was reasonable to assume the potential benefits (or losses) might be large enough in each case to outweigh considerations of sampling cost and necessitate retaining the least probability of Type 1 error. In this scenario a Type 1 error would lead a company to market a drug which does not work. A Type 2 error in this case implies a failure to market a drug which does work, which whilst it implies financial losses may not be seen as such a serious mistake.

Another possible interpretation is that there was little understanding of statistical principles and the choice of significance level was based on a feeling that a result at the level of $p < 0.01$ was somehow better than one at $p < 0.05$. There is some evidence for this latter view from the interaction between significance level and experimental result shown in Experiment 2 (See Figure 2.2). In this experiment the set significance level of $p < 0.01$ only had a positive influence on quality ratings when it was known in hindsight that the sampling procedure had achieved a positive result at this level. Thus, this would seem to be more an outcome bias effect (not only a significant result but a very significant result) than one due to considerations of statistical implications. Notably this significant interaction is not present in Experiments 3, 4 and 5 where the scenario and the use of percentage values makes this relationship much less obvious.

In Experiments 3, 4 and 5 the factor of pre-set significance levels was replaced by a measure of effect size expressed as a percentage. In this case subjects preferred sampling procedures which only accepted larger effects as evidence of significant differences. As before subjects also preferred larger sample sizes. This strategy has the same implications as that used by the subjects in Experiment 2 in that it minimises the probability of Type 1 error regardless of considerations of power or cost. Again this may not be an unreasonable strategy given that, in these scenarios from the fraudsters' perspective, a Type 1 error implied losing a large amount of money betting on a roulette wheel that you believed to be biased in your favour when it was not. A Type 2 error is less serious from this perspective as it only implies a missed opportunity to make a large amount of money and no real losses.

However, Experiment 5 also included a different (manufacturers') perspective. From this perspective a Type 1 error implied only the small unnecessary cost of replacing a wheel which was not in fact biased, whilst a Type 2 error implied large losses. If it were true that subjects' strategies were based on an analysis of the relative costs of Type 1 and Type 2 errors, this group should change their responses. In order to minimise Type 2 errors this group should favour the use of smaller percentage differences as evidence of biased tables. There was no evidence of this difference in the results.

Overall the most likely explanation for these results is one in which subjects faced with the very difficult task of making a quick intuitive estimate of the quality of a sampling procedure adopt a simple additive strategy without recourse to any form of formal calculation or cost benefit analysis. This strategy could be described as a sort of biggest is best approach; favouring larger sample sizes and more highly significant results or larger effects. To minimise Type 1 error regardless of cost and power involves the simplest analysis of these experimental factors and can be seen as reasonable in most scenarios. That this strategy remained unaffected by the introduction of clear definitions of quality and cues about sampling costs is not surprising as neither of these factors serve to make the task any simpler. Indeed a normative analysis of these sampling procedures performed intuitively in the time given to these subjects would be beyond the powers of any but the most expert statisticians. Given this task difficulty, it was not surprising that this strategy did not vary between subject groups with differences in statistical expertise. Thus, the overall picture is one of the use of a simple strategy in the face of a very difficult task and this would be a perfectly reasonable response were it not for the pervasive influences of irrelevant outcome information.

### 4.5.2 Influences of outcome information

The results of Experiments 2, 3, 4 and 5 showed outcome information to be the largest influence on subjects' quality judgements in every case. This effect was unchanged despite clear definitions of quality and varying statistical expertise. Not only did subjects fail to ignore this information but its influence overshadowed influences of the relevant experimental factors. The previous section discussed subjects' use of the relevant

experimental factors in generating quality judgements. The resulting strategy had some merits and could be seen as reasonable in the light of task difficulty. However, this strategy cannot be seen in isolation combined as it was with large influences of outcome information.

These results have important practical implications which are not evident from previous studies of hindsight effects. In the present studies where no long term outcome information (and thus no financial implication) was given, outcome effects were quite small with an average r = 0.11; in this case quality judgements were mostly influenced by considerations of sample size. This reflects the small and not particularly influential hindsight bias reported by Christensen-Szalanski and Willham (1991). (See Chapter 3, Section 3.4) Notably, the studies used in this meta-analysis were mostly of the standard hindsight paradigm in which the hindsight information has no particular salience to the subject.

However, when long term outcome information (and its financial implications) was introduced in these experiments the size of hindsight bias effects increased to an average r = 0.32 and became the most influential factor in influencing quality judgements. This result mirrors the practically important findings of Mitchell and Kalb (1981) where evaluations of nurses' performance were biased in hindsight given the particularly salient outcome of injury to patients. In the present experiments, the introduction of a salient outcome increased the size of outcome bias in quality judgements and this effect was also shown to be almost completely dependent on the implications of a given outcome from the point of view of the person making the judgement. Experiment 5 has shown that by changing subjects' perspectives, and thus the relative salience of particular outcomes, both the size and direction of resulting outcome biases are changed. In this experiment the introduction of estimates of outcome salience showed that this factor was an excellent predictor of the nature of resulting biases.

This result presents a somewhat different view of hindsight influences to that normally reported. The actual outcome of a given event appears to be of little importance without consideration of the salience of that outcome from the perspective of the subject. That a small and consistent bias towards positive (or confirming) results was present was shown by the bias towards significant results seen in the groups which did not receive long-term outcome

information. This effect would be the equivalent of the effects generally seen in experiments using the standard hindsight paradigm where the result has little or no salience to the subject. This effect is completely overshadowed, or even reversed, by the inclusion of the personal implications of an outcome. The effect of outcome salience is all the more impressive when it is considered that within these experiments subjects readily conform to a perspective suggested in the instructions and that a small change in wording is all that is needed to change this point of view. This implies that in real life where a strongly held personal conviction is at stake hindsight bias would be an extremely influential factor.

The present paradigm therefore has some clear advantages over the standard hindsight paradigm. The sort of judgements of quality which are made every day are generally salient to the person making the judgement. Judges have a point of view and different outcomes will be more or less favourable depending on that point of view. It will be the personal salience of these outcomes rather than the outcomes per se which influence judgements when hindsight information is available. The inclusion, and manipulation, of personal salience of outcomes demonstrates the presence of a much larger and practically important hindsight bias than has previously been seen using the standard paradigm.

Further advantages of the present paradigm include the fact that the outcome information is clearly not relevant to the judgement required. This is important because it refutes the theory that hindsight bias is based on the rational inclusion of available relevant information or the inability to ignore relevant information. (See Chapter 3, Sections 3.2.1, 3.3 and 3.3.1) In this paradigm all the relevant information required to make the judgement was available to the subjects in the form of sample sizes and significance levels or effect sizes. It could be argued that the initial result and long-term outcome of a sampling procedure may contain some relevant information even though it is patently clear that this information would be unavailable until after a sampling procedure was designed and run. However, there was no evidence in these experiments that this information was being used rationally. Indeed by manipulating the subjects' perspective a particular outcome can be given a positive or a negative outcome bias. The direction and size of outcome bias was clearly seen to be dependent not on the nature of the outcome but rather the salience of that outcome. Given this fact the view that these biases may be based on some form of rational

analysis of available relevant information is no longer tenable.

Given that the size of the resulting bias is largely dependent on outcome salience and can overshadow all other factors relevant to the judgement where this salience is high, this factor warrants serious consideration. For example, it may be possible to account for the differing effects seen in previous hindsight and outcome bias studies by an analysis of the outcome salience related to the tasks used in these studies. In addition the inclusion of the importance of a subjects' point of view may also have implications for other biases. These theoretical issues will be addressed in Chapter 6.

## 4.6 SUMMARY OF CHAPTER FOUR

Chapter 4 has described three experiments based on a scenario in which subjects were required to judge the quality of sampling procedures from a particular point of view. The basic scenario was one in which the subject was a member of a team of fraudsters who were testing roulette wheels to try to discover which of them were biased. This scenario served to answer a number of possible criticisms arising from the design of Experiment 2. Using this scenario it was possible to present subjects with a task closely related to that of judging the quality of scientific experiments without cueing the specific statistical nature of the problems. More importantly it was possible to manipulate the salience of particular outcomes from the point of view of the subjects in terms of their financial implications. In addition, these scenarios presented a specific definition of experimental quality and emphasised the need for subjects to take account of sampling costs.

The first of these experiments, Experiment 3, produced results directly comparable with those of Experiment 2 despite these changes in presentation. Again, outcome information had the greatest influence on subjects' judgements of quality. Experiment 4 replicated the long-term outcome condition of Experiment 3 using a more statistically expert subject group than that used in either of the previous experiments. This made no appreciable difference and again the same pattern of results was reported. The specific influences of the interaction between initial and long-term outcome information in both these experiments

were discussed. It was suggested that this persistent pattern of results could not be explained by differences in the nature of each possible outcome alone. It was hypothesised that both the size and direction of these outcome biases were mediated by the salience of each outcome from the point of view of the subject.

Experiment 5 tested this hypothesis by the introduction of a perspective change. In this experiment one group of subjects were given the fraudster perspective used in the long-term outcome groups of Experiments 3 and 4. Another group of subjects were given the same problems presented from a manufacturers' perspective. This manipulation had the effect of presenting exactly the same problems and outcomes, whilst altering the salience of each given outcome for the different subject groups. Results for the group given the fraudster perspective again replicated the pattern of results seen in the long-term outcome groups of Experiments 3 and 4. The group given the manufacturer perspective showed a different pattern of results related to the influences of the interaction between initial and long-term outcome information. As predicted, these results showed the size and direction of outcome biases to be largely dependent on outcome salience.

In an overall discussion, both the influences of experimental factors and the influences of outcome information on judgements of quality were discussed.

## Table of contents - Chapter five

# CHAPTER FIVE

## Additional experiments

## 5.1 CHAPTER INTRODUCTION

Chapter 5 consists of two additional studies which served to clarify a number of issues arising from the previous experimental research. The first of these studies, Experiment 6, explored the belief system underlying the outcome bias seen in previous experiments. In this experiment subjects rated the relevance of experimental and outcome information to potential judgements of quality. It was hoped that the results would clarify whether the outcome bias seen in previous experiments arose from the logical application of a false belief in the relevance of outcome information, or was truly a subconscious bias. A manipulation of subjects' point of view was included in this rather simple rating study in order to test the possibility that subjects' relevance ratings would themselves be influenced by outcome salience.

The second experiment presented in this chapter, Experiment 7, was designed to strengthen the theoretical links between outcome and hindsight effects. Hindsight bias has traditionally been demonstrated as a bias of memory, whereas outcome bias has always been demonstrated in on-line paradigms as a bias of judgement. Experiment 7 explored the possible existence of a memory bias coincident with outcome judgements. This experiment utilised a design similar to that of Experiments 3, 4 and 5, with the exception that long-term outcome information was presented some time after an original quality judgement and at this later time subjects were required to recall their earlier judgements. This change made the design compatible with the majority of hindsight studies which utilise the memory of earlier judgements.

## 5.2 EXPERIMENT SIX - PERCEIVED RELEVANCE OF EXPERIMENTAL AND OUTCOME FACTORS

### 5.2.1 Introduction

Experiments 2, 3, 4 and 5 have shown judgements of experimental quality to be biased by the presence of outcome information. The size and direction of this outcome bias effect was shown to be dependent on outcome salience; a measure of the subjective utility of a given outcome. Whether this effect is a true bias or simply an error of judgement is dependent on subjects' beliefs about the relevance of outcome information to the judgement they are required to make.

Firstly, subjects may be well aware that outcome information (and particularly information on the utility of a particular outcome) is not relevant to the quality of the original experimental design. In this case the bias would reflect unconscious influence on judgements of quality and would present great difficulty in terms of debiasing. This situation is suggested by the fact that emphasising and clarifying the definition of quality in these experiments has no apparent debiasing effect. Also the bias remains equally influential for more statistically experienced subjects who should have a greater understanding of the irrelevance of this information to the quality of the original design.

Secondly, it is possible that outcome bias stems from a false belief that outcomes and the utility of those outcomes <u>are</u> relevant to the quality of the design. In this case the resulting influence on quality judgements can be explained as a logical application of what were (falsely) believed to be relevant factors. If this were the case then debiasing could be achieved simply by explaining to subjects that outcomes and utilities were not relevant to the judgement being made. In this scenario it might also be possible to explain the differential effects of outcome salience in terms of resulting beliefs in the relevance of salient factors.

Altogether, over all the previous studies, six factors were varied; sample sizes, effect sizes, pre-set significance levels (in experiment two only), initial results, long term outcomes and the financial implications of those outcomes. The present study was designed to explore subjects' belief in the relevance of these factors to judgements of experimental quality. It

would have been possible to simply ask subjects to rate these factors according to their relevance to judgements of experimental quality, however, as quality judgements were influenced by outcome salience there was every possibility that relevance ratings would also be influenced by similar considerations. Thus in the present experiment it was decided to include a variation in the perspective from which a prospective judgement was being made. This manipulation had the effect of presenting the subject, either as the judge of someone else's experiment, or as the author of an experiment to be judged by someone else.

It was predicted that neither subject group would express a strong belief in the relevance of outcome information in general and that relevance scores would be particularly low for financial outcome information. However, if there were a tendency to believe in the relevance of outcome information, it was predicted that this would be greater for subjects who had the perspective of judging the experiments of others rather than those whose own experiments were being judged. These predictions were made for two reasons. Firstly, if outcome salience does affect beliefs about relevance, then an outcome which was potentially financially beneficial to the subject would be rated as more relevant than one which was not. Thus it was expected that this influence would be demonstrated specifically in the case of judgements of the relevance of financial outcomes. Secondly, the potential effect of these different perspectives would reflect the difference between the difficulty judges have in ignoring outcome information and the ease with which outside observers recognise this information as irrelevant. These perspective related differences are already clearly visible in both hindsight and outcome bias studies when the differences between the subjects of these studies and readers of subsequent articles are considered. Hindsight bias has been classically described as the 'I should have known it all along effect'. By comparison, earlier in this thesis the outcome bias effect was described as a 'they should have known it all along effect' referring as it does to judgements of the performance of others. (Chapter 3, Section 3.3) As there is an element of unreasonable overestimation of one's own abilities in hindsight biased memories, there is an element of unfair underestimation of others abilities in outcome biased judgements. From the perspective of an outside observer (either an experimenter or a reader of the article) judgements biased by outcome information seem clearly unreasonable and yet the subjects of these experiments continue to utilise this

information. It may be the case that subjects in these experiments also see these biases as unreasonable but are unable to prevent their occurrence. On the other hand, from the point of view of the judge the use of outcome information may appear reasonable.

It was decided not to include a definition of experimental quality in the scenario of this experiment in order to give subjects the widest possible range of interpretation of this concept. The use of a rigid statistical definition has been shown to make no difference to levels of bias in the previous experiments. In addition, if subjects believe outcome information to be irrelevant to undefined quality judgements they would be unlikely to find it relevant where quality was defined statistically. Thus, leaving quality undefined gives subjects every possibility to rate outcome factors as relevant.

### 5.2.2 Method

Subjects

Subjects were 36 first and second year Psychology undergraduates. All subjects were volunteers.

Materials and Design

Subjects were presented with one of two scenarios related to the reviewing of an experimental study submitted for publication in a journal. The two scenarios reflected different points of view. In one case subjects were told that they were reviewing a study submitted by someone else. In the other case subjects were told that someone else was reviewing a study submitted by them. In both cases they were asked to rate the relevance of six pieces of information about the study.

The subject instructions and scenario from the perspective of you reviewing someone else's experiment was as follows:

Below is a list of six pieces of information about an experiment.
Imagine that you were reviewing an experimental study submitted for publication in a journal and this study contained an experiment to which these six pieces of information were related.
If you were asked to judge the quality of the experiment in this study, how relevant would each of the pieces of information be to your judgement ?

Please estimate the relevance of each piece of information using a scale from **0 to 9**.
Where a piece of information which **would not influence** your judgement scores **0**
and a piece of information which **would strongly influence** your judgement scores **9**.

| INFORMATION | SCORE (0 TO 9) |
|---|---|
| The number of subjects used in the experiment | |
| The significance level set in the experiment | |
| Whether the experiment found a significant result or not | |
| The size of the experimental effect | |
| Whether the result of the experiment was replicated by later experiments | |
| The amount of money which you personally stood to gain or lose as a result of the experimental findings | |

The subject instructions and scenario from the perspective of someone else reviewing your experiment was as follows:

> Below is a list of six pieces of information about an experiment.
> Imagine that you had submitted an experimental study for publication in a journal and this study contained an experiment to which these six pieces of information were related.
> If the person reviewing the study was asked to judge the quality of your experiment, how relevant should each of the pieces of information be to their judgement ?
>
> Please estimate the relevance of each piece of information using a scale from **0 to 9**.
> Where a piece of information which **should not influence** their judgement scores **0**
> and a piece of information which **should strongly influence** their judgement scores **9**.

| INFORMATION | SCORE (0 TO 9) |
|---|---|
| The number of subjects used in the experiment | |
| The significance level set in the experiment | |
| Whether the experiment found a significant result or not | |
| The size of the experimental effect | |
| Whether the result of the experiment was replicated by later experiments | |
| The amount of money which the reviewer stood to gain or lose as a result of your experimental findings | |

Subjects were randomly allocated to one of the two perspective conditions. The resulting design had two independent variables. The factor being rated had six levels (see above) and was within subjects. The subjects' perspective had two levels (you reviewing anothers' study or another reviewing your study) and was between subjects. The dependent variable was subjects' ratings of relevance.

Procedure

The instructions, scenario and task was presented to subjects who completed ratings alone in their own time.

### 5.2.3 Results

There were significant main effects of both experimental factors. There was a significant difference between the two perspective groups ($F = 4.5$, $p = .04$). The subject group who had the perspective of reviewing someone elses' experimental study produced higher mean overall ratings of relevance than the group with the perspective of having their own study reviewed by another. There were also significant differences between the factors rated ($F = 425.1$, $p = .0001$). The details of these ratings for each factor can be seen in Table 5.1. A follow up analysis of this main effect showed the financial outcome to be rated as significantly less relevant than all other factors. The factors related to effect size, the initial experimental result and replication (long-term outcome) information were all rated as significantly less relevant than the factors related to subject numbers and significance level. (For Anova table and follow up analysis see Appendix 5.)

Table 5.1

Mean ratings of the relevance of each factor

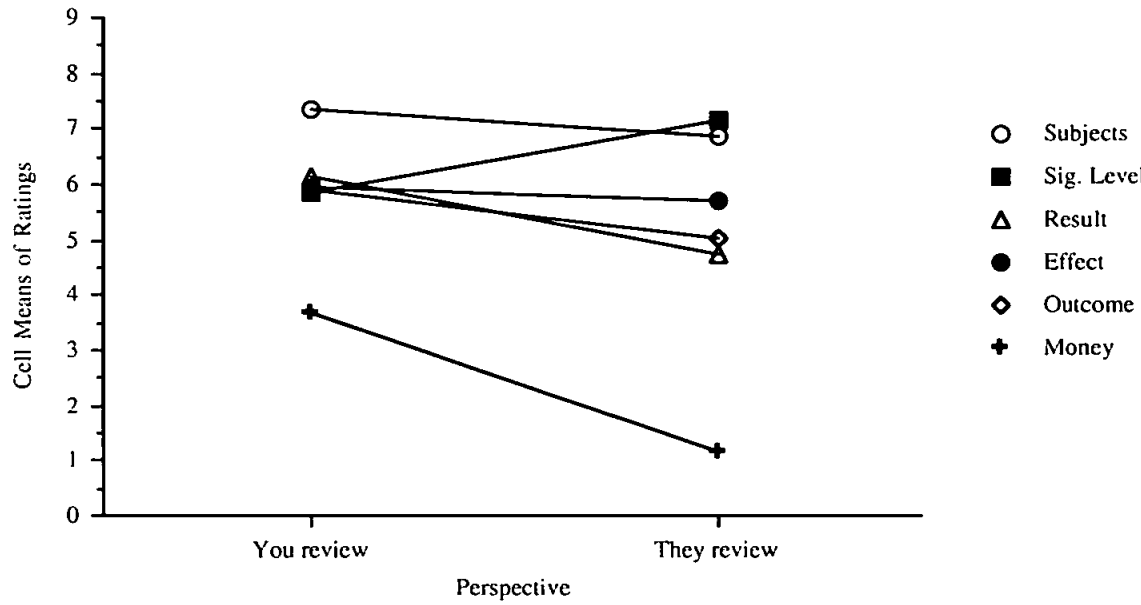
| Factor | Mean relevance rating (S.D.) |
|---|---|
| The number of subjects used in the experiment | **7.1** (1.2) |
| The significance level set in the experiment | **6.5** (1.8) |
| Whether the experiment found a significant result or not | **5.4** (2.6) |
| The size of the experimental effect | **5.8** (1.8) |
| Whether the result of the experiment was replicated by later experiments | **5.5** (2.4) |
| The amount of money which the reviewer stood to gain or lose as a result of your experimental findings | **2.4** (2.5) |

There was also a significant interaction between relevance ratings for different factors and the perspective from which these ratings were made ($F = 3.9$, $p = 0.002$). This effect was specifically due to the differences between ratings for the financial factor across the different perspectives. No significant differences across perspective were shown for any of

the other factors. The nature of this interaction can be seen in Figure 5.1.

Figure 5.1

Graph of the interaction between questions and perspective




### 5.2.4 Discussion

This experiment was designed to answer the question 'Do subjects believe that outcome information is relevant to judgements of the quality of experimental design ?'. The results suggest that the answer is more complicated than a straight yes or no. Both initial and long-term outcome information were rated as less relevant than the experimental factors of sample size and significance level. This difference should not be interpreted without reference to the mean ratings for these factors; all of which were above the mid-point of the relevance scale. Thus, subjects appear to believe that all these factors are to some degree relevant to judgements of experimental quality, with experimental factors being slightly more relevant than outcome information.

What is clear from these results, is that subjects believed that the financial implications of an outcome should be the least important factor in judgements of the quality of experimental designs. However, as predicted the strength of this belief was dependent on the subjects' point of view. The interaction between the factors rated and subjects' perspective shows the presence of bias in these ratings. Whilst subjects clearly felt it was

unreasonable for reviewers to utilise financial outcome when judging the quality of the subject's own studies, they were not prepared to be so scrupulous in their own judgements of others' studies. Notably the opposite occurs with judgements of the relevance of significance levels. Subjects felt that their own studies should be judged on the basis of the significance of their results, but believed it was not so important for them to use this factor when judging the studies of others. This effect can hardly be described as an instance of outcome bias as there is no real outcome in this experiment, and yet it is an example of self-centred information processing which has a lot in common with outcome bias effects and is clearly in line with the predictions based on outcome salience which were made in the introduction.

These results are interesting when seen in terms of the results of the previous studies. A comparison between the mean relevance ratings shown in Table 5.1, and the actual use of these factors shown in the previous judgement studies, shows a sharp contrast between what subjects believe they would do and what they actually do when judging experimental quality. In Experiments 3, 4 and 5 long-term outcomes and their financial implications have been shown to be the largest factor influencing quality judgements. These factors are rated as the least relevant in the present study.

Thus, the results of the present experiment uphold the view that outcome bias is an unconscious bias in that explicit judgements of relevance of information are in opposition to the use of this information in actual quality judgements. Subjects were aware that this information was the least relevant and yet the previous studies have shown it to be the most influential factor in actual quality judgements. This finding supports the view that judgements are the result of implicit Type 2 processes. In addition, the manipulation of perspective in this experiment has shown that even apparently straightforward relevance ratings were influenced by subjects' points of view. In the light of these results it may be necessary to re-evaluate both hindsight and outcome effects in terms of a potentially more general and wide reaching effect of self-centred information processing. This possible theoretical approach will be discussed in Chapter 6.

## 5.3 EXPERIMENT SEVEN - INFLUENCES OF OUTCOME INFORMATION ON MEMORY

### 5.3.1 Introduction

The standard hindsight bias paradigm involves subjects making a judgement of the likelihood of a particular outcome given certain information. At a later time these subjects are given further (hindsight) information (described as the actual outcome) and asked to recall their original judgement. Hindsight bias is demonstrated when the hindsight information biases this memory in favour of the actual outcome. Through numerous replications this paradigm has shown consistent if not particularly large effects. A number of general criticisms have been levelled at this research. (See section 3.4) These include the claim that as effect sizes are small hindsight bias is unlikely to have much influence on actual decisions. Also it has been claimed that as hindsight information is relevant to the original judgement it is reasonable to incorporate this additional information in judgements.

In Experiments 2, 3, 4 and 5 biases relating to outcome information were demonstrated using a paradigm which to some extent answered these criticisms. By manipulating levels of both positive and negative outcome information presented at the same time as the task information, it was possible to demonstrate large and highly influential biases in judgements of quality. Unlike biases seen in the original paradigm, these biases were not mediated by the ability to remember an earlier judgement. In addition, they were related to real life decisions (of experimental quality) and were based on outcome information which was not relevant to the judgement in question. The results of these studies have shown quality judgements to be largely dependent on the subjective utility of the outcome presented.

The relationship between these results and the generally accepted form of hindsight bias is not immediately clear. In the original hindsight paradigm the bias was due to the influence of subsequent outcome information on the memory of an earlier judgement. In the present research bias was due to the influence of irrelevant outcome information on an original judgement. In both cases subjects fail to treat the outcome as irrelevant to the task in hand. In the hindsight paradigm the outcome is relevant to the original judgement but not

to the memory of this judgement. In the present research the outcome is not relevant to the original judgement (quality of the design); firstly because outcome information was not available to the designer and secondly because the experimental result may be due to chance (Type 1 or Type 2 error). Despite these apparent differences, one of the contentions of this thesis was that hindsight bias and outcome bias were essentially the same effect seen from a different point of view. The point of view in question being the particular paradigm used to explore the effect. (See section 3.3) This contention implies that there is one underlying biasing effect related to the presence of outcome information (outcome bias) and that the effect traditionally known as hindsight bias is simply one form of this bias.

The findings from the present series of experiments show outcome salience to be the most important factor in determining the size and direction of biases related to outcome information. These experiments have used an on-line paradigm where the outcome is presented along with other relevant information. If this outcome bias effect is pervasive there is no reason not to suppose that any judgement required of subjects would be biased by salient outcome information whatever this judgement relates to. From this point of view, the traditional hindsight bias effect could be seen as an outcome bias effect where the judgement required is a memory of an earlier probability estimate. In this case whether outcome information biases judgements made at the time or memories of earlier judgements depends only on whether outcomes are presented at the time of the required judgement or at some later time.

In this scenario the outcome salience relevant to traditional memory based hindsight effects is not immediately apparent. Although there would inevitably be a positive salience related to the belief that one had 'known it all along' a number of other factors specific to the study in question might also influence outcome salience. The positive aspects of having 'known it all along' would include motivational factors such as self justification and an increased sense of control. These factors would result in a potential outcome salience in the in the traditional hindsight paradigm which would not be particularly strong relative to the outcome salience of most outcome bias paradigms. This fact serves to support the above contention as hindsight bias effects are generally much smaller than those found using outcome bias paradigms.

This view predicts that where outcome information is presented after the event, the memory of any number of factors may be biased. Where subjects were asked to recall an original likelihood estimate this would create a 'hindsight' effect. If subjects were asked to recall other relevant factors there is no reason to suppose that these memories would not also be biased in hindsight.

The present experiment tested these predictions by presenting subjects with four descriptions of experiments. These descriptions varied in terms of the number of subjects used, the significance level set and the initial result (significant or non significant). In a similar paradigm to that used in Experiments 3, 4 and 5, subjects were required to make a judgement of the quality of these experimental designs. However, this experiment also contained a second stage which reflected the type of paradigm normally utilised in hindsight experiments. This second stage consisted of a time delay, after which subjects were reminded of the earlier experimental descriptions and were given long-term outcome information (whether the original study was replicated or not). They were then required to remember their original quality judgements. In addition, they were also asked to recall the numbers of subjects and the significance levels used in these studies. It was predicted that memories of original judgements would be biased in line with the nature of the long-term outcome information. Thus, where an experiment had been replicated, memories would overestimate earlier quality judgements; this would be the similar to the effect normally reported in hindsight bias studies. Where an experiment had failed in replication it was predicted that memories would underestimate earlier judgements.

It has been suggested earlier that outcome bias effects may be extremely pervasive. If this is the case then it is possible that memories of any other relevant factor will also be biased in a direction which supports the subjects' (biased) view of the quality of that experiment. Thus, in the present study it was also predicted that memories of the other experimental factors (subject numbers and significance levels) would be biased in line with the nature of the long-term outcome. This would result in memories of increased subject numbers and smaller significance levels where outcome information showed experiments to have been replicated and memories of reduced subject numbers and larger significance levels where experiments were not replicated.

### 5.3.2 Method

Subjects

Subjects were 60 second and third year psychology undergraduates. All subjects were paid £1.50 for participation. Subjects were allocated to experimental conditions by random distribution of question sheets; ensuring only that equal numbers of each type of question sheet were distributed. All responses were anonymous.

Materials

Each subject received an initial instruction sheet and four brief experimental descriptions which they were asked to rate for quality. These descriptions were given somewhat striking titles and subject matter to enable subjects to clearly differentiate them in the memory stage of the experiment. Subjects were then presented with a filler task consisting of a number of questions in which they were required to estimate the frequency of various features in the environment; this task had no relevance to the experimental descriptions or tasks. In the second part of the experiment, subjects were given new instructions containing reminders of the original descriptions with additional outcome information and a series of questions relating to their memory of the initial experimental descriptions and their initial responses.

An example of the initial subject instructions and task is shown below;

INSTRUCTIONS

You will be presented with four descriptions of experiments. You should read each description carefully and then make a judgement of the quality of each experimental design.

**An experiment of high quality is one which has a high probability of finding an effect if one exists without incurring excessive cost.** Quality judgements should be recorded by marking an X at the relevant point on the line under each description.

When you have completed this part of the experiment there will be a short break in which you will be set other tasks. After these are complete you will be given a question sheet to test your memory of the four experimental descriptions. If you have any questions please ask the experimenter before you start the task.

**Drug test**

Stephens and Bell (1989) performed an experiment in which the drug 'Epsadrine' was predicted to reduce blood pressure in human subjects. The subjects (twenty eight volunteers from Chicago medical school) were split into an experimental group and a control group. The experimental group were each given a five millilitre dose of the drug intravenously. The control group recieved a placebo. Subsequent blood pressure measures were taken two hours later. A significance level of $p < 0.01$ was set. A t test showed a significant difference between the groups. The hypothesis was supported.

Low Quality _____ High Quality

**Teddy bears**

An experiment was reported in which Goodburn, Jones and Larch (1991) tested the hypothesis that children who owned teddy bears would more readily talk to an adult if their teddy bear (an object of safety) was present. A total of forty two pre-school children (aged between three and four) were tested by means of a structured interview. Half the children were interviewed with their teddy bears present. The other half acted as a control group and were interviewed without their teddy bears. A significance level of $p < 0.05$ was set. A t test on the number of answers given was significant; supporting the hypothesis that children with bears present did talk more readily.

Low Quality _____ High Quality

**Bicycle riding**

Naesmith and Barton (1972) performed an experiment on the acquisition of motor skills. They tested the hypothesis that regular cyclists would learn the skills required for a complex balancing task faster than non-cyclists. Subjects were thirty four psychology undergraduates (half cyclists, half non-cyclists). Measurement involved the number of attempts (up to a maximum of ten) needed for the successful completion of the balancing task. A significance level of $p < 0.01$ was chosen. A t test showed **no** significant difference in task learning. The hypothesis was not supported.

Low Quality _____ High Quality

**Depression**

A study by Whittaker et al. (1988) tested the effects of a behavioural regime on clinically depressed patients. Subjects were thirty eight adult patients in the long stay ward of a Manchester hospital. All subjects had been clinically depressed for more than two months. The regime was given to half the subjects and involved positive reinforcement for a number of positive social behaviours. The other half of the subjects acted as a control group. A significance level of $p < 0.05$ was chosen. A t test showed **no** significant difference between the groups. The hypothesis was not supported.

Low Quality _____ High Quality

An example of subject instructions and tasks for the second (memory) phase of the experiment is shown below;

**Part 2 INSTRUCTIONS**

On this sheet are brief reminders of the four experiments which were presented earlier. And some additional information on the long term outcome of these studies.

You should try in each case to remember your original judgement of quality and mark an X at the same point on the line under each description. Please also try to answer the questions from memory. If you cannot remember please make your best guess. It is important to answer every question.

If you have any questions please ask the experimenter before you start the task.

**Bicycle Riding**

Naesmith and Barton (1972) found **no** significant differences when testing the hypothesis that regular cyclists would learn the skills required for a complex balancing task faster than non-cyclists. Although the hypothesis was not supported subsequent replications of this experiment have shown this original result to have been false. There was an effect which this experiment missed.

**What was your original quality judgement ?** Low _____ High Quality Quality

**How many subjects were used in this experiment ?** _____

**What level of significance was set ?** _____

**Teddy bears**

Goodburn, Jones and Larch (1991) found a significant result supporting the hypothesis that children who owned teddy bears would more readily talk to an adult if their teddy bear (an object of safety) was present. However subsequent replications of this experiment have shown this original result to have been false. There was no effect.

**What was your original quality judgement ?** Low _____ High Quality Quality

**How many subjects were used in this experiment ?** _____

**What level of significance was set ?** _____

**Depression**

A study by Whittaker et al. (1988) found no significant improvement in depression when testing the effects of a three month behavioural regime on clinically depressed patients. The hypothesis was not supported and subsequent replications of this experiment have shown this original result to have been correct. There was no effect.

**What was your original quality judgement ?** Low _____ High Quality Quality

**How many subjects were used in this experiment ?** _____

**What level of significance was set ?** _____

**Drug test**

Stephens and Bell (1989) performed an experiment in which the drug 'Epsadrine' was shown to significantly reduce blood pressure in human subjects. The hypothesis was supported and subsequent replications of this experiment have shown this original result to have been correct. There was an effect.

**What was your original quality judgement ?** Low _____ High Quality Quality

**How many subjects were used in this experiment ?** _____

**What level of significance was set ?** _____

Design

In the first part of the experiment, four experimental descriptions were rated by the subjects in terms of their quality. These ratings were made were made by marking a point on a line with a range from low quality at one end to high quality at the other. This change to an analogue response was made to prevent the second memory stage becoming too easy. As in Experiments 3, 4 and 5 quality was rigidly defined in the experiment instructions (see above). The fictitious experimental descriptions presented to subjects varied in terms of three independent variables:

1. Pre-set significance level - With two levels (0.05 and 0.01)
2. Subject numbers - With four levels (28,32, 38 and 42)
3. Experimental result - With two levels (Significant or not significant)

Levels of these factors were varied across question type using a greco-latin square creating four different question presentations, one example of which is shown above. A fully balanced design would have required sixteen questions and this would have made the memory stage of the experiment excessively difficult.

The second (memory) stage of the experiment consisted of brief reminders of the original experimental descriptions. Reminders consisted of the original experimental title followed by a description of the experiment including the names of the experimenters, the hypothesis which they had tested and the original result. To this was added outcome information describing whether the original experimental result had been replicated or not. This outcome information created a fourth independent variable; replication, with two levels (replicated and not replicated). This factor was also counterbalanced across question type.

There were four dependent variables. In the first stage; the initial quality judgements. In the second (memory) stage; the remembered quality judgements, the remembered significance levels and the remembered subject numbers.

Procedure

After a brief introduction explaining that the experiment was in three parts subjects

were presented with the first instruction sheet and task. Subjects were given five minutes to read the descriptions and complete the four quality judgements. They were informed when they had one minute remaining. At the end of this time the instruction and description sheets were removed. Quality ratings were recorded on the task sheets by marking a cross on a line the extremes of which were low quality at one end and high quality at the other.

Subjects were then presented with a filler task and told they had ten minutes to complete it. At the end of this time these sheets were removed and the part two instructions and question sheets were presented. There was no time limit for this section. Subjects were reminded that it was important to fill in all the sections on the sheet and left to complete it in their own time. Remembered quality ratings were recorded by marking a cross on a line identical to that in the first part of the experiment.

Subsequent scoring of quality judgements was by measuring the distance of crosses along each line and converting to a 0 to 100 point scale. Blind scoring was achieved by scoring remembered quality judgements without reference to original judgement scores which could only be compared by matching question types and answer sheet and question sheet identifying codes.

### 5.3.3 Results

Initial stage

In the first stage of the experiment subjects were required to judge the quality of the fictitious experiments. An analysis of the effects of pre-set significance level, subject numbers and experimental result on these judgements of experimental quality was performed. An analysis of variance showed only initial outcome information (experimental result) to have a significant effect on quality judgements ($F = 9.43$, $p < 0.01$). Where the fictitious experiments were described as having achieved a significant result, they were rated as having higher quality than those described as having no significant result. This difference is shown in Table 5.2.

Table 5.2

Mean initial quality judgements
(Standard deviations in brackets)

| Initial outcome | Mean quality |
|---|---|
| Significant result | **56.6** (25.9) |
| No significant result | **46.5** (24.9) |

Memory stage

Three analyses were completed on the data from the memory stage of the experiment.

In order to analyse differences between subjects' remembered significance levels and the actual levels, values were categorised and actual levels were subtracted from remembered levels. This created a scoring scheme where an actual level of 0.05 remembered as 0.01 would score +1. An actual level of 0.01 remembered as 0.05 would score -1, and a correct remembered level would score 0. Analysis of these scores against both initial experimental result and hindsight information (replication) showed no significant effects.

Memory for the numbers of subjects described in the fictitious experiments was also analysed against both initial experimental result and long-term outcome information. In this case the difference between the number of subjects remembered and the actual numbers of subjects described in the fictitious experiments was the dependent variable. The main effect of long-term outcome information was significant ($F = 4.21$, $p < 0.05$)[$r = 0.12$]. Where subjects were told in hindsight that an experiment had been replicated they remembered the experiment to have had a higher number of subjects. This difference is shown in Table 5.3.

Table 5.3

Mean difference between remembered and actual subject numbers
(Standard deviations in brackets)

| Long-term outcome | Mean difference |
|---|---|
| Replicated | **+ 4.5** (16.2) |
| Not replicated | **+ 0.7** (15.0) |

The differences between original and remembered quality judgements were also analysed against both initial experimental result and long-term outcome information. Again the main effect of long-term outcome information was significant ($F = 8.32$, $p < 0.01$)[$r = 0.16$]. Where subjects were told in hindsight that an experiment had been replicated they remembered their original quality judgement to have been higher than it was. When an experiment was not replicated they remembered a lower quality rating. This difference is shown in Table 5.4.

Table 5.4

Mean difference between remembered and actual quality judgements
(Standard deviations in brackets)

| Long-term outcome | Mean difference |
|---|---|
| Replicated | **+ 1.4** (5.1) |
| Not replicated | **- 0.6** (6.9) |

For all Anova tables and complete tables of means see Appendix 6.

### 5.3.4 Discussion

The results of Experiment 7 demonstrated that the effects of outcome information on quality judgements were still present in a memory based paradigm. In addition it was shown that outcome information biased not only memories of subjects' original judgements but also biased memories of other relevant factors in hindsight. These findings strongly supported the predictions made in the introduction (Section 5.3.1) based on the view of outcome bias as a pervasive effect of which hindsight effects are a sub-set.

The analysis of subjects' quality judgements of the initial experimental descriptions showed a biasing effect of the experimental result. This was a small effect and mirrored the effects of initial experimental results seen in the earlier experiments. Unlike the earlier experiments, in this analysis none of the relevant experimental factors (subject numbers and significance levels) had a significant influence on quality judgements. This was not

surprising in the case of subject numbers. In previous experiments differences in sample sizes were large. In this experiment these differences were much smaller in order to prevent the memory stage of the experiment from being too simple. In the case of the different significance levels presented in the fictional experimental descriptions, there was no clear reason why this factor should have failed to have any significant influence on subjects' original quality judgements. However, this initial stage of the experiment was not fully counterbalanced within subjects and as a consequence the resulting analysis was not particularly powerful.

In the memory stage of the experiment the predicted effects of outcome information on the memory of earlier quality judgements were demonstrated. Where outcome information was positive (replication) earlier judgements were overestimated in a similar manner to traditional hindsight bias. Where outcome information was negative (failure to replicate) earlier judgements were underestimated. Thus, the 'I knew it all along" effect was polarised into 'I knew it was good all along' or 'I knew it was bad all along' depending on the nature of the long-term outcome information. These biases of memory are all the more notable when it is remembered that subjects only had a time delay of ten minutes before the recall task.

These effects were smaller than those seen in the on-line outcome bias paradigm used in Experiments 3, 4 and 5. This difference in effect size is consistent with the proposed mediating influence of outcome salience (see Section 4.5.2). In these earlier experiments there were high levels of outcome salience clearly stated in financial terms. In contrast in the present experiment there was no clearly stated outcome salience beyond subjects' desire to be seen as correct in their judgements. This is comparable with the majority of hindsight bias studies; as are the smaller effect sizes.

Subjects' memories of their own earlier judgements were not the only factor to be biased in hindsight. Memory for sample sizes was also influenced by the nature of long-term outcome information. Where an experiment was shown to have been replicated subjects remembered it as having had larger subject numbers. Thus it seems that subjects reconstruct their memories in line with their current knowledge; if an experiment was good then it must have had a large number of subjects. Again this suggests more widespread and

pervasive effects of outcome information than those originally reported in outcome bias studies. A similar effect of reconstructive remembering of the scientific literature has been noted by Vincente and Brewer (1993), potential links between these effects will be discussed in Chapter 6.

There was no such effect for subjects' memories of significance levels. This was not surprising as only the two standard levels, $p < 0.05$ and $p < 0.01$ were used in the fictitious experimental descriptions. This made the memory stage of the experiment rather simple. In a more extensive study it would have been possible to present various exact levels (such as $p = 0.037$) in which case the memory task would be difficult enough to demonstrate an effect.

## 5.4 SUMMARY OF CHAPTER FIVE

Chapter 5 reported two experiments which served to clarify a number of questions arising from outcome bias effects seen in the previous experiments. Experiment 6 explored subjects' beliefs about the relevance of information to prospective quality judgements. The results demonstrated that subjects were clearly aware that information relating to experimental factors was more relevant than that relating to outcomes when judging the quality of experiments. Further, they were aware that financial outcomes should not be relevant to these judgements. This result supported the view that the effects of outcome information seen in earlier experiments, were due to an unconscious bias in which subjects were unable to ignore information that they knew to be irrelevant.

This experiment also included a manipulation of subjects' point of view. In one case subjects were asked how relevant factors would be to their judgements of the quality of someone else's experiment. In the other case subjects were asked how relevant factors should be to someone else's judgement of the quality of their experiment. This difference in perspective had a marked influence on subjects' ratings of the relevance of financial outcome information. Subjects clearly believed that others should not use financial outcome information to judge their experiments. When subjects were judging the experiments of others they were not so adamantly against the use of financial outcome information. It was suggested that this effect was based in some form of self-centred information processing

which might also be the underlying cause of outcome biases.

Experiment 7 linked the on-line paradigms used in outcome bias studies with the memory based paradigms usually associated with hindsight studies. In this experiment initial outcome information (experimental result) was presented along with fictitious descriptions of experiments. Subjects were required to judge the quality of these experiments. After a time interval subjects were reminded of the original experimental descriptions and given long-term outcome (replication) information. They were then required to remember their original quality judgements and the number of subjects and the significance level associated with the former experimental descriptions.

Results showed outcome bias related to the on-line initial outcome information and hindsight bias related to the long-term outcome information. This hindsight effect was shown to be dependent on the nature of the long-term outcome information. Positive information led to overestimation of earlier quality judgements whilst negative information led to underestimation of earlier quality judgements. This effect mirrored the dependency of outcome bias effects on the nature of the related outcome information seen in Experiments 3, 4 and 5.

The memory stage of this experiment demonstrated other effects beyond those usually reported in hindsight bias studies. Memory for sample sizes was also shown to be biased by the nature of long-term outcome information presented in hindsight. Where a study was shown to have been replicated subjects remembered it as having used a higher number of subjects. These results confirmed the pervasive effects of outcome information, not only on original judgements, but also on associated memories where the outcome was presented in hindsight.

**Table of contents - Chapter six**

# CHAPTER SIX

## Discussion and Conclusions

## 6.1 INTRODUCTION

This final chapter begins with a review of the results of the present research followed by a discussion of some of the direct practical implications of these findings. Specific implications for the design process are discussed as are more general implications for any judgements which are made in the presence of outcome information. It is shown that the present findings also allow for quite specific predictions relating to areas in which future research would be most beneficial.

Unlike previous hindsight and outcome bias research the present research is able to directly demonstrate the influences of motivational factors in biases related to the presence of outcome information. A definition of these motivational factors in terms of outcome salience is presented and its role as the determining factor in the size and direction of the resulting biases is discussed.

The results of the present research suggest an account in which the previously separate fields of hindsight and outcome bias are unified. To this end an overview of hindsight and outcome biases is presented. In the light of the present results and by accounting for methodological differences, this overview argues that hindsight effects are a sub-set of outcome bias effects.

In addition to unifying these previously separate research fields this account is also able to clarify previously unexplained effects from the hindsight literature. The next section analyses these effects in terms of the concept of outcome salience generated by the present research. The differences explained in this section include variations in effect size across different paradigms, the differential effects of negative hindsight information and reverse hindsight effects.

Potential relationships between the concept of outcome salience and the concepts of Subjective Expected Utility and relevance are discussed. The following section considers

the wider theoretical implications of the present experimental results. The implications of these results for theoretical accounts of reasoning in general and hindsight bias in particular are explored.

## 6.2 A SUMMARY OF THE PRESENT RESEARCH

The first experiment in this series was a prospective study of the design of psychology experiments. Subjects experienced in the design of psychology experiments were given the task of designing experiments to test a variety of hypotheses and their subsequent verbal protocols were recorded. It was the intention of this experiment to explore the extent and nature of subjects' evaluations of the quality of their putative designs. The results showed the clear failure of subjects to consider explicit evaluation factors during the design process.

Despite the fact that Experiment 1 demonstrated a lack of explicit evaluation in the design process, it was clear that researchers do routinely make evaluations of the quality of experimental designs in published research. Thus, researchers must be capable of some form of design evaluation even if this evaluation is not explicit in the design process. Experiment 2 was designed to explore the factors influencing these judgements of existing designs. A number of descriptions of experiments were presented to subjects. These varied in terms of the number of subjects used and the pre-set significance level. As experimental results are necessarily present in published research, the experimental results were also presented in these descriptions. The inclusion of this outcome information led to the probability that subsequent quality judgements would be biased in line with this outcome. To further explore the effect of outcome information half the subjects were given long-term outcome information in terms of replication (or non-replication) of the initial result. In addition this group also received financial outcome information. The results of Experiment 2 showed little understanding of the statistical principles of good design. Long-term outcome was the most influential factor in judgements of quality.

This extremely influential form of outcome bias was further explored in subsequent experiments. Given the possibility that the results of Experiment 2 were influenced by the

obvious statistical complexity of the task, Experiment 3 presented subjects with an equivalent task based on a different scenario. This scenario was based on a team of fraudsters who were testing roulette wheels to try to discover which of them were biased. Thus subjects were presented with a task closely related to that of judging the quality of scientific experiments without cueing the specific statistical nature of the problems. In addition this scenario presented a specific definition of experimental quality and emphasised the need for subjects to take account of sampling costs. Experiment 3 produced results directly comparable with those of Experiment 2 despite these changes in presentation. Again, outcome information had the greatest influence on subjects' judgements of quality.

Experiment 4 replicated the long-term outcome condition of Experiment 3 using a more statistically expert subject group than that used in either of the previous experiments. This made no appreciable difference and again the same pattern of results was reported. The experimental paradigm used in Experiments 3 and 4 demonstrated consistent and high levels of outcome bias. This paradigm had a number of advantages over the paradigms in which hindsight and outcome biases had been previously reported. Most importantly the outcome information had no direct relevance to the judgement required, unlike all previous hindsight studies. Also in this paradigm, all the information which was relevant to the judgement was available to the subjects. The presence of bias under these conditions argued strongly against the theory that hindsight bias was based on the rational inclusion of any relevant information which becomes available.

In Experiments 2,3 and 4 the interaction between initial and long-term outcome information led to a specific pattern of results. This persistent pattern could not be explained by differences in the nature of each possible outcome alone. It was hypothesised that both the size and direction of these outcome biases were mediated by the salience of different outcomes from the point of view of the subject, particularly in terms of their financial implications. Experiment 5 tested this hypothesis by the introduction of a perspective change. In this paradigm it was possible to manipulate the financial salience of particular outcomes by changing the point of view of the subjects In this experiment one group of subjects were given the fraudster perspective used in the long-term outcome groups of Experiments 3 and 4. Another group of subjects were given the same problems presented

from a manufacturers' perspective. This manipulation had the advantage of differentiating between the actual nature of a given outcome and its financial implications from the point of view of the subject. Results for the group given the fraudster perspective again replicated the pattern of results seen in the previous experiments. The group given the manufacturer perspective showed a different pattern of results related to the influences of the interaction between initial and long-term outcome information. As predicted these results showed the size and direction of outcome biases to be largely dependent on outcome salience. Unlike previous outcome bias studies this result clearly demonstrated the role of motivational factors in this form of bias.

It was suggested that one of the advantages of the experimental paradigm used in Experiments 3,4 and 5 was the irrelevance of the outcome information (and its financial implications) to the judgement required. However, the conclusion that subjects' responses are based on an unconscious bias depends on whether subjects believe this information to be irrelevant to their decision. Experiment 6 explored subjects' beliefs about the relevance of this information to prospective quality judgements. The results demonstrated that subjects were clearly aware that information relating to experimental factors was more relevant than that relating to outcomes when judging the quality of experiments. Further, they were also clearly aware that financial outcomes were the least relevant factor in these judgements. Overall subjects' ratings of relevance for the factors which had been varied in the earlier experiments were diametrically opposed to subjects' actual use of these factors in their earlier quality judgements. This result supported the view that the effects of outcome information seen in earlier experiments were likely to be the result of implicit Type 1 processes in which subjects were unaware of the specific influences of information on their judgements.

This experiment also included a manipulation of subjects' point of view. Subjects were either asked how relevant factors would be to their judgements of the quality of someone else's experiment, or subjects were asked how relevant factors should be to someone else's judgement of the quality of their experiment. This difference in perspective strongly influenced subjects' ratings of the relevance of financial outcome information. Whilst subjects believed that others should not use financial outcome information to judge their experiments, they were not so averse to using this information when judging the

experiments of others. Again personal motivation was shown to influence judgement.

The outcome bias demonstrated in the earlier experiments was a result of on-line presentation of outcome information. Both the size of the effect and its direction were shown to be dependent on the salience of the outcome from the point of view of the subject. It was proposed that this outcome bias was pervasive and would affect any related judgement or memory. Thus the only difference between this effect and the effect traditionally described as hindsight bias was that in hindsight bias studies outcome information was presented after the event and the task was to remember an earlier judgement. Experiment 7 tested this hypothesis. In this experiment subjects were required to judge the quality of four fictitious experiments. After a time interval they were given long-term outcome (replication) information and were required to remember their original quality judgements. They were also asked to remember the numbers of subjects and the significance level in the former experimental descriptions.

Results showed the predicted hindsight effect, the direction of which was dependent on the nature of the long-term outcome information. Positive information led to overestimation of earlier quality judgements whilst negative information led to underestimation of earlier quality judgements. This effect mirrored the dependency of outcome bias effects on the nature of the related outcome information. This experiment also demonstrated more general effects of outcome bias on memory. Where a study was shown to have been replicated subjects remembered it as having used a higher number of subjects. These findings supported the hypothesis that hindsight effects were a particular subset of outcome effects.

## 6.3 PRACTICAL IMPLICATIONS OF THE PRESENT RESEARCH

The present experimental results include a number of findings which have not previously been demonstrated in this research field. The paradigm used in this research was able to demonstrate that the presence of irrelevant outcome information will influence judgements of the quality of experimental designs. Also an irrelevant outcome presented in

hindsight was shown to influence the memory of earlier judgements. This differs from previous hindsight and outcome research in which the outcome information presented to subjects was relevant to their memory or judgement task. In addition the present research has demonstrated that the size (and direction) of the resulting bias will depend on the salience of the outcome from the point of view of the subject. Overall the effects of the presence of outcome information were shown to be extremely pervasive and inevitably linked to motivational factors. These findings have important practical implications.

One of these implications is that the bias demonstrated in a laboratory setting will always be smaller than a bias resulting from a real life situation. As the effect demonstrated in a given experiment depends on the salience of the outcome information used, its size will always be limited by the hypothetical nature of the scenario. For example, in the present research, telling subjects that they would gain a large amount of money from a particular outcome generated effect sizes twice as large as those resulting from presenting outcome information alone. These effects would presumably be even larger if the subjects had actually received the money. Actual success or failure would necessarily produce stronger motivation than the hypothetical case. These results completely refute the arguments of Christensen-Szalanski and Willham (1991) who suggest that hindsight bias is of no practical significance. (Section 3.4) Their argument is based on an average effect size generated from a meta-analysis of hindsight studies in which the salience of outcomes was not considered. The present findings enable the prediction of situations in which the effects of outcome information will be very large.

It has also been argued that the fact that people believe that they "knew it all along" merely creates a feeling of overconfidence which has few damaging practical implications. The present results have shown biases in judgements which go far beyond the simple "knew it all along effect". For example, biases in on-line quality judgements have been demonstrated at the same time as additional biases in the memory of those quality judgements and biases in the memory of factors relating to those judgements. This creates a situation in which a biased judgement will be supported after the event by a biased view of the facts.

This tendency to bias the memory of experimental details in line with a current

viewpoint explains the mechanism behind the reconstructive remembering reported by Vincente and Brewer (1993). When a researcher refers to a famous experiment from the literature, the memory of the details of that experiment will be biased in line with the belief that it was an experiment of high quality.

### 6.3.1 Implications for the design process

In general the effects of outcome bias demonstrated in the present research imply a situation in which researchers will be unable to learn effective experimental design from experience. The results have shown outcome information to be the overriding factor in judgements of design quality. Where initial outcome information was present these judgements were biased in favour of significant results. Where subsequent long-term outcome information was present judgements were biased in favour of desirable outcomes.

Judging the quality of research studies by their long-term outcomes, although unreasonable in individual cases, will lead to an accurate view of correct quality criteria over time. Over a large number of experiments the long-term outcomes will tend to reflect the true quality of the experimental designs. However, long-term outcome (replication) information is very rarely available and judgements based on the initial experimental result will be erroneous. In an individual case the experimental result may be the result of Type 1 or Type 2 error or may truly reflect the state of the world. Given that there is no objective measure of the truth of a given hypothesis; there is no way to tell whether a significant or a non-significant result reflects the truth or experimental error. This would suggest that outcome bias in experimental design will affect the ability of designers to learn accurately to judge quality even if they have extensive experience. Outcome bias results in the continuing belief that the only good experiment is one with a significant result. This belief can be seen in the literature where experiments with non-significant outcomes are almost never published. Thus the literature itself presents a biased range of experiments. This leads to a situation where the effects of insufficient experimental power in design (which inevitably leads to non-significant results) will not be reflected in this literature. This biased presentation of experiments will create a belief that experimental power is not an important issue as every example in the literature has sufficient power. This effect can be seen in

Experiment 1 where there was little or no consideration of power. It is also entirely consistent with the generally inadequate power of experiments reported by Cohen (1962) and Sedlmeier and Gigerenzer (1989) (See Section 1.8.5).

One of the most important of the present findings is that outcome bias is not dependent on the nature of the outcome itself (in this case replication or non-replication of the initial experimental result). Rather it depends on the salience of the outcome from the point of view of the subject. This does not imply particular problems in terms of the design of psychology experiments where the designers are not generally subject to financial motivations, although any design fault which leads to increased possibilities of successful outcome (such as biased sampling) may be less likely to be noted. However, this finding does suggest the probability of severe biases in those design areas where either the designer or the final arbiter of design quality is subject to financial motivations. A similar outcome bias effect in these cases would lead judges to conclude that the design which was most financially beneficial to them was in fact the design with the highest quality. In architecture or civil engineering, for example, this potential bias would lead to the acceptance of designs for relatively cheap buildings or bridges which were potentially inadequate. It should be noted that this would not be a cynical conscious decision by designers to line their own pockets at the expense of adequate quality. This self-motivated outcome bias is unconscious and, as such, all the more difficult to deal with. The present research demonstrates that a change of perspective will in some cases reverse the effects of outcome bias and thus it may be possible to debias subjects by requiring them to consider a number of different points of view before making a quality judgement. The problem with this potential debiasing approach is that the personal salience of a hypothetical perspective would always be considerably less than the salience of a real financial outcome. Thus in real life situations the only possibility for removing bias is to ensure an independent judge of quality who has no personal interest in the outcome.

### 6.3.2 Implications for quality judgements where outcomes are known

From the present research and from the outcome bias literature it would appear that any relevant judgement made after an outcome is known will be influenced by that outcome. A

number of cases have been highlighted in the literature where this form of outcome bias has potentially damaging effects. These include judgements of the blame attributed to victims (Janoff-Bulman et al. 1985), the quality of consumer goods (Mazursky and Ofir, 1990), the appropriateness of others' decisions (Lipshitz, 1989), evaluation of performance (Mitchell and Kalb, 1981) and the quality of monetary gambles and medical decisions (Baron and Hershey, 1988) to name but a few.

The present findings show that the nature of the influence of a known outcome will be dependent on the personal salience of that outcome. Thus the amount of bias present in any of these situations will depend on the personal motivations of the judge. This factor has not been directly manipulated in previous research. The present findings enable the prediction of quite specific situations in which outcome bias will have a major influence, for example, judgements of the quality of investment decisions which have a direct financial result. It is also possible to predict more biased medical decisions where the decisions have financial consequences for the judge, for example, where a doctor is responsible for a medical budget or alternatively is paid for specific treatment. These predictions clearly suggest areas in which further research will be most beneficial.

Practical methods of reducing bias are also suggested by the present research. Clearly where it is not possible to conceal the outcome from the judge until after the judgement has been made, the best method of debiasing is the introduction of an independent judge who could be told the facts without being told the outcome. Where neither of these solutions is possible bias can still be reduced by ensuring that a judge has the minimum of personal motivation to favour any particular outcome.

## 6.4 MOTIVATIONAL FACTORS AND OUTCOME SALIENCE

The present research differs from previous hindsight and outcome research in two important ways. Firstly, the outcome information presented to subjects was not relevant to the required judgement and all the information necessary for the judgement was presented in the task. Secondly, a potential motivational factor, the financial implication of the outcome,

was made explicit and systematically varied within the experiment. This manipulation enabled a differentiation to be made between effects which were due to the nature of the outcome and those which were due to the implications of that outcome. These differences led to results which showed that even when subjects are aware that the financial implications of outcome information are the least relevant factor they still have the greatest influence on their judgements.

Subjects' point of view was found to be easily manipulated. Although subjects were not able to ignore irrelevant outcome information, they were seen to reverse the effects of this information when another point of view was suggested to them. By manipulating subjects' points of view in this way it was possible to separate the nature of an outcome from its financial implications. Thus, the size and direction of the resulting bias was shown to depend largely on the implications of the given outcome from the point of view of the subject. These findings clearly demonstrate the central role of motivational factors in hindsight and outcome biases; a position which had been suggested by previous research but never directly tested. In the light of this important finding it is necessary to discuss general motivational accounts and the present motivational factors in some detail.

### 6.4.1 Motivational accounts of hindsight bias

Motivational accounts of hindsight bias are not in competition with information processing accounts. Rather, they should be seen as an attempt to define the adaptive advantages which are served by the type of information processing which leads to bias. The present research, by explicitly defining motivational factors, has demonstrated their overriding importance in determining the size and direction of biases relating to outcome information. Previous hindsight and outcome bias studies have not explicitly defined these factors, nevertheless, their influence may still be inferred from previous research.

Campbell and Tesser (1983) describe motivational accounts as those in which hindsight bias serves to satisfy some basic human needs. They suggest two basic types of motive which may be influential; the predictability motive and the motive to maintain self-evaluation. In a correlational study they report positive associations between measures related to both these motives and the amount of hindsight bias exhibited.

The predictability motive derives from a basic human desire for control in one's interactions with the environment. (White, 1959, Wortman, 1976) This basic need is reflected in numerous psychological effects. Campbell and Tesser (1983) quote; "the desire for certainty (Brim and Hoff, 1957), the need to know and to be able to predict the environment (Kelly, 1971, Kelly, 1955, Pervin, 1963) and the need to experience an integrated and meaningful world (Cohen, Stotland and Wolfe, 1955)" (Page 607) In support of these examples is the whole literature on the illusion of control. Discussing the illusion of control Langer (1975) reports "While people may pay lip service to the concept of chance, they behave as though chance events are subject to control". This has been a familiar phenomenon in gambling research since both Goffman (1967) and Henslin (1967) reported subjects' apparent belief in their ability to control random events whilst gambling. These influences lead to a position where people are reluctant to believe that chance has influenced an outcome, particularly where the consequences of the outcome are severe (Lerner and Miller, 1978). This not only suggests a motivation to believe in hindsight that the given outcome was predictable, it also explains the increased effect sizes in outcome bias studies where the consequences of outcomes are severe (present research, Mitchell and Kalb, 1981). The need to maintain public or private self-evaluation is also a possible motivational factor in hindsight bias. Fischhoff (1975) recognises this factor when he points out that it is flattering to the subject to believe that they knew it all along. Both the concept of predictability and that of self-evaluation are vague. Whilst predictability is based on the need for control it may also reflect elements of just world theory. This would be particularly applicable in outcome bias studies where there might be a tendency to believe that a "good" outcome must have resulted from good causes and vice versa. This is to some degree suggested by studies in the outcome bias literature where the attribution of blame is exaggerated by strong negative outcomes (Mitchell and Kalb, 1981, Janoff-Bulman and Timko, 1985).

Up to now the evidence linking these forms of motivation with hindsight bias has been correlational. In addition, the majority of these proposed motivational factors are vague and cannot be easily quantified. The present research has added the factor of financial gain to these motivational influences, and by manipulating this factor has experimentally demonstrated its influence as a defining factor in subsequent bias.

### 6.4.2 The role of outcome salience

Outcome salience was first defined in Section 4.4.1 as "the relative importance of the implications of an outcome from the point of view of the subject". It was noted that outcome salience was related to the subjective utility of a given outcome and could vary in size in both positive and negative directions. In those experiments where outcomes had direct financial implications, positive or negative outcome salience was strongly related to these gains or losses of money. Although, in these experiments, financial outcome was the major factor influencing outcome salience, other motivational factors relating to outcome information must also be taken into account. Motivational accounts of hindsight bias have previously focused on factors relating to the personality of the subject. The more commonly proposed motivational factors were reviewed in the previous section, these factors are concerned with the need to maintain public or private self-evaluation and the need for a sense of control. Outcome salience must include these factors where a judgement based on the implications of a given outcome will serve to support these motivational needs.

In the present experiments subjects were given a hypothetical role (Fraudster or roulette wheel manufacturer) which was related to a clear hypothetical goal, to make (or avoid losing) money. This manipulation artificially created a motivational factor to which the outcome information was directly relevant. This creation of a hypothetical goal may not have been entirely necessary to demonstrate the influence of the resulting outcome salience. The adoption of this goal without the need for role definition was shown in Experiment 2. This experiment had a scenario based on drug companies testing new drugs. In this case subjects were not given a particular role but simply asked to judge the quality of the sampling procedures these companies had used. Even from the point of view of an objective outsider, subjects were still biased in favour of procedures which had made money for the companies. This result suggests that subjects will adopt a goal of success in general, even where no personal success is implied. In these experiments the hypothetical implications of the outcome can be seen to interact with the subjects' personal motivations. From the subjects' point of view, the outcome now presents a measure of personal success or failure. In this case a personal motivation such as the need for control in general would be reflected

in the need to control your own (hypothetical) success. A motivation for successful self-evaluation could also be reflected in this direct measure of hypothetical success or failure.

By manipulating this form of goal relevant outcome information in the experimental paradigm, the present experiments have shown the resulting outcome salience to have a systematic and therefore predictable influence. This finding implies a model of bias in which the extent to which outcome information satisfies goals based on a subject's motivation (outcome salience) determines the relevance and the weighting of that information in the subsequent judgement. Thus, this motivational influence may serve to explain previously unexplained differences in the literature. The next section explores some of the factors related to outcome salience and reviews previous hindsight research taking these factors into account.

In order to maintain an overall view of subjects' behaviour in these experiments it is necessary to remember that there will be a percentage of subjects whose judgements are either not biased or only biased to a small degree. The bias arising from outcome salience is only related to those motivations which are influenced by the presence of outcome information. Other motivations which act in opposition to these biasing motivations will also be present in any experimental paradigm, for example, motivations to form correct and fair judgements.

### 6.4.3 Factors relating to outcome salience

The present results have shown that the size and direction of a given bias is largely dependent on the salience of the outcome from the point of view of the subject. Following these findings, if outcome salience can be reliably estimated in any given study, the nature of the resulting bias can be predicted. Unlike the present research, in the majority of hindsight and outcome studies the implications of outcome information are not presented explicitly. Despite this there is no reason to suppose that the effects of outcome bias are limited to those studies which explicitly define the implications of outcomes. Outcome salience has been defined as "the relative importance of the implications of an outcome from the point of view of the subject". Using this definition it is possible to form a gross estimate of outcome salience in any experiment by considering the implications of a given outcome in respect of a

subject's motivations.

A typical hindsight bias experiment provides an example where implications are not explicitly defined. In this type of experiment the subject is required to remember an earlier likelihood judgement after being told the 'real' outcome. There is no clear benefit or loss in the content of this memory task, nevertheless in the experiment as a whole, two contrasting motivational factors can be predicted based on the motivational accounts presented in Section 6.4.1. Where the earlier judgement can't be remembered accurately (and these experiments are specifically designed to make the memory tasks difficult to avoid ceiling effects) there is a certain amount of leeway in responses. In this case a subject who biases their responses in favour of the 'true' outcome achieves a number of benefits. The biased subject is able to believe that they are usually right when predicting future events, resulting in a positive sense of self-evaluation. In the case of almanac questions a bias towards true outcomes enables the subject to believe that they are knowledgeable. Also in the case of real life events correct prediction will also support a world view that in general events are predictable. This belief will support a subject's motivation for a sense of control in their environment. Conversely an unbiased subject has to cope with the conclusion that they are less knowledgeable than they thought or that events in the world are less predictable than they supposed.

The motivational factor of a sense of control over the environment can also be seen in the light of just world theory (Lerner and Miller, 1978). It is clear that people prefer to believe in a world where chains of cause and effect are predictable and that consequences are not random. This belief is ecologically necessary in the drive to explain the world; a person who believed that events occurred at random would have no motivation to practice inductive or deductive reasoning. There is also the related question of the reversibility of cause and effect. It seems to be the case that people believe firstly, that any effect must have a definable cause (can't be random) and secondly, that if the effect is beneficial the cause must have been good. Note the drive to explain why when any significant event happens and the need to attribute blame when any bad event happens (Walster, 1967; Janoff-Bulman et al.,1984).

By estimating the implications of a given outcome in terms of motivational goals an approximate measure of outcome salience can be generated. The next section utilises these

estimates of outcome salience in a comparison of hindsight and outcome biases. This form of analysis derived from the present findings serves to produce a unified account of these two biases and to explain previously unexplained effects from the hindsight literature.

## 6.5 HINDSIGHT AND OUTCOME BIAS IN THE LIGHT OF THE PRESENT RESEARCH

### 6.5.1 A comparison of hindsight and outcome effects

This thesis adopts the viewpoint that there is no fundamental difference between hindsight and outcome biases and, as was first argued in Section 3.3, that outcome bias is a pervasive effect within which hindsight bias is a special case. In this account it is proposed that the presentation of outcome information will influence any subsequent related judgement. This is supported by the fact that outcome bias studies using on-line presentation of outcome information have demonstrated influences on a wide range of different judgements (See Section 6.3.2). In the present studies outcome information has been shown to be the largest influence on judgements of the quality of experimental design. Notably, in Experiment 7 when the same outcome information was presented at a later time subjects' memories of their earlier judgements of experimental quality were biased in hindsight. These hindsight effects were not limited to subjects' memories of earlier judgements; subjects' memories of the sample sizes in the original experimental descriptions were also biased. These results strongly suggest the pervasive influences of outcome information on any related judgement, whether the outcome is presented on-line or in hindsight.

In order to make a clearer comparison between hindsight and outcome biases it is necessary to take a closer look at differences in methodology in these studies. Before a theoretical account of either of these biases can be proposed the differences between within and between subject designs must be considered. In general within subject designs rely on a memory component. In these cases the judgement task is usually described as a memory task. The differences between these memory tasks and the more direct on-line judgement

tasks usually associated with between subject designs must also be considered.

### 6.5.2 Within and between subject designs

In any hindsight study using a within subjects design, the subjects have to make an initial judgement without outcome information. It is then necessary for the experimenter to introduce a time gap in order for the subjects to forget these judgements, they then repeat a similar judgement having been given outcome information. In a between subjects design one group of subjects makes a judgement having been given outcome information whilst another group makes the same judgement without outcome information, in this case there is no need for a time delay.

Because of the use of within subjects methodology hindsight bias is traditionally seen as a bias of memory. In hindsight bias studies it is usually accepted that the memory of an earlier judgement (usually likelihood) is biased by outcome information presented at some later time. In this paradigm two judgements are made. The first judgement is made without outcome information. The second judgement is made some time later when the outcome is known. In this experimental paradigm this second outcome judgement is then compared, within subjects, to the earlier no outcome baseline.

In contrast, the outcome bias paradigm is generally a between subjects design. In this case comparisons are made between subject groups, with and without (or with different) outcome information. If the variations in judgement tasks are ignored, the only difference in this paradigm is that subjects don't act as their own controls. The paradigm used in present research was an exception to this rule as it enabled on-line outcome effects to be compared within subjects for the first time. Thus, the reason that it is difficult to make a clear distinction between hindsight and outcome paradigms is that if a hindsight bias study was run between subjects, rather than within subjects, it would become an outcome bias study. In both cases the comparison is between a judgement made without outcome information and the same judgement made with outcome information. Notably, if this methodological difference is used to define whether a given effect is hindsight bias or outcome bias, then the seminal studies reported by Fischhoff (1975) are in fact outcome bias studies. In the three experiments he reported, effects were all based on differences between subjects who had

been presented with different on-line outcome information (or no outcome information). The first study to actually use a memory based hindsight paradigm was that of Fischhoff and Beyth (1975). It should be noted that in this experiment the time interval between the initial judgements and the "memory" of these judgements was anything from 2 weeks to 6 months.

A direct comparison of between subjects and within subjects methods is available in the hindsight literature. In a study principally designed to test motivational effects in hindsight bias, Campbell and Tesser (1983) used both methods with the same subject groups and with counterbalanced tasks. The results of this study show significant levels of hindsight bias in both Memory and Hypothetical judgements. However, the between subjects 'Hypothetical' condition produced a bias four times larger than the bias in the within subjects 'memory' condition. It seems extremely likely that this difference in effect size is due to subjects' ability to recall their earlier judgements in the 'Memory' condition, particularly when it is considered that the time delay before recall was only 30 minutes. Nevertheless, there is still a significant hindsight bias in this memory condition, even if it is small. Thus it can be seen that in the within subjects paradigm the size of the resulting bias is reduced by subjects' ability to remember their earlier judgements.

For the purposes of argument, if the first judgement is ignored then this second judgement can be seen as no different from any on-line outcome judgement. By treating these two judgements as separate, this description of the hindsight bias paradigm assumes there is little or no influence from subjects' memories of their earlier judgements. If this is the case then the difference between the hindsight effect and the outcome effect is only a difference in methodology. It is, therefore, necessary to assess the role of memory in within subjects hindsight bias studies in some detail and to clarify the effects of the different tasks involved in within and between subjects designs.

### 6.5.3 Differences in judgement tasks and the effects of memory

As described above the within and between subjects designs lead to different judgement tasks. Comparison between studies using these different methodologies is limited by the belief that in each case the subject groups with outcome information are making different judgements. In the between subjects outcome bias paradigm, the subjects

with outcome information are making an original one-off judgement. In the within subjects hindsight bias paradigm subjects are instructed to try to recreate their earlier judgement after receiving outcome information. The clear difference being that in one case the task is to make an original judgement and in the other the task is to remember.

For this reason hindsight bias is often described as a memory effect. There is considerable evidence to suggest that this view is wrong. In fact it is more likely that the opposite is true, the effect will only occur where the original judgement has been forgotten. Where this is the case the second judgement with outcome information is as much a new judgement as that in the between subjects outcome paradigm. When subjects are requested to recall their earlier judgements, if these judgements have been forgotten, the only way they can complete the task is to recreate the judgement. No hindsight studies have given subjects the option to say they have forgotten, for example, Fischhoff's (1977) instructions actually tell subjects to either "remember (or reconstruct if you have forgotten) your original responses" (Page 350).

This proposal, that subjects who exhibit hindsight bias are the ones who are reconstructing rather than remembering, is strongly supported by those studies which have shown that the effect is reduced by the ability to recall the earlier judgement. Hell et al. (1988) showed an effect of memory in a within subjects study of hindsight which used almanac questions. For a group of subjects with no particular motivation to recall their original judgement (as is normally the case in hindsight experiments) hindsight bias was doubled when the time at which outcome information was presented was increased by a week. They also reported that, overall, 35% of their subjects correctly recalled their original responses and thus demonstrated no hindsight bias. Unfortunately the percentages of correct recall related to the different times of presentation of outcome information were not reported. Nevertheless, these findings show that a significant proportion of subjects do not demonstrate any bias, and that increasing the time gap between the initial judgement and the hindsight judgement increases the size of the overall bias. This leads Hell et al. to state "We can, therefore, safely conclude that a prerequisite for a hindsight bias to occur is a weak memory trace of the original response" (Page 537).

That some subjects fail to remember their original responses is not surprising if within

subjects hindsight studies are considered as if they were memory tasks. For example, in the most difficult condition of the Hell et al. (1988) experiment outlined above, subjects would have to remember 88 numeric responses for one week. Fischhoff's (1977) almanac study required a memory of 75 probability judgements for one hour. In contrast, Fischhoff and Beyth's (1977) study, based on real life events, required a memory for fewer items (approximately ten probability judgements) but the time delay was between two weeks and two months. As a final example, in the memory condition of the Campbell and Tesser (1983) experiment reported earlier, subjects would have to remember 80 judgements for up to 45 minutes. Despite the difficulty of this task, correct recall reduced the hindsight bias by a factor of four in this condition. It would be an extremely difficult task to recall one's earlier judgements in any of these experiments, even if one knew in advance that recall would be required and which factors would need to be recalled. Given that subjects are not told to remember anything in the original phases of these experiments, it is clear that these potential memory tasks are beyond the ability of the experimental subjects. This is not an accident, within subjects hindsight experiments are designed to ensure these tasks are difficult in order to avoid ceiling effects. Hindsight biases only operates under uncertainty; where you have clear memory you have no bias.

Even considering the difficulty of accurate recall, the within subjects hindsight paradigm consistently produces smaller effects than the between subjects outcome bias paradigms. All other things being equal, this suggests that in the hindsight paradigm the effect is being attenuated by, either the ability for a few subjects to clearly recall and reproduce some of their earlier judgements, or the ability of a number of subjects to vaguely recall their earlier judgements and thus limit their responses to a smaller range.

### 6.5.4 Conclusions for hindsight and outcome effects

As shown above, accurate or partial recall attenuates the effects of outcome information in within subject paradigms. Therefore, it seems unlikely that the mechanism underlying hindsight bias is one in which trace memories are adjusted by the assimilation of new outcome information. If this were the case then even subjects with strong trace memories would be biased to some extent. Campbell and Tesser (1983) have shown this

not to be the case when they report 35% of subjects demonstrating no bias. As hindsight bias is only demonstrated by those subjects with little or no recall, then it can be concluded that the judgement made in hindsight is a re-judgement rather than one of adjusted recall. In addition, if hindsight bias were based on the reconstruction or adjustment of existing memory traces, then it would not occur in any of the between subjects experiments. The hypothetical hindsight studies (where subjects are asked to respond as if they had not been given outcome information) and the on-line outcome bias studies, do not rely on the attempt to recall or reconstruct an earlier judgement.

None of the remaining information processing accounts (see Section 6.3.1) differentiates between the influences of outcome information presented on-line with the original information, or presented at some later time than the original information. These remaining accounts are based on the way in which outcome information is combined with other relevant information to create a unitary judgement. In these accounts any relevant judgement based on the available information will be biased by the presence of outcome information. This enables the consideration of one comprehensive theoretical account to explain all effects related to the presence of outcome information. This account should relate to both within and between subject designs. Clearly, however, to cover the within subjects hindsight condition any theoretical account needs to allow for the effects of residual memory traces of the original judgement as an additional factor influencing the subsequent judgement.

This view leaves us with an account of hindsight bias in which it is a specific form of outcome bias. Outcome bias implies the fact that the presence of irrelevant outcome information will bias any subsequent judgement. Where this judgement is the recreation of an earlier (largely forgotten) judgement, or where this judgement is an estimate of what might have been judged if the outcome wasn't known, this effect is usually deemed hindsight bias.

## 6.6 A RE-ANALYSIS OF HINDSIGHT EFFECTS IN TERMS OF OUTCOME SALIENCE

The present research has demonstrated both hindsight and outcome effects within the same experimental paradigm (Experiment 7). This finding supported the hypothesis that hindsight bias was a special case of a more pervasive outcome bias effect. In the light of this result the previous section has discussed hindsight bias as a form of outcome effect and has argued that these biases have only previously appeared as separate effects because of methodological differences. Given this unified approach the dependency of outcome bias effects on outcome salience will also apply to hindsight bias effects. Thus by analysing the outcome salience present in previous hindsight studies this section attempts to clarify all the previously unexplained effects present in the hindsight literature.

### 6.6.1 Variations in effect size

There are a number of consistently reported differences in effect size in the literature which may be explained in terms of differences in outcome salience. Overall outcome bias studies demonstrate much larger biases than hindsight studies (Hawkins and Hastie, 1990). This general finding is clearly in line with differences in outcome salience. In hindsight bias studies outcome salience depends on the implications of the outcome in terms of the subjects' personal motivational factors alone. In outcome bias studies these personal motivational factors are also present and, in addition, motivational goals and outcome implications are explicit in the task content. This additional factor will serve to increase outcome salience.

Another commonly reported effect size difference is between different types of hindsight study. Those studies using almanac questions result in unusually large hindsight biases (Hawkins and Hastie, 1990, Christensen-Szalanski and Willham, 1991) As has been noted outcome salience in hindsight studies is entirely dependent on personal motivations. The use of almanac questions will enhance the motive of need for self-esteem. In this case a subject is aware that their general knowledge is being tested and that their responses are a

measure of their personal ability. Under these conditions this enhanced motivation would create an increased outcome salience and thus an increased effect.

#### 6.6.1.1 The effect of negative outcome information

In hindsight bias studies negative outcomes always generate smaller biases, Cristensen - Szalanski and Willham (1991) recognise the existence of smaller effects in those cases where in hindsight subjects were told that an event did not occur. In these cases subjects reduced their earlier likelihood judgements. It should be noted that this is not reverse hindsight bias; the bias is still in the same direction; confirming the outcome. Subjects still 'knew it all along' the difference being that they knew it wouldn't happen (when told that it didn't) rather than knowing it would happen (when told that it did). The present studies demonstrate that this asymmetry between positive and negative outcome information is not limited to hindsight studies. In the on-line outcome bias studies presented here negative outcome information has always produced smaller relative effects than positive information. (Experiments 2,3,4 and 5).

Fischhoff (1977) proposed an explanation of this reduction in hindsight bias (when subjects were told that an effect did not occur) in terms of the cognitive difficulty involved in processing negative information. The difficulty of processing a negative outcome is clarified from the outcome salience viewpoint. With a positive outcome (an event has happened) there is a direct relationship between the outcome information and motivational goals. Here the implications of the outcome support the goals and thus outcome salience is clear. A subject only has to bias their likelihood judgement in line with the outcome to satisfy motivational goals. Where there is a negative outcome (an event did not happen) this results in a negative outcome salience associated with that outcome. This negative outcome salience cannot be used to satisfy motivational goals directly. Here rather than trying to achieve their goals directly, subjects are trying to avoid failing in their goals more than is necessary. In order to do this subjects must infer that two negatives will make a positive; where outcome salience is negative, a bias against that particular outcome will result in a positive influence in terms of motivational goals. This is clearly a more complex process. In the case of outcome bias studies, where the implications of the outcome are explicit, there is the added factor that

subjects have already failed in their main goal. When the outcome is negative their motivation is now a form of damage limitation.

### 6.6.2 Reverse hindsight effects

Reverse hindsight bias involves subjects biasing their memory of likelihood judgements in the opposite direction to the known outcome i.e. 'I never thought that might happen'. This is a completely different case to the effects of negative outcome information outlined above. This is a contentious effect which some observers consider to be caused by methodological irregularities (Mark and Mellor, 1990) and very few clear cases of reverse hindsight exist in the literature. These studies were initially described in Section 3.2.2, but are worth looking at in more detail in terms of the outcome salience associated with the experimental designs in each case.

Mazursky and Ofir (1990) claimed a reversal of hindsight bias in three rather unusual experiments. Rather strangely for a hindsight study none of these experiments had a foresight condition, pre-exposure quality judgements (which the authors misleadingly term recall) were constructed in hindsight from a post-exposure questionnaire in which subjects were asked what their initial quality judgement might have been. Clearly in this case the outcome salience usually associated with hindsight judgements does not apply. There is no motive for self evaluation or for sense of control where an initial (pre-exposure) judgement has not been made. Therefore there is no associated outcome salience. Incidentally, in all these studies post-exposure quality judgements were biased in line with the outcome of the actual exposure. Thus the outcome salience of the actual exposure had the usual effects seen in outcome bias studies.

Another study claiming reverse hindsight bias is that of Verplanken and Pieters (1988). who had completed a study of attitudes to nuclear power plants before the Chernobyl disaster. After the disaster they asked the subjects to recall their earlier judgements of the probability of this type of accident. Remembered probabilities were lower than they had been before the disaster in direct opposition to the usual hindsight effect. This result may also be explained in terms of outcome salience. The potential outcome salience where the outcome is a major disaster will be different from that in normal hindsight studies.

In this case the outcome itself is an extremely negative event and the implications of the outcome are all negative. If the motivations for self-evaluation and sense of control are satisfied the results are unpleasant. Enhanced self-evaluation and sense of control in this case lead to the suggestion that the disaster was predictable and that you knew it was going to happen. A motivation for "It was nothing to do with me" would be more appropriate when the outcome is a disaster. These reversed motivations would lead to a negative outcome salience which in turn would reverse the direction of the bias in exactly the way that has been reported.

### 6.6.3 Conclusions from hindsight studies

The present research suggests that motivational factors cannot be ignored in any account of hindsight or outcome bias. The concept of outcome salience derived from the present research has served to elucidate the previously unexplained effects in hindsight research. This concept implies the existence of an underlying drive based on the implications of outcome information for motivational goals. This does not mean that motivation is necessarily a causal factor, only that in any model of cognitive processing motivational goals must be considered. When discussing a variety of influences of outcome information Fischhoff (1977) noted "Detailed information about how these biases work in their most general form should improve our understanding of how information is stored, altered and retrieved" (Page 350). This remains true even if a motivational account is accepted. The following section considers the concept of outcome salience in terms of theoretical accounts of information processing.

## 6.7 OUTCOME SALIENCE AND POTENTIALLY RELATED CONCEPTS

This thesis has described outcome salience as a motivational factor involved in hindsight and outcome biases. By definition outcome salience is a measure of the utility of a given outcome from the point of view of the subject. As such this approach may have a great deal in common with the concept of subjective expected utility which has been used to

explain choice behaviour.

A second, and quite different, approach explains the utilisation of information in decision making in terms of relevance. A relevance account based on personal goal driven measures of relevance would also have a number of similarities to the concept of outcome salience.

This section attempts to clarify the position of outcome salience in terms of these potentially related concepts.

### 6.7.1 Outcome salience and Subjective expected utility

Outcome salience is in part a measure of the utility of outcome information in terms of personal goals. In this respect it has a great deal in common with the concept of subjective utility. Subjective expected utility theory was initially formulated as a theory of choice behaviour. Normative decision theory, as described by Von Winterfeldt and Edwards (1986) describes an approach to selecting goals and the actions required to attain them. This approach is based on maximising subjective expected utility (SEU) in choices. In this approach choice preferences based on individual goals are rated in terms of their subjective utility and the probability of their possible outcome. SEU is then calculated by summing the various outcomes and their probabilities using the following formula;

$$SEU = \sum_i s_i U_i$$

where $s_i$ represents the subjective probability of the ith outcome and $U_i$ represents it's subjective utility. The calculation of SEU is, therefore, the probability of a given outcome multiplied by its subjective utility, summed over all possible outcomes (the range of i). In order to make the innumerable unconscious everyday choices that allow human beings to cope with their environment this formula would have to be the basis of a very rapid and complex implicit processing system.

If S.E.U. theory does describe choice behaviour, it should be possible to adapt this theory to explain the use of information in judgement. As choice depends on the respective subjective utilities of available actions, then judgement would depend on the respective

subjective utility of available information. In this case the probability of an outcome occurring has no meaning. Therefore, this term could been replaced by believability; a term reflecting the probability that a piece of information is true. Given these changes, it is possible to generate the following model of subjective expected utility of information [SEU(i)].

$$SEU(i) = \sum_z b_z U_z$$

where $b_z$ represents the subjective believability of a piece of information and $U_z$ represents its subjective utility in relation to the zth goal . The calculation of SEU(i) is, therefore, the summation of believability multiplied by goal relevance over all personal goals (the range of z).

In order to differentiate between these two different SEU calculations, the calculation relevant to choice behaviour will be termed SEU(c) and that relating to the utility of information SEU(i). There is a possible difference in these two descriptions if the view is taken that in choice behaviour only one final choice is selected from many possibilities. Thus the choice with the highest SEU(c) is acted upon and all the others are discarded. This is not the case in judgement where any number of pieces of information may be included in the final judgement. In this case it is necessary to consider the calculation of SEU(i) as a means of generating a hierarchy of information which can then be utilised in weighting the influences of individual sources. Given the necessary generation of an SEU hierarchy in judgement, it is also possible that in choice a hierarchy of preferred choices may be generated from this form of processing. This intuitively seems a better view, in a situation where the preferred choice became unavailable there would be an immediate alternative. This situation would allow for more flexible response and remove the need for unnecessary re-processing of SEU(c).

Whilst this seems an acceptable general model, in the case of outcome bias experiments this calculation has an unusual result. If it is assumed that subjects believe the information they are presented with, then the believability factor will always have a value of 1.00. This effectively removes this factor from the formula. In this case the subjective

expected utility of a piece of information is equal to the sum of the subjective utilities of that information for each goal. Thus, in the case of outcome information the SEU(i) is equal to the sum of the implications of the outcome in respect of a subjects' (motivational) goals. This is, of course, exactly the same as the definition of outcome salience.

Given this correspondence between the SEU(i) of outcome information and outcome salience, the present experimental results have implications for this potential model. It has been seen that outcome salience influences subsequent judgements in both positive and negative directions. Therefore, the range of possible SEU(i) values must also include negative values where a piece of information will have a negative influence on a subsequent judgement. Thus, the SEU(i) value of a piece of information can be used not only to determine if that information is relevant to a judgement but also the way in which it is relevant to that judgement.

As believability is not relevant in most experimental paradigms this produces a special case in which considerably less processing is required. This ease of processing may also have implications in the hindsight bias paradigm. From the point of view of hindsight the selection of relevant information would seem obvious as the SEU(i) calculation is easy (you have been told the outcome). In order to see from the point of view of someone without outcome information it would be necessary to ignore this knowledge and reconstruct hypothetical outcomes. It would then be necessary to estimate their believability without being influenced by what we know, and proceed with the calculation of SEU(i) based on our hypothetical estimates. As this calculation utilises the full model (by including hypothetical believability values) the processing involved would be more complex than that required when the outcome is known and accepted. It is easy to see why subjects might avoid this more complex processing task, especially when it would rarely serve any goal that could be described as useful in terms of personal motivations.

This is not an argument for or against SEU theory. This rather simplistic reinterpretation of the basic theory serves only to clarify the concept of outcome salience in terms of utility. In doing so no account is taken of existing research on epistemic utility which is considered beyond the scope of the present discussion. Nevertheless, the picture of SEU created by this limited consideration of the selection and use of information in

judgement raises a number of interesting factors not usually considered in choice related SEU models. In attempting to explain the differential use of the implications of information the concept of negative utility has to be considered. A piece of information with a high negative SEU(i) is as important as one with a high positive SEU(i) to the subsequent judgement. This case is not usually considered in the case of choice behaviour where the choice in a given situation depends only on maximising SEU. Yet it seems reasonable that in choice situations people would not only be aware of the best choice in terms of maximising the probability of achieving their goals, they would also be aware of the need to avoid those choices which will have disastrous results. In fact an awareness of these potentially disastrous choices would have a considerably higher survival value.

In discussing the limitations of SEU theory, Simon (1983) notes that it necessitates a view in which "the decision maker contemplates, in one comprehensive view, everything that lies before him" (Page 13). Simon also reports, "the SEU model finesses completely the origins of the values that enter into the utility function; they are simply there" (Page 14). These concerns are reiterated by Evans and Over (1996) who note the impossibility of including all potential outcomes in a calculation of SEU. However, this problem remains even if SEU theory is discarded. Whatever system of processing is used to determine choice or judgement there must be some method of selecting relevant information from the infinite number of possibilities available before subsequent processing takes place. This question of relevance is discussed in the next section.

### 6.7.2 Outcome salience and relevance

The question of how relevant information is selected from the myriad sources available has become an area of increasing interest in decision making. It is not intended that this section will present an account of relevance in decision making. It will focus only on the implications of the present research findings for any potential account of relevance. In the experimental paradigms used in this research the influences of relevance related choices are severely curtailed. Within these experiments a very limited amount of information was presented. To some extent the presence of this information will suggest to subjects that it is relevant. Sperber and Wilson (1986) argue, in line with Grice's (1975) maxim of relevance,

that in any form of goal-directed communication relevance is implied by content. This presents a conflict in hindsight studies where subjects are presented with outcome information which is relevant to their decision and then instructed to ignore it. Similarly in outcome bias studies subjects are presented with information which they can logically deduce is not relevant to their specific task (although it is relevant to the problem content).

The clearest finding from both hindsight and outcome studies is that subjects are unable to ignore this outcome information. Telling them to ignore the information does not work. Telling them to put themselves in the position of someone who did not have this information does not work. Telling them to remember what it was like before they had the information does not work. Also, as demonstrated in Experiment 6, subjects' ratings of the respective relevance of available information to quality judgements is in direct opposition to their actual use of this information in quality judgements.

Using Evans and Over's (1996) dual process theory it is possible to explain why this is the case. Explicit beliefs about the relevance of information based on normative analysis, or explicit instructions to ignore this information, will only influence explicit Type 2 processes. Type 1 processes will be unaffected by this explicit knowledge. This suggests that those subjects who are biased by outcome information are relying to a considerable extent on Type 1 processing to derive their judgement. It follows that what is relevant in a Type 2 process is not the same as what is relevant in a Type 1 process.

In the dual process theory a Type 2 process is explicit and reflects the use of conscious resources such as normative rules. Information would only be relevant to this process where it was relevant to the appropriate normative rules. The explicit rating of information in Experiment 6 reflects this judgement of relevance to some degree. In this case outcome information would not be judged as relevant. Conversely, Type 1 processes involve the tacit identification of goals and the formation of inferences based on these goals. In this form of processing the relevance of information would depend on its implications in terms of these tacit goals. As outcome salience is a measure of the implications of outcome information in terms of a subject's motivational goals, salient outcome information would necessarily be relevant to Type 1 processes. This reflects the automatic inclusion of outcome information in biased judgements.

It has been suggested that the results of Experiment 6 referred to explicit ratings of relevance, as such these ratings reflected the relevance of information in respect of normative rules; i.e. the relevance that relates to Type 2 processes. However, these ratings were in themselves a form of judgement and as such they will also be biased by the influences of motivational goals on Type 1 processes. This bias was evident in the differential effects of subject perspective. That this bias was due to the influences of motivational goals, was shown by the fact that only ratings of the relevance of the potential financial benefits of an outcome (the outcome salience) were affected to any great extent by perspective change.

Sperber et al. (1995) suggest the connection between the salience of information and its perceived relevance when they define relevance in terms of a trade off between the cognitive effect and the cognitive effort resulting from the processing of this information. The form of cognitive processing required to determine relevance would necessarily have to be the sort of rapid preconscious and tacit process suggested by Evans (1989). If this were a complex form of processing this account would only work in experimental environments where information is sparse. In complex, information rich environments it is necessary to pre-select relevant information to prevent subsequent cognitive processing becoming infinitely complex. The nature of these processes is not clear. One of the most interesting questions raised by this research is; What determines a subject's (implicit) belief in the relevance of a given piece of information ? The effects of outcome salience particularly in terms of financial implications suggest that the answer may be based on motivational factors, self-interest above and beyond all else.

## 6.8 IMPLICATIONS OF THE PRESENT RESULTS FOR THEORETICAL APPROACHES TO BIAS

This section considers the most common theoretical explanations for hindsight and outcome biases in relation to the present results. In order to place these theories in a wider context the section will begin with a short introduction to theories of rationality and bias in human reasoning and a brief outline of the relationship of these theories to outcome effects.

Following this the implications of the present results for theoretical accounts of hindsight bias will be presented. No theoretical accounts specific to outcome bias are presented because, as yet, no theories separate from those used to explain hindsight bias have been proposed in the literature.

### 6.8.1 Reasoning, rationality and bias

Before considering theoretical approaches specific to hindsight bias, it is necessary to consider some of the more fundamental approaches to reasoning. It is also useful at this point to clarify the implications of biased judgement in terms of rationality. In order to do this two definitions from Evans and Over (1996) will be adopted. The first of these is a definition of bias as, "a departure from an apparently appropriate normative system" (Page 6). As Evans and Over point out, this definition avoids the pejorative use of the term bias to imply error or irrationality. The second definition concerns what the authors refer to as personal rationality "rationality$_1$" and impersonal rationality "rationality$_2$" these are defined as follows:

" **Rationality$_1$** - Thinking, speaking, reasoning, making a decision, or acting in a way that is generally reliable and efficient for achieving one's goals.

**Rationality$_2$** - Thinking, speaking, reasoning, making a decision, or acting when one has a reason for what one does sanctioned by a normative theory"

Evans and Over (1996), Page 11.

These two definitions of rationality are particularly appropriate when applied to examples of outcome bias studies. From the impersonal point of view (that of the experimenter), using rationality$_2$, the outcome bias reported by Mitchell and Kalb (1981) can be interpreted as supervisors making irrationally biased evaluations of the performance of nurses. These evaluations were based not on the nurse's actions but on the chance outcomes of these actions. From the point of view of rationality$_1$, the subjects' point of view, these results look quite different. In this experiment the subjects were asked to take on the role of a supervisor. The job of a supervisor has a specific goal; to protect patients from harm arising from negligence. Given this overriding goal, a supervisor must punish a nurse whose actions have led to patient harm, even if the supervisor feels that this nurse has been

unlucky compared to a nurse who has made the same mistake without a harmful outcome. Punishment in this case is necessary to warn others to be vigilant and to consider the possible consequences of their actions, also a supervisor with the goal of keeping his or her job would not be successful for long if they took no action when a nurse's negligence led to patient injury. In this case outcome bias is irrational from the point of view of rationality$_2$ but clearly serves the goal directed purposes of rationality$_1$. It should be noted that this form of goal directed reasoning typically appears inappropriate when applied to single cases. Over a large number of cases this form of reasoning inevitably supports long term goals.

This distinction is not so clear in the present research. Experiments 2,3,4 and 5 show subjects judging the quality of sampling procedures on the amount of money they make rather than using the appropriate normative theory. This is clearly irrational from the point of view of rationality$_2$. Especially when it is considered that these subjects have all been taught the appropriate normative theory and some of them are even studying for a degree in this normative theory (B.Sc. Statistics students). Again from the point of view of rationality$_1$ these results look different. In all these experimental scenarios the goal from the subjects' perspective was ultimately to make money. Therefore, on the surface it seems reasonable to favour those sampling procedures which fulfilled this goal. However, in this case, if this strategy was repeated in real life even over a large number of cases it would fail. Section 6.3 describes a number of the potentially damaging practical implications of the use of this strategy.

In the present experimental scenarios subjects had enough information to determine when a good outcome was the result of chance rather than the result of a good sampling procedure. In these cases the outcome was clearly irrelevant and this should have served to reduce the resulting bias. The fact that this explicit knowledge of the irrelevance of the outcome had no debiasing effect is explained by Evans and Over's dual process theory of reasoning.

Evans and Over (1996) contend that human cognition depends on two systems and "the...implicit system is primarily responsible for rationality$_1$ while...the explicit one mainly affects the extent of peoples rationality$_2$" (Page 13). This leads to their view that in good practical reasoning the task of identifying a goal and adopting a reliable way of attaining it is

accomplished "almost immediately with little awareness on our part of how it is done" (Page 17). Evans and Over present a dual process theory of reasoning based on this dual definition of rationality. In this theory there are two distinct forms of cognitive processing; goal directed implicit inferential processes which relate to rationality$_1$ (Type 1 processes) and explicit reasoning processes which relate to rationality$_2$ (Type 2 processes). They suggest that these two processes interact in reasoning and decision making tasks. In this theory Type 1 processes involve the tacit identification of goals and the formation of inferences based on these goals. This process may be extremely computationally complex and is very rapid. It is also entirely unconscious and automatic. It is suggested that these "intuitive" processes reflect the sort of learning required to achieve everyday goals. Type 2 processes are explicit and reflect the use of conscious resources such as normative rules.

Applying this theory to the earlier contention, that subjects would be less biased by an outcome they know to be irrelevant, we see that this explicit knowledge of irrelevance would only influence Type 2 processes. In Type 1 processes where identifying goals and favouring information which potentially supports these goals is implicit and automatic, then explicit knowledge of irrelevance will have little or no debiasing effect. Thus, from this point of view outcome biases can be seen as the result of Type 1 processes in which the outcome information has been automatically cued as relevant to the judgement. If outcome salience is seen as a method of defining Type 1 goals, then the present findings fit perfectly within this description.

The next section will review some of the main information processing and motivational accounts of hindsight bias. In many ways these different approaches to hindsight bias have a great deal in common with the dual processes outlined above. The personality based motivational accounts of hindsight bias are related to goal directed Type 1 processes. These attempts to define a subject's motivation in a given context are no more than an attempt to define the goals a subject is trying to achieve in Type 1 processing. As in rationality$_1$, motivational accounts take the point of view of the subjects' personal rationality in a given situation. Biases arising from this situation are explained in terms of how they meet the needs of the subject. It may also be useful to consider information processing accounts in the light of Type 2 processes. These accounts focus on possible mechanisms by which a

process of reasoning based on normative theory could be influenced by the automatic assimilation of outcome information. In this view the resulting bias is due to errors in the reasoning process.

### 6.8.2 Information-processing and motivational accounts of hindsight bias

There are two main classes of account which have been proposed to explain the processes underlying hindsight bias. These are information processing accounts and motivational accounts. Information processing accounts focus on the way that information is assimilated, stored and retrieved. In these accounts hindsight bias is explained in terms of the influences of subsequent assimilation of hindsight information on one or more of these processes. Motivational accounts focus on those factors present in the task which would lead subjects to derive some personal benefit from biasing their judgements. (See Section 6.4.1)

It has been the tendency for researchers to focus on either one or the other of these accounts. This has to some extent created the misapprehension that these accounts are mutually incompatible; either subjects are motivated to produce bias or it is the result of implicit cognitive processes. This belief may in some part be due to the false impression that motivation must be explicit and that subjects consciously form biased judgements. In fact when motivations are seen as implicit these two accounts are entirely compatible and will fit within the sort of dual process theory described above. Nevertheless, the influences of motivational factors defined by the present research clearly demonstrate that an account based on information-processing alone will not be adequate. The following section re-analyses information processing accounts in the light of the present results.

#### 6.8.2.1 Information-processing accounts of hindsight bias

An information-processing account was first suggested by Fischhoff (1975) who suggested hindsight bias was due to a form of 'creeping determinism'. Fischhoff suggests "that on the receipt of outcome knowledge judges immediately assimilate it with what they already know about the event in question" (Page 279). This immediate assimilation serves to make a coherent whole from all the information available. Creeping determinism is the

direct result of using this coherent (biased) description in any subsequent hindsight judgement. In this account judges remain unaware of both the assimilation process and its subsequent influences. Fischhoff (1977) suggests that this assimilation process "may involve both reinterpreting previously held information to make sense out of it in light of the reported answer and strengthening associative links with reasons supporting the reported answer". (Page 356) Thus this description relies on a theoretical process in which some form of coherent memory of an event is automatically revised (updated) upon the receipt of subsequent relevant information. This account would need considerable revision in order to explain the effects of on-line outcome information and the effects of outcome salience demonstrated in the present research.

Fischhoff (1975) also presented an alternative explanation based on Tversky and Kahneman's (1974) anchoring and adjustment heuristic. In this explanation it was proposed that on receipt of outcome information a subject assigns this outcome a probability of 1.00. They then look for reasons to adjust this value downwards. In general this type of adjustment has been shown to be inadequate (Slovic and Lichenstein, 1971, Tversky and Kahneman, 1974). In the case of judgements based on outcome information, this inadequate adjustment would lead to the overestimation of hindsight probabilities. This explanation has since been largely disregarded due to its inability to explain differences in hindsight effects (Fischhoff and Beyth, 1975 and Fischhoff, 1977). In addition to these failures this account cannot explain the differential influences of motivational factors shown in the present research. Thus, it seems reasonable to discount anchoring and adjustment as a possible explanation.

There are a number of other possible mechanisms on which an information processing account may be based. These mechanisms fall into two basic categories: memory updating and re-judgement. Fischhoff (1977) focused mainly on a processing account in which a global memory was automatically updated by the automatic assimilation of outcome information. This global memory consisted of the memory of an earlier likelihood judgement and memories of the factors which were related to that judgement. In this account, the adjustment of this global memory included both biasing the memory of likelihood in the direction of the known outcome (creating hindsight bias) and increasing the

strength of associative links with evidence supporting the known outcome (creating post hoc rationalisation). In order to account for biases produced in between subjects experiments where outcome information is presented on-line, the same effect must occur in the creation of global memory traces as well as in the updating of existing memory traces. Again there is no clear mechanism by which this account could be adapted to allow for the effects of variations in outcome salience seen in the present research.

The other category of information processing accounts involves re-judgement. In these accounts it is assumed that on the receipt of outcome information subjects form a new likelihood judgement. This new judgement then replaces, or is to some extent integrated with, their earlier judgement. Hawkins and Hastie (1990) describe a number of alternative strategies on which re-judgement could be based. The first of these theoretical strategies involves the availability of evidence on which a re-judgement is made. When the new judgement is made relevant evidence from the environment and from long-term memory must be reviewed. In this case hindsight bias is the result of a reduction in the accessibility of information which does not fit the known outcome. Slovic and Fischhoff (1977) imply this strategy when they suggest that in hindsight only the stated outcome is used as a cue to information retrieval.

A second theoretical strategy is based on the evaluation of available information. In this strategy it is suggested that once evidence has been selected as relevant further processing takes place to combine this evidence with other available evidence. In this further processing the implications of raw evidence are assessed in the light of what is known about the whole topic. This process is suggested by Fischhoff's (1975) concept of the creation of a coherent whole in which a judge attempts to make sense of all the information presented. Given the context of a particular outcome, information which can be interpreted in terms of causal relations would be easily combined into this coherent whole. This inferential process would serve to create a consistent world view in which relevant information clearly caused the given outcome. In this system any available information which did not fit within the context of the outcome would not be combined within the coherent whole and would thus be disregarded. As Hawkins and Hastie (1990) point out, it is unlikely that this process is as easy, automatic or unconscious as Fischhoff's original account suggests, relying as it does

on the formation of causal inferences. Nevertheless, they accept that the necessary inference processes "could easily occur in the time frame" (Page 322). This form of account is considerably more open to the possible inclusion of motivational factors. If the assessment of raw evidence was based on its implications in terms of motivational goals, then it is possible to account for the effects of outcome salience. However, if this adaptation is made the role of causal inferences (on which this account was originally based) become considerably weaker.

A third theoretical strategy by which information processing could account for hindsight biases is based on the combination of the implications of available information to produce a unitary judgement. This account relies on more complex and, thus, less automatic cognitive processes. In this case a judge attempts to derive an appropriate response by combining the implications of the available evidence. Thus, each piece of evidence must first be weighted on some form of quantitative scale and then using these weights the implications of the evidence must be summed in order to form a unitary judgement. Hawkins and Hastie (1990) suggest the factors used for weighting evidence would be; credibility, authority, relevance and importance. They emphasise that in this process hindsight bias is a side effect of an adaptive learning process. It is easy to see that in this strategy outcome information would inevitably receive a high weighting, given that its credibility is usually not in question and that its relevance would always be high. Indeed, if you believe that the end justifies the means, the relevance of the outcome overshadows that of all other factors. In this potential model the meanings of the factors "authority" and "importance" are not so clear. It seems probable that the "authority" of a given piece of information would only serve to influence its credibility, thus, this factor could be subsumed under the credibility factor. If the term "importance" refers to the importance of the information to the final judgement then it is no more than the sum of all the other factors (or possibly another

measure of relevance). If, on the other hand, it refers to the importance of the information from the point of view of the judge then the term "personal salience" would be clearer. Thus, this information processing account is entirely compatible with the present findings, and motivational accounts in general, if it is assumed that the factors used for weighting

information will vary with the subjects' motivation.

## 6.9 SUMMARY AND CONCLUSIONS

In this thesis the effects of the presence of outcome information were shown to be an important source of bias to judgements under uncertainty. These biases were seen to affect on-line judgements, memories of judgements and even memories of facts relating to judgements. In the majority of the experiments reported here these judgements were of quality, however, it was concluded that there was no reason to suppose these effects are limited to a specific type of judgement or memory. The present experiments also demonstrated that the size and direction of outcome biases were mediated by a factor described as outcome salience. This factor was defined as "the relative importance of the implications of an outcome from the point of view of the subject".

It was shown that the present results had practical implications in terms of experimental design where specific failures of designers to learn from experience were predicted. Specific predictions were also made relating to areas in which potentially damaging biased quality judgements would be expected.

Following a discussion of outcome salience it was possible to present an account which unified hindsight and outcome biases. A comparison was made between hindsight and outcome bias studies based on the present findings and an analysis of the methodological differences in these studies. This comparison concluded that the effect known as hindsight bias is simply a special case of a more wide ranging and pervasive outcome bias effect. In support of this contention previously unexplained variations in hindsight bias were analysed in terms of the potential outcome salience present in these paradigms. This analysis supported the view that hindsight and outcome effects arise from a common source, and that variations in the size and direction of all biases related to outcome information can be explained by outcome salience. The relationship between this concept of outcome salience and subjective expected utility was explored. By adapting SEU theory to account for preferential use of information in decision making, the similarity between the

resulting SEU values and outcome salience was highlighted. The implications of this rather simplistic model in terms of the need to consider negative SEU values were discussed. In addition a brief account of the role of relevance was presented in which the similarities between outcome salience and relevance were noted.

In a more general theoretical discussion definitions of bias and rationality and the dual process model of thinking, proposed by Evans and Over (1996), were reviewed. This theory was shown to provide a good explanatory framework for both the present results and outcome bias effects in general. Information processing accounts of hindsight bias were also considered in the light of the present results. The concept of outcome salience was shown to derive from the implications of outcome information in terms of a subject's motivational goals. It was demonstrated that motivational factors such as this must be considered in any complete account of biases related to outcome information.

The present results were in general support of the dual process theory of thinking proposed by Evans and Over (1996). Within the general framework of this theory, the importance of the implications of motivational goals (for example outcome salience) in Type 1 processing was emphasised.

## REFERENCES


Abdolmohammadi, M.J. and Shanteau, J. (1992) Personal attributes of expert auditors. Organizational Behaviour and Human Decision Processes 53 158-172

Adelman, L. and Bresnick, T. (1992) Examining the effect of information sequence on patriot air defense officers' judgements. Organizational Behaviour and Human Decision Processes 53 204-228

Baron, J. and Hershey, J.C. (1988) Outcome bias in decision evaluation. Journal of Personality and Social Psychology 54(4) 569-579.

Berry, D.C. and Broadbent, D.E. (1988) Interactive tasks and the implicit-explicit distinction. British Journal of Psychology 79 251-272

Berry, D.C. and Broadbent, D.E. (1984) On the relationship between task performance and associated verbalizable knowledge. The Quarterly Journal of Experimental Psychology 36A 209-231

Bhaskar, R. and Simon, H.A. (1977) Problem solving in semantically rich domains: An example from engineering thermodynamics. Cognitive Science 1 193-215

Brim, O.G.Jr. and Hoff, D.B. (1957) Individual and situational differences in desire for certainty. Journal of Abnormal and Social Psychology 54 225-229

Byrne, R. (1977) Planning Meals: Problem-solving on a real data base. Cognition 5 287-322

Carette, T.R. and Moreland, R.L. (1983) The direct and indirect effects of inadmissable evidence. Journal of Applied Social Psychology 13 291-309

Campbell, J.D. and Tesser, A. (1983) Motivational interpretations of Hindsight bias:An individual difference analysis. Journal of Personality 51(4) 605-620

Carette, T.R. and Moreland, R.L. (1983) The direct and indirect effects of inadmissable evidence. Journal of Applied Social Psychology 13 291-309

Chan, C. (1990) Cognitive processes in architectural design problem solving. Design Studies 11(2) 60-80

Christensen-Szalanski, J.J.J. and Willham, C.F. (1991) The hindsight bias: A meta-analysis.
Organizational Behaviour and Human Decision Processes 48 147-168.

Cohen, A., Stotland, E. and Wolfe, D.M. (1955) An experimental investigation of need for cognition. Journal of Abnormal and Social Psychology 51 291-297

Cohen, J. (1962) The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology 65 145-153

Cohen, J. (1977) Statistical power analysis for the behavioural sciences.
Earlbaum, London.

Cohen, J. (1988) Statistical power analysis for the behavioural sciences - 2nd Edition.
Earlbaum, London.

Cohen, J. (1990) Things I have learned (so far).
American Psychologist 45(12) 1304-1312

Doob, A.N. and Kirshenbaum, H.M. (1972) Some empirical evidence on the effect of section 12 of the Canada Evidence Act upon the accused.
Criminal Law Quarterly 15 88-96

Eckersley, M. (1988) The form of design processes: a protocol analysis study.
Design Studies 9(2) 86-94

Eekels, J. and Roozenburg, N.F.M. (1991) A methodological comparison of the structures of scientific research and engineering design: their similarities and differences. Design Studies 12(4) 197-203

Ericsson, K.A. and Simon, H.A. (1980) Verbal reports as data.
Psychological Review 87(3) 215-251

Ericsson, K.A. and Simon, H.A. (1984) Protocol analysis: Verbal reports as data.
MA, MIT Press

Evans, J. St. B.T. and Bradshaw, H. (1986) Estimating sample size requirements in research design: A study of intuitive statistical judgement.
Current Psychological Research and Reviews 5 10-19

Evans, J.St.B.T. and Dusoir, A.E. (1977) Proportionality and sample size as factors in intuitive statistical judgement. Acta Psychologica 41 129-137.

Evans, J.St.B.T. (1983) Thinking and reasoning, psychological approaches. London, Earlbaum.

Evans, J.St.B.T. (1981) A reply to morris. British Journal of Psychology 72 469-470

Evans, J.St.B.T. (1989) Bias in human reasoning: Causes and consequences. Hove (U.K.), Earlbaum

Evans, J. St. B.T. and Wason, P.C. (1976) Rationalisation in a reasoning task. British Journal of Psychology 67 479-486

Evans, J.St.B.T. and Over, D.E. (1996) Rationality and Reasoning. London, Earlbaum.

Fischhoff, B. and Beyth, R. (1975) "I knew it would happen"- Remembered probabilities for once future things. Organizational Behaviour and Human Performance 13 1-16

Fischhoff, B. (1975) Hindsight =/= Foresight: The effect of outcome knowledge on judgement under uncertainty. Journal of Experimental Psychology: Human Perception and Performance 1 288-299.

Fischhoff, B. (1977) Percieved informativeness of facts. Journal of Experimental Psychology: Human Perception and Performance 3 349-358

Gick, M.L. and Holyoak , K.J. (1980) Analogical problem solving. Cognitive Psychology 12 306-355

Gill, M.A. (1987) Fuzziness and loss of information in statistical problems. IEEE Transactions on Systems Man and Cybernetics 17(6) 1016-1025

Goffman, E. (1967) Interaction ritual . New York, Anchor

Grice, P. (1975) Logic and conversation. In P. Cole and J. Morgan (Eds) Studies in syntax Vol 3: Speech acts. New York, Academic Press

Hammond, K.R., Hamm, R.M., Grassia, J. and Pearson, T. (1987) Direct comparison of the efficacy of intuitive and analytical cognition in expert judgement. IEEE Transactions on Systems Man and Cybernetics 17(5) 753-770

Hans, V.P. and Doob, A.N. (1976) Section 12 of the Canada Evidence Act and the deliberations of simulated juries. Criminal Law Quarterly 18 235-253

Hasher, L., Attig, M.S. and Alba, J.W. (1981) I knew it all along: Or did I ? Journal of Verbal Learning and Verbal Behaviour 20 86-96

Hastie, R. and Park, B. (1986) The relationship between memory and judgement depends on whether the judgement task is on-line or memory based. Psychological Review 93 258-268

Hawkins, S.A. and Hastie, R. (1990) Hindsight: Biased judgements of past events after the outcomes are known. Psychological Bulletin 107(3) 311-327.

Hell, W., Gigerenzer, G., Gauggel, S., Mall, M. and Mueller, M. (1988) Hindsight bias: An interaction of automatic and motivational factors ? Memory and Cognition 16 533-538

Henslin, J.M. (1967) Craps and magic. American Journal of Sociology 73 316-330

Hoch, S.J. and Lowenstein, G.F. (1989) Outcome feedback: Hindsight *and* information. Organizational Behaviour and Human Decision Processes 15 605-619

Hoiberg, B.C. and Stires, L.K. (1973) The effects of several types of pretrial publicity on the guilt attributions of simulated jurors. Journal of Applied Social Psychology 3 267-275

Horowitz, A., Bordens, K.S. and Feldman, M.S. (1980) A comparison of verdicts obtained in severed and joined criminal trials. Journal of Applied Social Psychology 10 444-456

Hubka, V. and Eder, W.E. (1987) A scientific approach to engineering design. Design Studies 8(3) 123-132

Janoff-Bulman, R., Timko, C. and Carli, L.L. (1985) Cognitive biases in blaming the victim. Journal of Experimental Social Psychology 21 161-177

Kahneman, D. and Tversky, A. (1972) Subjective probability: a judgement of representativeness. Cognitive Psychology 3 430-454.

Kahney, H. (1993) Problem solving - Current issues (2nd. ed.) Open University Press, Buckingham Philadelphia.

Kelley, G.A. (1955) The psychology of personal constructs. New York, Norton

Kelley, H.H. (1971) Attribution in social interaction. General Learning Press, Morristown, N.J.

Klahr, D., Fay, A.L. and Dunbar, K. (1993) Heuristics for scientific experimentation: A developmental study. Cognitive Psychology 25 111-146

Kotovsky, K., Hayes, J.R. and Simon, H.A. (1985) Why are some problems hard ? Evidence from the Tower of Hanoi. Cognitive Psychology 17 248-294

Langer, E. J. (1975) The illusion of control. Journal of Personality and Applied Psychology 32 311-328

Leary, M.R. (1982) Hindsight distortion and the 1980 presidential election. Personality and Social Psychology Bulletin 8 257-263

Lerner, M.J. and Miller, D.T. (1978) Just world research and the attribution process: Looking back and ahead. Psychological Bulletin 85 1030-1051

Lichenstien, S., Fischhoff, B. and Phillips, L.D. (1982) Calibration of probabilities: The state of the art to 1980. London, Earlbaum.

Lipsey, M.W. (1990) Design Sensitivity; Statistical power forexperimental research. London, Sage

Lipshitz, R. (1989) "Either a medal or a Corporal": The effects of success and failure on the evaluation of decision making and decision makers. Organizational Behaviour and Human Decision Processes 44 380-395

Luger, G.F. and Bauer, M.A. (1978) Transfer effects in isomorphic problem situations. Acta Psychologica 42 121-131

Mark, M.M. and Mellor, S. (1990) "We don't expect it happened": On Mazursky and Ofir's (1990) purported reversal of the hindsight bias. Organizational Behaviour and Human Decision Processes 57 247-252

Mazursky, D. and Ofir, C. (1990) "I could never have expected it to happen": The reversal of the hindsight bias. Organizational Behaviour and Human Decision Processes 46 20-33

Mitchell, T.R. and Kalb, L.S. (1981) Effects of outcome knowledge and outcome valence on supervisors' evaluations. Journal of Applied Psychology 66 604-612.

Nisbett, R.E and Wilson, T.D (1977) Telling more than we can know: Verbal reports on mental processes. Psychological Review 84(3) 231-259

Pennington, D.C., Rutter, D.R., McKenna, K. and Morely, I.E. (1980) Estimating the outcome of a pregnancy test: Women's judgements in hindsight and foresight. British Journal of Social and Clinical Psychology 19 317-324

Pennington, D.C. (1981) The British firemans strike of 1977/78: An investigation of judgements in hindsight and foresight. British Journal of Social Psychology 20 89-96

Pervin, L.A. (1963) The need to predict and control under conditions of threat. Journal of Personality 31 570-587

Powell, J.L. (1988) A test of the knew it all along effect in the 1984 presidential statewide elections. Journal of Applied Social Psychology 18 760-773

Reagan, R.T. (1989) Variations on seminal demonstration of people's insensitivity to sample size.
Organizational Behaviour and Human Decision Processes 43 52-57

Reed, S.K., Ernst, G.W. and Banerji, R. (1974) The role of analogy in transfer between similar problem states. Cognitive Psychology 6 436-450

Roozenburg, N.F.M. and Cross, N.G. (1991) Models of the design process: Integrating across the disciplines. Design Studies 12(4) 215-220

Ross, M and Conway, M. (1986) Remembering one's own past. In, R.M. Sorrentino and E.T. Higgins (Eds) The handbook of motivation and cognition.
Guildford press, New York

Sanbonmatsu, D.M., Kardes, F.R. and Herr, P.M. (1992) The role of prior knowledge and missing information in multiattribute evaluation.
Organizational Behaviour and Human Decision Processes 51 76-91

Schul, Y. and Burnstein, E. (1985) When discounting fails: Conditions under which individuals use discredited information in making a judgement.
Journal of Personality and Social Psychology 49 894-903

Sedlmeier, P. and Gigerenzer, G. (1989) Do studies of statistical power have an effect on the power of studies ? Psychological Bulletin 105(2) 309-316

Shanteau, J. (1992) Competence in experts: The role of task characteristics.
Organizational Behaviour and Human Decision Processes 53 252-266

Shanteau, J. and Stewart, T.R. (1992) Why study expert decision making ? Some historical perspectives and comments.
Organizational Behaviour and Human Decision Processes 53 95-106

Simon, H.A. (1978) Information processing theories of human problem solving. In Estes, W.K. (Ed.) Handbook of learning and cognition processes.
Earlbaum, London

Simon, H.A. (1983) Reason in Human affairs.
Stanford, Stanford university press

Slovic, P. and Fischhoff, B. (1977) On the psychology of experimental surprises. Journal of Experimental Psychology: Human Perception and Performance 3(4) 544-551

Slovic, P. and Lichenstien, S. (1971) Comparison of Bayesian and regression approaches to the study of information processing in judgement. Organizational Behaviour and Human Performance 6 649-744

Smith, G.F. (1989) Representational effects on the solving of an unstructured decision problem. IEEE Transactions on Systems Man and Cybernetics 19(5) 1083-1090

Sperber, D. and Wilson, D. (1986) Relevance. Oxford, Blackwell

Sperber, D., Cara, F. and Girotto, V. (1995) Relevance theory explains the selection task. Cognition 57 31-95

Stauffer, L.A. and Ullman, D.G. (1988) A comparison of the results of studies into the mechanical design process. Design Studies 9(2) 107-114

Stone, D.N. and Schkade, D.A. (1991) Numeric and linguistic information representation in multi-attribute choice. Organizational Behaviour and Human Decision Processes 49 42-59

Sue, S., Smith, R.E. and Caldwell, C. (1973) Effects of inadmissible evidence on the decisions of simulated jurors: A moral dilemma. Journal of Applied Social Psychology 3 345-353

Tanford, S. and Penrod, S.D. (1982) Biases in trials involving defendants charged with multiple offences. Journal of Applied Social Psychology 12 453-470

Tempczyk, H. (1986) A survey of research and studies on design. Design Studies 7(4) 199-215

Thompson, W.C., Fong, G.T. and Rosenhan, D.L. (1981) Inadmissible evidence and juror verdicts. Journal of Personality and Social Psychology 40 453-463

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

Ullman, D.G. (1991) The status of design theory research in the United States.
Design Studies 12(4) 204-207

Verplanken, B. and Pieters, R.G.M. (1988) Individual differences in reverse hindsight bias: I never thought something like Chernobyl would happen. Did I ?
Journal of Behavioral Decision Making 1 131-147

Vincente, K.J. and Brewer, W.F. (1993) Reconstructive remembering of the scientific literature. Cognition 46 101-128

Von Winterfeldt, D. and Edwards, W. (1986) Decision analysis andbehavioural research. Cambridge, Cambridge University Press

Walster, E. (1967) 'Second guessing' important events.
Human Relations 20 239-249

Ward, A. (1989) Phenomenological analysis in the design process.
Design Studies 10(1) 53-61

Wason, P.C. and Evans, J. St. B.T. (1975) Dual processes in reasoning.
Cognition 3 141-154

Well, A.D., Pollatsek, A. and Boyce, S.J. (1990) Understanding the effects of sample size on the variability of the mean.
Organizational Behaviour and Human Decision Processes 47 289-312

Werner, C.M., Kagehero, D.K. and Strube, M.J. (1982) Conviction proneness and the authoritarian juror: Inability to disregard information or attitudinal bias.
Journal of Applied Psychology 12 629-636

White, R. W. (1959) Motivation reconsidered: The concept of competence.
Psychological Review 66 297-333

Wolf, S. and Montgomery, D.A. (1977) Effects of inadmissible evidence and level of judicial admonishment to disregard on the judgements of mock jurors.
Journal of Applied Social Psychology 7 205-219

Wood, G. (1978) The knew it all along effect.
Journal of Experimental Psychology: Human Perception and Performance 4 345-353

Wortman, C.B. (1976) Causal attributions and personal control. In, J.H.Harvey, W.J.Ickes and R.F.Kidd (Eds) New Directions in Attribution Research. Vol.1. Hillsdale,N.J., Erlbaum

Wyer, R.S.Jr. and Budesheim, T.L. (1987) Person memory and judgements: The impact of information that one is told to disregard. Journal of Personality and Social Psychology 53 14-29

Wyer, R.S.Jr. and Unversagt, W.H. (1985) The effects of instructions to disregard information on its subsequent recall and use in making judgements. Journal of Personality and Social Psychology 48 533-549

## Appendices - Table of Contents

# APPENDIX 1

## EXPERIMENT ONE MATERIALS

### Task Instructions

You will be presented with four problems, in the form of experimental hypotheses. Your task, in each case, is to design a study to test the hypothesis.

Your design should attempt to find the best possible method of testing the given hypothesis.

During the task it is important that you talk aloud describing what you are thinking about at that moment. If you fall silent you will be prompted to continue talking.

You may jot down any notes you need on the question sheet, you will have fifteen minutes to consider each question. After the first ten minutes you will be given an answer sheet requesting specific information about the study.

### Questions

**A** Design a study to test the hypothesis:

**Differences in table height affect writing speed.**

**B** Design a study to test the hypothesis:

**Physical pain affects reaction time to non painful stimuli.**

**C** Design a study to test the hypothesis:

**Background noise will increase solution times for quadratic equations for more mathematically able subjects, less able subjects will not be affected.**

**D** Design a study to test the hypothesis:

**Faces of close friends in photographs are more readily recognisable than faces of people only seen very rarely.**

**Answer sheet**

Please read each of the following headings aloud and if they are relevant, briefly jot down any conclusions you have come to in the design of your study, they may be done in any order, please continue to talk aloud.

**Independent Variables**

**Dependent Variables, types of measurement**

**Subjects (and subject groups)**

**Within, between subjects or mixed design** .....

**Numbers of subjects** .....

**Controlled Variables**

**Advantages of this design**

**Problems with this design**

## EXPERIMENT ONE CODING SCHEME AND CODED RESULTS

### Coding scheme

Utterances were coded into the following categories:

Q - Any clarification of the original question

I - Consideration of the independent variable

D - Consideration of the dependent variable

C - Consideration of confounds

WB - Consideration of within or between subjects design

SN - Consideration of subject numbers

SL - Consideration of significance levels

SG - Consideration of subject groups or subject type

Po - Consideration of power

P - Consideration of the practicality of a design

CO - Consideration of cost

M - Consideration of materials

ST - Consideration of statistical analysis

ET - Consideration of ethics

EV - Any form of evaluation (Either general evaluation or added to any of the previous categories to show which factor is being evaluated)

* - Any form of decision (Added to one of the previous categories to show which factor has been decided)

## Experiment one - categorical coding of verbal protocols

**First question attempted**

| Subject | 3 | 7 | 5 | 6 | 10 | 4 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| Question | A | A | B | B | B | C | C | D |
| | D | I | Q | I | Q | Q | Q | Q |
| | C | SG | I | Q | D | I | SG | P |
| | C | WB | Q | I | D | D | I | D |
| | I | I | I | I | I | SG | WB* | I |
| | Q | P | Q | P | P | Q | EV | P |
| | I* | C | Q | I | P | I | P | Q |
| | C | C | Q | WB* | SG | SG | I | I |
| | SN* | D | I | SN* | C | C | I | P |
| | C | P | I | D | SN | P | EV | C |
| | WB | I | Q | I | SN* | C | SG | Q |
| | CO | C | I | I | SNEV | E | SGEV | C |
| | WB* | WB | Q | I | I | E | IEV | C |
| | Q | C | SG | I | I | C | EV | Q |
| | P | Q | I | WBE | C | C | C | I |
| | C | ST | I | Q | WB | Q | WB* | STEV |
| | Q | ST | SG | WB | WBEV | C | C | P |
| | C | | I | C | C | SG | SG | DEV |
| | P | | | SN | C | WB* | | D |
| | M | | | C | WB* | I | | Q |
| | C | | | | ST | D | | I |
| | P | | | | EV | | | D |
| | P | | | | ST | | | I |
| | CO | | | | EV | | | EV |
| | | | | | EV | | | I* |
| | | | | | EV | | | P |
| | | | | | C | | | D* |
| | | | | | C | | | P |
| | | | | | SGEV | | | SG |
| | | | | | STEV | | | P |

## Second question attempted

| Subject | 4 | 5 | 8 | 3 | 7 | 9 | 6 | 10 |
|---|---|---|---|---|---|---|---|---|
| Question | A | A | A | B | B | C | D | D |
| | Q | SG | Q | Q | I | I | I | Q |
| | C | C | I | I | Q | I | D | D |
| | WB* | SG | WB | E | ET | WB | C | D |
| | P | I | EV | SG | Q | I | D | D |
| | I | D | EV | D | SG | P | WB* | D |
| | I* | P | I | I | I | D | I | C |
| | I | I | I | Q | WB | SG | D | P |
| | SG | P | C | I | WB* | SG | C | I |
| | D | P | EV | I | I | SG | C | D |
| | D | C | WB | E | I | Q | C | D |
| | C | | WB | I | E | SG | Q | SN |
| | WBE | | CO | D* | D | C | C | WB |
| | I | | SG | I | Q | C | WB | WB |
| | C | | | WB | SN | C | D | C |
| | WBE | | | I | ET | EV | SG | C |
| | COE | | | WB | ET | ST | C | WB* |
| | C | | | C | SN* | ST | C | STEV |
| | C | | | Q | I | STEV | C | SNEV |
| | ST | | | C | ET | CEV | C | ST |
| | C | | | I* | ET | ST | C | CO |
| | C | | | WB* | I | EV | | SN* |
| | D | | | SN | | EV | | P |
| | D | | | SG | | | | SG |
| | P | | | P | | | | C |
| | C | | | I | | | | CO |
| | | | | E | | | | C |
| | | | | I | | | | C |
| | | | | I | | | | P |
| | | | | C | | | | C |
| | | | | I | | | | C |
| | | | | SG | | | | |
| | | | | P | | | | |
| | | | | CO | | | | |

## Third question attempted

| Subject | 9 | 3 | 6 | 7 | 10 | 4 | 5 | 8 |
|---|---|---|---|---|---|---|---|---|
| Question | B | C | C | C | C | D | D | D |
| | Q | Q | Q | SG | Q | Q | I | Q |
| | Q | I | I | SG | Q | D | D | Q |
| | ET | D | I | SG* | Q | D | Q | I |
| | D | D | D | I | SG | I | I | P |
| | I | D* | SG | WB* | I | Q | E | P |
| | P | Q | C | I | SG | P | I | D |
| | I | SG | C | D | ST | P | C | I* |
| | I | C | I | I | SG | C | I | WB* |
| | P | WB | P | D | ST | I | D | C |
| | I | Q | I | D | WB | D | Q | P |
| | ST | SG | WB* | D | I | Q | | EV |
| | QEV | I | C | C | C | P | | |
| | D | Q | WB | C | SG | C | | |
| | I | SG | I | D | SGEV | D | | |
| | I | C | SG | I | ST | D | | |
| | WB | SG* | SG | P | ST | D | | |
| | ET | C | SG* | C | I | P | | |
| | P | D | I | SGEV | IEV | I | | |
| | ET | I | I | C | WB | P | | |
| | WB | I | I | | QEV | C | | |
| | WB* | C | I* | | | | | |
| | I | C | | | | | | |
| | C | SG | | | | | | |
| | I | | | | | | | |
| | ST | | | | | | | |
| | ST | | | | | | | |
| | DEV | | | | | | | |
| | EV | | | | | | | |
| | ET | | | | | | | |

**Fourth question attempted**

| Subject | 6 | 9 | 10 | 4 | 8 | 5 | 3 | 7 |
|---|---|---|---|---|---|---|---|---|
| Question | A | A | A | B | B | C | D | D |
| | I | P | Q | Q | Q | SG | Q | Q |
| | C | Q | C | D | I | SG | I | I |
| | D | I | D | E | I | I | Q | P |
| | I | D | D | Q | ET | ST | P | D |
| | WB | P | C | E | ET | C | I | D |
| | WB | P | WB | E | SG | C | C | C |
| | P | D | ST | SG | ET | P | Q | D* |
| | I | P | I | I | I | SG | I | Q |
| | C | CO | WB | SG | ET | I | I | WB* |
| | SG | I | C | WB | ET | E | Q | Q |
| | SG | WB | WB | I | SG | I | I | Q |
| | C | C | CO | D | ET | SG | D | Q |
| | SN | WBEV | I | E | SG | | P | SN |
| | WB* | I | C | SG | ET | | D | I |
| | SN* | I | WB | I | SG | | P | D |
| | I | ST | WBEV | D | I | | D | P |
| | C | C | WB* | WB | | | P | EV |
| | I* | C | I | SG | | | I | |
| | C | C | CO | I | | | ST | |
| | P | EV | I* | SG | | | I* | |
| | I | C | C | I | | | SN* | |
| | I | SG | C | SG | | | SN* | |
| | SG | ST | C | I | | | I | |
| | E | SG | C | E | | | P | |
| | P | C | C | I | | | | |
| | SG | C | C | | | | | |
| | C | C | C | | | | | |
| | | STEV | C | | | | | |
| | | | SN* | | | | | |
| | | | C | | | | | |

## APPENDIX 2

## EXPERIMENT TWO MATERIALS and ANALYSES

**Experiment two - Questions (with outcome information)**

1.

In an experiment to test the effects of the drug Largacil 15 adult subjects were each given the standard therapeutic dose. Resulting blood pressure levels were compared with a placebo group of 15 subjects, using a t - test. Significance levels were set at 0.01. A significant reduction in blood pressure was found.

**Outcome :** Subsequent clinical trials confirmed this effect on blood pressure. The drug was licenced for clinical use and has made profits for this company.

| Quality score | % |
|---|---|

2.

The compound Ectarin B was tested for its effect on blood pressure using an experimental group of 50 adult subjects who each received a standard dose of the drug. Their resulting blood pressure levels were compared with those of a placebo control group of 50 adults, using a t - test. It was decided to use a statistical significance level of 0.01. The results showed a significant reduction in the blood pressure of the experimental group, compared to the control group.

**Outcome :** Subsequent clinical trials confirmed this effect on blood pressure. The drug was licenced for clinical use and has made profits for this company.

| Quality score | % |
|---|---|

3.

In the March 15th drug trial, Reprobin, a potential hypertension agonist was administered to 15 human subjects at therapeutic dose levels, (15 subjects recieved placebos). Significance levels were set at 0.05. An analysis of blood pressure measurements showed a significant reduction for the drug condition when compared to the placebo condition (t - test).

**Outcome :** Subsequent clinical trials confirmed this effect on blood pressure. The drug was licenced for clinical use and has made profits for this company.

| Quality score | % |
|---|---|

4.

Experimental testing of HTF 7 (Hypoteflin) demonstrated a significant reduction in blood pressure levels. The trial was conducted using adult subjects 50 in an experimental group, who were given HTF 7, and 50 in a placebo group. Blood pressure scores for each group were statistically compared using a t - test; the critical significance level had been set at 0.05.

**Outcome :** Subsequent clinical trials confirmed this effect on blood pressure. The drug was licenced for clinical use and has made profits for this company.

| Quality score | % |
|---|---|

5.

In an experimental trial to determine the efficiency of a new drug to reduce the severity of hypertension (high blood pressure), normal dose levels of Hypolax were given orally to 15 adult subjects; 15 received a placebo substitute. Resulting blood pressure measures were compared between these groups; significance was set at 0.01. Statistical analysis (t - test) failed to show a significant difference in the blood pressure scores.

**Outcome :** Subsequent clinical trials run by a competing company also found no effect on blood pressure. The drug was not licenced for clinical use.

| Quality score | % |
|---|---|

6.

In the first of a potential series of blood pressure trials Minigen was tested for its effects on human subjects. The experimental design used 50 subjects in an experimental (drug) condition and 50 control (placebo) subjects. The experimental hypothesis was that the experimental group would exhibit lower blood pressure at a significance level of 0.01. This hypothesis was not proved and the null hypothesis of no difference was accepted.

**Outcome :** Subsequent clinical trials also found no effect on blood pressure. The drug was not licenced for clinical use.

| Quality score | % |
|---|---|

7.

The initial clinical trial of Trimazole consisted of administration of doses at therapeutic levels, (5 mg orally), to 15 adult subjects in an experimental condition (and 15 placebo). Resulting blood pressure was compared with levels from a placebo group. Using t - test statistical analysis with a pre-set significance level of 0.05 no significant reduction in blood pressure was found for the drug condition.

**Outcome :** Subsequent clinical trials run by a competing company also found no effect on blood pressure and the drug was not licenced for clinical use.

| Quality score | % |
|---|---|

8.

An experiment was designed to test the hypothesis that 10 mg doses of Disigamine would significantly reduce blood pressure in human subjects. The experimental procedure consisted of either the drug or a placebo being given to two groups of 50 subjects each. The resulting blood pressure scores for drug and placebo groups were compared statistically; significance levels of 0.05 were chosen. No significant difference was found (t- test).

**Outcome :** Subsequent clinical trials run by a competing company also found no effect on blood pressure and the drug was not licenced for clinical use.

| Quality score | % |
|---|---|

9.

In an experiment to test the drug Lessophin, 15 adult subjects were each given the standard therapeutic dose. Resulting blood pressure levels were compared with a placebo group of 15 subjects, using a t - test. Significance levels were set at 0.01. A significant reduction in blood pressure was found.

**Outcome :** Subsequent clinical trials however showed there was no effect on blood pressure and the clinical licencing application failed at a significant cost.

| Quality score | % |
|---|---|

10.

Ectofligen was tested for its effect on blood pressure using an experimental group of 50 adult subjects these subjects each received a standard dose of the drug. Resulting blood pressure levels were compared with those of a placebo control group of 50 adults, using a t - test. It was decided to use a statistical significance level of 0.01. The results showed a significant reduction in the blood pressure of the experimental group compared to the control group.

**Outcome :** Subsequent clinical trials however, showed there was no effect on blood pressure and the clinical licencing application failed at a significant cost.

| Quality score | % |
|---|---|

11.

In an exploratory drug trial, Restanic, a potential treatment for hypertension was administered to 15 human subjects at therapeutic dose levels (15 subjects received placebos). Significance levels were set at 0.05. An analysis of blood pressure measurements showed a significant reduction for the drug condition when compared to the placebo condition (t - test).

**Outcome :** Subsequent clinical trials however showed there was no effect on blood pressure and the clinical licencing application failed at a significant cost.

| Quality score | % |
|---|---|

12.

A trial to test the effects of PRZ (Prozalin) was conducted using adult subjects, 50 in an experimental group who recieved PRZ, and 50 in a placebo group. Blood pressure scores for each group were statistically compared using a t - test, the critical significance level had been set at 0.05. The trial demonstrated a significant reduction in blood pressure levels.

**Outcome :** Subsequent clinical trials however showed there was no effect on blood pressure and the clinical licencing application failed at a significant cost.

| Quality score | % |
|---|---|

13.

In an experimental trial of a new drug to reduce the severity of high blood pressure normal dose levels of Hyperstem were administered to 15 subjects. A placebo substitute was administered to another 15 subjects. Resulting blood pressure measures were compared between these groups using a t - test; significance was set at 0.01. Statistical analysis failed to show a significant difference in the blood pressure scores.

**Outcome :** The company involved cancelled testing. Subsequent clinical trials run by a competitor did find an effect on blood pressure. The drug was licenced for clinical use and has made significant profits for the competing company.

| Quality score | % |
|---|---|

14.

Mestamine was tested for its effects on blood pressure in human subjects. The experimental design involved 100 subjects, 50 experimental (drug) and 50 control (placebo). The experimental hypothesis was that Mestamine would lower blood pressure, a significance level of 0.01 was set. This hypothesis was not proved the null hypothesis was accepted (t - test).

**Outcome :** The company involved cancelled testing. Subsequent clinical trials run by a competitor did find an effect on blood pressure. The drug was licenced for clinical use and has made significant profits for the competing company.

| Quality score | % |
|---|---|

15.

An experimental trial of Triole-1 consisted of administration of doses at therapeutic levels to 15 adult subjects in an experimental group. Resulting blood pressure measures were compared with those of a placebo group (15 subjects). Using a t - test for statistical analysis, with a pre-set significance level of 0.05 no significant reduction in blood pressure was found for the drug condition.

**Outcome :** The company involved cancelled testing. Subsequent clinical trials run by a competitor did find an effect on blood pressure. The drug was licenced for clinical use and has made significant profits for the competing company.

| Quality score | % |
|---|---|

16.

The hypothesis that 10 mg doses of Lipramine would significantly reduce blood pressure in human subjects was experimentally tested. The trial consisted of either the drug or a placebo being given to two groups of 50 subjects each. In this design significance levels were set at 0.05. When the blood pressure scores for drug and placebo groups were compared statistically results failed to show a significant difference (t - test).

**Outcome :** The company involved cancelled testing. Subsequent clinical trials run by a competitor did find an effect on blood pressure. The drug was licenced for clinical use and has made significant profits for the competing company.

| Quality score | % |
|---|---|

**Experiment two - Long-term outcome information**

All four possible outcomes were as follows:

**Outcome :** Subsequent clinical trials confirmed this effect on blood pressure. The drug was licensed for clinical use and has made profits for this company.

**Outcome :** Subsequent clinical trials run by a competing company also found no effect on blood pressure. The drug was not licensed for clinical use.

**Outcome :** Subsequent clinical trials however showed there was no effect on blood pressure and the clinical licensing application failed at a significant cost.

**Outcome :** The company involved cancelled testing. Subsequent clinical trials run by a competitor did find an effect on blood pressure. The drug was licensed for clinical use and has made significant profits for the competing company.

The subject group given long term outcome information received the same experimental descriptions (see above) with the addition of one of these sections counterbalanced across question type.

## Experiment two - anova table for complete analysis of all subjects

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Outcome Info | 1 | 2269.695 | 2269.695 | .714 | .4010 |
| Subject(Group) | 70 | 222475.797 | 3178.226 | | |
| Outcome | 1 | 6646.084 | 6646.084 | 15.887 | .0002 |
| Outcome * Outcome Info | 1 | 6342.195 | 6342.195 | 15.161 | .0002 |
| Outcome * Subject(Group) | 70 | 29282.533 | 418.322 | | |
| Result | 1 | 2559.105 | 2559.105 | 5.968 | .0171 |
| Result * Outcome Info | 1 | 73.508 | 73.508 | .171 | .6801 |
| Result * Subject(Group) | 70 | 30018.450 | 428.835 | | |
| Level | 1 | 6646.084 | 6646.084 | 16.080 | .0001 |
| Level * Outcome Info | 1 | 44.730 | 44.730 | .108 | .7432 |
| Level * Subject(Group) | 70 | 28932.748 | 413.325 | | |
| Ssnumber | 1 | 35366.918 | 35366.918 | 59.048 | .0001 |
| Ssnumber * Outcome Info | 1 | 250.320 | 250.320 | .418 | .5201 |
| Ssnumber * Subject(Group) | 70 | 41926.325 | 598.947 | | |
| Outcome * Result | 1 | 265.459 | 265.459 | .950 | .3332 |
| Outcome * Result * Outcome Info | 1 | 1514.793 | 1514.793 | 5.419 | .0228 |
| Outcome * Result * Subject(Group) | 70 | 19567.561 | 279.537 | | |
| Outcome * Level | 1 | 82.883 | 82.883 | .560 | .4569 |
| Outcome * Level * Outcome Info | 1 | .459 | .459 | .003 | .9558 |
| Outcome * Level * Subject(Group) | 70 | 10366.470 | 148.092 | | |
| Result * Level | 1 | 529.480 | 529.480 | 2.885 | .0938 |
| Result * Level * Outcome Info | 1 | 336.918 | 336.918 | 1.836 | .1798 |
| Result * Level * Subject(Group) | 70 | 12846.165 | 183.517 | | |
| Outcome * Ssnumber | 1 | 5.980 | 5.980 | .038 | .8466 |
| Outcome * Ssnumber * Outcome Info | 1 | 47.938 | 47.938 | .302 | .5841 |
| Outcome * Ssnumber * Subject(Group) | 70 | 11095.894 | 158.513 | | |
| Result * Ssnumber | 1 | 471.501 | 471.501 | 2.427 | .1237 |
| Result * Ssnumber * Outcome Info | 1 | 175.001 | 175.001 | .901 | .3458 |
| Result * Ssnumber * Subject(Group) | 70 | 13597.061 | 194.244 | | |
| Level * Ssnumber | 1 | 219.626 | 219.626 | 1.420 | .2374 |
| Level * Ssnumber * Outcome Info | 1 | 15.355 | 15.355 | .099 | .7536 |
| Level * Ssnumber * Subject(Group) | 70 | 10824.582 | 154.637 | | |
| Outcome * Result * Level | 1 | 275.147 | 275.147 | 2.353 | .1296 |
| Outcome * Result * Level * Outcome Info | 1 | 34.376 | 34.376 | .294 | .5894 |
| Outcome * Result * Level * Subject(Group) | 70 | 8185.790 | 116.940 | | |
| Outcome * Result * Ssnumber | 1 | 8.168 | 8.168 | .057 | .8127 |
| Outcome * Result * Ssnumber * Outcome Info | 1 | 187.695 | 187.695 | 1.300 | .2581 |
| Outcome * Result * Ssnumber * Subject(Group) | 70 | 10107.450 | 144.392 | | |
| Outcome * Level * Ssnumber | 1 | 14.001 | 14.001 | .102 | .7502 |
| Outcome * Level * Ssnumber * Outcome Info | 1 | 453.758 | 453.758 | 3.312 | .0731 |
| Outcome * Level * Ssnumber * Subject(Group) | 70 | 9591.054 | 137.015 | | |
| Result * Level * Ssnumber | 1 | 18.251 | 18.251 | .134 | .7155 |
| Result * Level * Ssnumber * Outcome Info | 1 | 115.647 | 115.647 | .849 | .3601 |
| Result * Level * Ssnumber * Subject(Group) | 70 | 9538.165 | 136.259 | | |
| Outcome * Result * Level * Ssnumber | 1 | 216.147 | 216.147 | 2.296 | .1342 |
| Outcome * Result * Level * Ssnumber * Outc... | 1 | 19.793 | 19.793 | .210 | .6480 |
| Outcome * Result * Level * Ssnumber * Subj... | 70 | 6590.873 | 94.155 | | |

Dependent: Quality

## Experiment two - anova table for analysis of subjects recieving only initial experimental result

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 35 | 116268.375 | 3321.954 | | |
| Outcome | 1 | 1.778 | 1.778 | .013 | .9112 |
| Outcome * Subject | 35 | 4928.597 | 140.817 | | |
| Result | 1 | 1750.028 | 1750.028 | 3.575 | .0669 |
| Result * Subject | 35 | 17131.597 | 489.474 | | |
| Level | 1 | 2800.174 | 2800.174 | 6.323 | .0167 |
| Level * Subject | 35 | 15498.951 | 442.827 | | |
| Ssnumber | 1 | 20784.028 | 20784.028 | 40.752 | .0001 |
| Ssnumber * Subject | 35 | 17850.347 | 510.010 | | |
| Outcome * Result | 1 | 256.000 | 256.000 | 2.160 | .1506 |
| Outcome * Result * Subject | 35 | 4148.375 | 118.525 | | |
| Outcome * Level | 1 | 47.840 | 47.840 | .387 | .5378 |
| Outcome * Level * Subject | 35 | 4324.535 | 123.558 | | |
| Result * Level | 1 | 855.563 | 855.563 | 6.066 | .0188 |
| Result * Level * Subject | 35 | 4936.562 | 141.045 | | |
| Outcome * Ssnumber | 1 | 10.028 | 10.028 | .085 | .7728 |
| Outcome * Ssnumber * Subject | 35 | 4145.097 | 118.431 | | |
| Result * Ssnumber | 1 | 36.000 | 36.000 | .189 | .6664 |
| Result * Ssnumber * Subject | 35 | 6664.875 | 190.425 | | |
| Level * Ssnumber | 1 | 175.563 | 175.563 | .966 | .3323 |
| Level * Ssnumber * Subject | 35 | 6357.813 | 181.652 | | |
| Outcome * Result * Level | 1 | 57.507 | 57.507 | .467 | .4991 |
| Outcome * Result * Level * Subject | 35 | 4313.868 | 123.253 | | |
| Outcome * Result * Ssnumber | 1 | 58.778 | 58.778 | .420 | .5214 |
| Outcome * Result * Ssnumber * Subject | 35 | 4903.847 | 140.110 | | |
| Outcome * Level * Ssnumber | 1 | 154.174 | 154.174 | 1.180 | .2849 |
| Outcome * Level * Ssnumber * Subject | 35 | 4574.451 | 130.699 | | |
| Result * Level * Ssnumber | 1 | 21.007 | 21.007 | .183 | .6714 |
| Result * Level * Ssnumber * Subject | 35 | 4017.868 | 114.796 | | |
| Outcome * Result * Level * Ssnumber | 1 | 52.562 | 52.562 | .687 | .4128 |
| Outcome * Result * Level * Ssnumber * ... | 35 | 2677.562 | 76.502 | | |

Dependent: Quality

## Experiment two - means table for analysis of subjects recieving only initial experimental result

**Means Table**
**Effect: Result * Level * Ssnumber**
**Dependent: Quality**

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Significant, Sig .01, Ss-30 | 72 | 56.111 | 17.213 | 2.029 |
| Significant, Sig .01, Ss-100 | 72 | 69.347 | 20.332 | 2.396 |
| Significant, Sig .05, Ss-30 | 72 | 49.986 | 18.187 | 2.143 |
| Significant, Sig .05, Ss-100 | 72 | 61.778 | 20.696 | 2.439 |
| Non Sig, Sig .01, Ss-30 | 72 | 50.306 | 18.883 | 2.225 |
| Non Sig, Sig .01, Ss-100 | 72 | 63.306 | 21.469 | 2.530 |
| Non Sig, Sig .05, Ss-30 | 72 | 49.819 | 20.966 | 2.471 |
| Non Sig, Sig .05, Ss-100 | 72 | 59.847 | 20.504 | 2.416 |

## Experiment two - analysis of interaction between experimental result and pre-set significance level (subjects recieving only initial experimental result)

**Least Squares Means Table**
**Effect: Result * Level**
**Dependent: Quality**

| | Vs. | Diff. | Std. Error | t-Test | P-Value |
|---|---|---|---|---|---|
| Significant, Sig .01 | Significant, Sig .05 | 6.847 | 1.400 | 4.892 | .0001 |
| | Non Sig, Sig .01 | 5.924 | 1.400 | 4.232 | .0002 |
| | Non Sig, Sig .05 | 7.896 | 1.400 | 5.641 | .0001 |
| Significant, Sig .05 | Non Sig, Sig .01 | -.924 | 1.400 | -.660 | .5136 |
| | Non Sig, Sig .05 | 1.049 | 1.400 | .749 | .4587 |
| Non Sig, Sig .01 | Non Sig, Sig .05 | 1.972 | 1.400 | 1.409 | .1676 |

## Experiment two - anova table for analysis of subjects recieving initial experimental result and long-term outcome information

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 35 | 106207.422 | 3034.498 | | |
| Outcome | 1 | 12986.502 | 12986.502 | 18.663 | .0001 |
| Outcome * Subject | 35 | 24353.936 | 695.827 | | |
| Result | 1 | 882.585 | 882.585 | 2.397 | .1306 |
| Result * Subject | 35 | 12886.852 | 368.196 | | |
| Level | 1 | 3890.641 | 3890.641 | 10.137 | .0030 |
| Level * Subject | 35 | 13433.797 | 383.823 | | |
| Ssnumber | 1 | 14833.210 | 14833.210 | 21.564 | .0001 |
| Ssnumber * Subject | 35 | 24075.977 | 687.885 | | |
| Outcome * Result | 1 | 1524.252 | 1524.252 | 3.460 | .0713 |
| Outcome * Result * Subject | 35 | 15419.186 | 440.548 | | |
| Outcome * Level | 1 | 35.502 | 35.502 | .206 | .6530 |
| Outcome * Level * Subject | 35 | 6041.936 | 172.627 | | |
| Result * Level | 1 | 10.835 | 10.835 | .048 | .8280 |
| Result * Level * Subject | 35 | 7909.602 | 225.989 | | |
| Outcome * Ssnumber | 1 | 43.891 | 43.891 | .221 | .6412 |
| Outcome * Ssnumber * Subject | 35 | 6950.797 | 198.594 | | |
| Result * Ssnumber | 1 | 610.502 | 610.502 | 3.082 | .0879 |
| Result * Ssnumber * Subject | 35 | 6932.186 | 198.062 | | |
| Level * Ssnumber | 1 | 59.418 | 59.418 | .466 | .4995 |
| Level * Ssnumber * Subject | 35 | 4466.769 | 127.622 | | |
| Outcome * Result * Level | 1 | 252.016 | 252.016 | 2.278 | .1402 |
| Outcome * Result * Level * Subject | 35 | 3871.922 | 110.626 | | |
| Outcome * Result * Ssnumber | 1 | 137.085 | 137.085 | .922 | .3435 |
| Outcome * Result * Ssnumber * Subject | 35 | 5203.602 | 148.674 | | |
| Outcome * Level * Ssnumber | 1 | 313.585 | 313.585 | 2.188 | .1480 |
| Outcome * Level * Ssnumber * Subject | 35 | 5016.602 | 143.331 | | |
| Result * Level * Ssnumber | 1 | 112.891 | 112.891 | .716 | .4033 |
| Result * Level * Ssnumber * Subject | 35 | 5520.297 | 157.723 | | |
| Outcome * Result * Level * Ssnumber | 1 | 183.377 | 183.377 | 1.640 | .2087 |
| Outcome * Result * Level * Ssnumber * Subject | 35 | 3913.311 | 111.809 | | |

Dependent: Quality

## Experiment two - means table for analysis of subjects recieving initial experimental result and long-term outcome information

**Means Table**
**Effect: Outcome * Result * Level * Ssnumber**
**Dependent: Quality**

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Replicated, Significant, Sig .01, Ss-30 | 36 | 56.389 | 24.862 | 4.144 |
| Replicated, Significant, Sig .01, Ss-100 | 36 | 72.000 | 17.180 | 2.863 |
| Replicated, Significant, Sig .05, Ss-30 | 36 | 54.611 | 22.967 | 3.828 |
| Replicated, Significant, Sig .05, Ss-100 | 36 | 66.472 | 18.767 | 3.128 |
| Replicated, Non Sig, Sig .01, Ss-30 | 36 | 54.500 | 22.120 | 3.687 |
| Replicated, Non Sig, Sig .01, Ss-100 | 36 | 64.528 | 22.122 | 3.687 |
| Replicated, Non Sig, Sig .05, Ss-30 | 36 | 51.111 | 23.688 | 3.948 |
| Replicated, Non Sig, Sig .05, Ss-100 | 36 | 56.417 | 22.435 | 3.739 |
| Not replicated, Significant, Sig .01, Ss-30 | 36 | 47.333 | 17.752 | 2.959 |
| Not replicated, Significant, Sig .01, Ss-100 | 36 | 59.194 | 24.371 | 4.062 |
| Not replicated, Significant, Sig .05, Ss-30 | 36 | 41.222 | 17.925 | 2.988 |
| Not replicated, Significant, Sig .05, Ss-100 | 36 | 50.722 | 21.530 | 3.588 |
| Not replicated, Non Sig, Sig .01, Ss-30 | 36 | 49.611 | 19.494 | 3.249 |
| Not replicated, Non Sig, Sig .01, Ss-100 | 36 | 55.278 | 21.663 | 3.611 |
| Not replicated, Non Sig, Sig .05, Ss-30 | 36 | 42.667 | 21.652 | 3.609 |
| Not replicated, Non Sig, Sig .05, Ss-100 | 36 | 54.028 | 18.913 | 3.152 |

## Experiment two - analysis of interaction between experimental result and long-term outcome (subjects recieving long-term outcome information)

**Least Squares Means Table**
**Effect: Outcome * Result**
**Dependent: Quality**

| | Vs. | Diff. | Std. Error | t-Test | P-Value |
|---|---|---|---|---|---|
| Replicated, Significant | Replicated, Non Sig | 5.729 | 2.474 | 2.316 | .0265 |
| | Not replicated, Significant | 12.750 | 2.474 | 5.154 | .0001 |
| | Not replicated, Non Sig | 11.972 | 2.474 | 4.840 | .0001 |
| Replicated, Non Sig | Not replicated, Significant | 7.021 | 2.474 | 2.838 | .0075 |
| | Not replicated, Non Sig | 6.243 | 2.474 | 2.524 | .0163 |
| Not replicated, Significant | Not replicated, Non Sig | -.778 | 2.474 | -.314 | .7551 |

# APPENDIX 3

## EXPERIMENTS THREE and FOUR MATERIALS and ANALYSES

### Experiment three (and four) - task instructions

#### Introduction

The following description and series of questions are designed to study understanding of the concept of quality in experimental sampling.

Your co-operation in this experiment is greatly appreciated, and the results should be of use to you as they will demonstrate general levels of understanding of sampling principles. Results and explanations will be made available as soon as possible.

You do not need to put your name on the answer sheets but please print your name on this instruction sheet and hand it back separately at the end.

#### Scenario

A worker in a roulette wheel factory had been bribed to produce a number of wheels in which the slots for even numbers were slightly larger than those for odd numbers.

It was the intention of a well financed team of fraudsters to bet heavily on even numbers on the biased wheels. Half the casinos ordering new wheels had received the biased wheels. However due to security measures it was impossible for the team to find out which of these casinos received true or biased wheels.

It was decided that each member of the team would go to a different casino and make a number of small bets whilst counting the number of even wins to see whether the new wheel was one of the biased ones. Simply standing at the table without betting was regarded as too suspicious as casino security consistently video and study behaviour at roulette tables. If it was decided that the wheel was biased towards even numbers then the rest of the team would return to make a great number of large bets on even numbers.

Tests involving extensive counting of even wins can be very expensive to run and dangerous in terms of alerting the casino to a specific interest in the roulette wheel. The potential cost of the larger bets is high.

#### Test bets

In initial tests sampling methods varied between individuals. Different numbers of

spins were counted and there were also differences in the percentage of even number wins which were regarded as sufficient evidence of a biased table. On an unbiased table even numbers have a 50% probability of winning if zero wins are disregarded; biased tables have a greater probability of even wins. Below are a series of descriptions of the tests used by different team members.

**Task**

Your task is to read each description and give an intuitive estimate of the quality of the test for a biased roulette wheel by writing a percentage mark (from 0 to 100) in the space provided.

**A test of high quality is one which has a high probability of differentiating whether a table is truly biased or not without incurring excessive cost.**

Please try to use the whole range of marks.

**Experiment three (and four) - Questions (without outcome information)**

1

In the Monte Carlo Grand casino 300 spins of the roulette wheel were counted. The team member involved decided to regard anything higher than 56% even wins as sufficient evidence of a biased table. A bias to even numbers was found.

| Quality score | % |
|---|---|

2

In the Paris Metropole casino 300 spins of the roulette wheel were counted. The team member involved decided to regard anything higher than 54% even wins as sufficient evidence of a biased table. No bias to even numbers was found.

| Quality score | % |
|---|---|

3

In the London Alhambra casino 700 spins of the roulette wheel were counted. The team member involved decided to regard anything higher than 56% even wins as sufficient evidence of a biased table. A bias to even numbers was found.

| Quality score | % |
|---|---|

4

In the Frankfurt Main casino 700 spins of the roulette wheel were counted. The team member involved decided to regard anything higher than 54% even wins as sufficient evidence of a biased table. No bias to even numbers was found.

| Quality score | % |
|---|---|

**Experiment three (and four) - Long-term outcome information**

All four possible outcomes were as follows:

**Outcome :** Subsequent large bets confirmed this wheel to be biased. The wheel has made large profits for the team.

**Outcome :** Subsequent betting by customers confirmed this finding. The table was not biased.

**Outcome :** Subsequent large bets by the team however showed no evidence of bias. A significant amount of money was lost.

**Outcome :** The team did not go ahead with large scale bets. However subsequent betting by customers did show a bias towards even numbers and the wheel was eventually replaced.

The subject group given long term outcome information received the same experimental descriptions (see above) with the addition of one of these sections counterbalanced across question type.

## Experiment three - anova table for complete analysis of all subjects

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Group | 1 | 15277.131 | 15277.131 | 4.549 | .0364 |
| Subject(Group) | 71 | 238424.774 | 3358.095 | | |
| Outcome | 1 | 12831.721 | 12831.721 | 25.486 | .0001 |
| Outcome * Group | 1 | 16269.976 | 16269.976 | 32.314 | .0001 |
| Outcome * Subject(Group) | 71 | 35747.796 | 503.490 | | |
| Result | 1 | 32.359 | 32.359 | .056 | .8131 |
| Result * Group | 1 | 2473.220 | 2473.220 | 4.305 | .0416 |
| Result * Subject(Group) | 71 | 40790.038 | 574.508 | | |
| Level | 1 | 10083.047 | 10083.047 | 29.360 | .0001 |
| Level * Group | 1 | 1448.415 | 1448.415 | 4.217 | .0437 |
| Level * Subject(Group) | 71 | 24383.768 | 343.433 | | |
| Spins | 1 | 16870.768 | 16870.768 | 24.403 | .0001 |
| Spins * Group | 1 | 1826.584 | 1826.584 | 2.642 | .1085 |
| Spins * Subject(Group) | 71 | 49085.105 | 691.340 | | |
| Outcome * Result | 1 | 2628.089 | 2628.089 | 12.542 | .0007 |
| Outcome * Result * Group | 1 | 1687.080 | 1687.080 | 8.051 | .0059 |
| Outcome * Result * Subject(Group) | 71 | 14877.794 | 209.546 | | |
| Outcome * Level | 1 | 90.341 | 90.341 | 1.481 | .2276 |
| Outcome * Level * Group | 1 | .408 | .408 | .007 | .9351 |
| Outcome * Level * Subject(Group) | 71 | 4329.850 | 60.984 | | |
| Result * Level | 1 | 51.660 | 51.660 | .533 | .4678 |
| Result * Level * Group | 1 | 59.812 | 59.812 | .617 | .4348 |
| Result * Level * Subject(Group) | 71 | 6882.668 | 96.939 | | |
| Outcome * Spins | 1 | 44.957 | 44.957 | .420 | .5190 |
| Outcome * Spins * Group | 1 | 233.493 | 233.493 | 2.182 | .1441 |
| Outcome * Spins * Subject(Group) | 71 | 7598.689 | 107.024 | | |
| Result * Spins | 1 | 39.653 | 39.653 | .791 | .3769 |
| Result * Spins * Group | 1 | 544.254 | 544.254 | 10.850 | .0015 |
| Result * Spins * Subject(Group) | 71 | 3561.339 | 50.160 | | |
| Level * Spins | 1 | 108.315 | 108.315 | .792 | .3766 |
| Level * Spins * Group | 1 | 2.375 | 2.375 | .017 | .8956 |
| Level * Spins * Subject(Group) | 71 | 9714.876 | 136.829 | | |
| Outcome * Result * Level | 1 | .032 | .032 | 3.042E-4 | .9861 |
| Outcome * Result * Level * Group | 1 | 4.226 | 4.226 | .040 | .8418 |
| Outcome * Result * Level * Subject(Group) | 71 | 7476.916 | 105.309 | | |
| Outcome * Result * Spins | 1 | 16.664 | 16.664 | .273 | .6032 |
| Outcome * Result * Spins * Group | 1 | 22.381 | 22.381 | .366 | .5470 |
| Outcome * Result * Spins * Subject(Group) | 71 | 4339.900 | 61.125 | | |
| Outcome * Level * Spins | 1 | 36.105 | 36.105 | .524 | .4715 |
| Outcome * Level * Spins * Group | 1 | 215.480 | 215.480 | 3.128 | .0812 |
| Outcome * Level * Spins * Subject(Group) | 71 | 4890.624 | 68.882 | | |
| Result * Level * Spins | 1 | 67.010 | 67.010 | .649 | .4233 |
| Result * Level * Spins * Group | 1 | 96.902 | 96.902 | .938 | .3360 |
| Result * Level * Spins * Subject(Group) | 71 | 7333.739 | 103.292 | | |
| Outcome * Result * Level * Spins | 1 | 318.596 | 318.596 | 5.375 | .0233 |
| Outcome * Result * Level * Spins * Group | 1 | .283 | .283 | .005 | .9451 |
| Outcome * Result * Level * Spins * Subjec... | 71 | 4208.091 | 59.269 | | |

Dependent: Quality

## Experiment three - means table for overall analysis

| | Count | Mean | Std. Dev. | Std. Erro |
|---|---|---|---|---|
| Correct, Bias found, %56, Spin300, No-outcome information | 44 | 46.852 | 20.658 | 3.11 |
| Correct, Bias found, %56, Spin300, Outcome information | 29 | 67.241 | 16.440 | 3.05 |
| Correct, Bias found, %56, Spin700, No-outcome information | 44 | 57.000 | 22.788 | 3.43 |
| Correct, Bias found, %56, Spin700, Outcome information | 29 | 69.069 | 16.966 | 3.15 |
| Correct, Bias found, %54, Spin300, No-outcome information | 44 | 43.432 | 21.427 | 3.23 |
| Correct, Bias found, %54, Spin300, Outcome information | 29 | 57.552 | 21.030 | 3.90 |
| Correct, Bias found, %54, Spin700, No-outcome information | 44 | 53.886 | 23.385 | 3.52 |
| Correct, Bias found, %54, Spin700, Outcome information | 29 | 65.310 | 18.486 | 3.43 |
| Correct, No Bias, %56, Spin300, No-outcome information | 44 | 43.318 | 16.884 | 2.54 |
| Correct, No Bias, %56, Spin300, Outcome information | 29 | 62.448 | 15.968 | 2.96 |
| Correct, No Bias, %56, Spin700, No-outcome information | 44 | 52.398 | 21.612 | 3.25 |
| Correct, No Bias, %56, Spin700, Outcome information | 29 | 70.138 | 15.172 | 2.81 |
| Correct, No Bias, %54, Spin300, No-outcome information | 44 | 40.102 | 16.536 | 2.49 |
| Correct, No Bias, %54, Spin300, Outcome information | 29 | 53.103 | 15.532 | 2.88 |
| Correct, No Bias, %54, Spin700, No-outcome information | 44 | 49.659 | 21.178 | 3.19 |
| Correct, No Bias, %54, Spin700, Outcome information | 29 | 61.931 | 20.683 | 3.84 |
| Incorrect, Bias found, %56, Spin300, No-outcome information | 44 | 46.034 | 19.685 | 2.96 |
| Incorrect, Bias found, %56, Spin300, Outcome information | 29 | 47.000 | 21.883 | 4.06 |
| Incorrect, Bias found, %56, Spin700, No-outcome information | 44 | 59.295 | 23.871 | 3.59 |
| Incorrect, Bias found, %56, Spin700, Outcome information | 29 | 50.793 | 26.140 | 4.85 |
| Incorrect, Bias found, %54, Spin300, No-outcome information | 44 | 42.659 | 21.358 | 3.22 |
| Incorrect, Bias found, %54, Spin300, Outcome information | 29 | 40.655 | 23.657 | 4.39 |
| Incorrect, Bias found, %54, Spin700, No-outcome information | 44 | 54.159 | 22.120 | 3.33 |
| Incorrect, Bias found, %54, Spin700, Outcome information | 29 | 41.034 | 22.138 | 4.11 |
| Incorrect, No Bias, %56, Spin300, No-outcome information | 44 | 47.068 | 17.447 | 2.63 |
| Incorrect, No Bias, %56, Spin300, Outcome information | 29 | 55.103 | 18.135 | 3.36 |
| Incorrect, No Bias, %56, Spin700, No-outcome information | 44 | 53.011 | 22.020 | 3.32 |
| Incorrect, No Bias, %56, Spin700, Outcome information | 29 | 60.517 | 16.013 | 2.97 |
| Incorrect, No Bias, %54, Spin300, No-outcome information | 44 | 39.307 | 17.678 | 2.66 |
| Incorrect, No Bias, %54, Spin300, Outcome information | 29 | 45.241 | 18.458 | 3.42 |
| Incorrect, No Bias, %54, Spin700, No-outcome information | 44 | 51.943 | 21.019 | 3.16 |
| Incorrect, No Bias, %54, Spin700, Outcome information | 29 | 51.241 | 19.244 | 3.57 |

## Experiment three - anova table for analysis of subjects recieving only initial experimental result

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 43 | 186266.848 | 4331.787 | | |
| Outcome | 1 | 128.267 | 128.267 | 1.440 | .2367 |
| Outcome * Subject | 43 | 3829.748 | 89.064 | | |
| Result | 1 | 1932.844 | 1932.844 | 3.062 | .0873 |
| Result * Subject | 43 | 27140.422 | 631.173 | | |
| Level | 1 | 2446.955 | 2446.955 | 12.033 | .0012 |
| Level * Subject | 43 | 8743.936 | 203.347 | | |
| Spins | 1 | 18753.299 | 18753.299 | 23.333 | .0001 |
| Spins * Subject | 43 | 34560.092 | 803.723 | | |
| Outcome * Result | 1 | 65.355 | 65.355 | .849 | .3619 |
| Outcome * Result * Subject | 43 | 3308.660 | 76.946 | | |
| Outcome * Level | 1 | 64.748 | 64.748 | 1.285 | .2632 |
| Outcome * Level * Subject | 43 | 2166.018 | 50.373 | | |
| Result * Level | 1 | .188 | .188 | .002 | .9647 |
| Result * Level * Subject | 43 | 4086.078 | 95.025 | | |
| Outcome * Spins | 1 | 46.279 | 46.279 | .757 | .3891 |
| Outcome * Spins * Subject | 43 | 2628.487 | 61.128 | | |
| Result * Spins | 1 | 182.560 | 182.560 | 3.226 | .0795 |
| Result * Spins * Subject | 43 | 2433.706 | 56.598 | | |
| Level * Spins | 1 | 89.847 | 89.847 | .603 | .4417 |
| Level * Spins * Subject | 43 | 6407.044 | 149.001 | | |
| Outcome * Result * Level | 1 | 2.216 | 2.216 | .048 | .8276 |
| Outcome * Result * Level * Subject | 43 | 1985.549 | 46.176 | | |
| Outcome * Result * Spins | 1 | 48.878 | 48.878 | .839 | .3648 |
| Outcome * Result * Spins * Subject | 43 | 2504.887 | 58.253 | | |
| Outcome * Level * Spins | 1 | 47.310 | 47.310 | .749 | .3916 |
| Outcome * Level * Spins * Subject | 43 | 2716.206 | 63.168 | | |
| Result * Level * Spins | 1 | 204.574 | 204.574 | 1.929 | .1721 |
| Result * Level * Spins * Subject | 43 | 4560.941 | 106.068 | | |
| Outcome * Result * Level * Spins | 1 | 188.722 | 188.722 | 3.721 | .0603 |
| Outcome * Result * Level * Spins * Subject | 43 | 2180.794 | 50.716 | | |

Dependent: Quality

## Experiment three - anova table for analysis of subjects recieving initial experimental result and long-term outcome information

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 28 | 52157.927 | 1862.783 | | |
| Outcome | 1 | 24056.640 | 24056.640 | 21.104 | .0001 |
| Outcome * Subject | 28 | 31918.047 | 1139.930 | | |
| Result | 1 | 804.571 | 804.571 | 1.650 | .2094 |
| Result * Subject | 28 | 13649.616 | 487.486 | | |
| Level | 1 | 7953.106 | 7953.106 | 14.238 | .0008 |
| Level * Subject | 28 | 15639.832 | 558.565 | | |
| Spins | 1 | 3150.175 | 3150.175 | 6.073 | .0201 |
| Spins * Subject | 28 | 14525.013 | 518.750 | | |
| Outcome * Result | 1 | 3536.554 | 3536.554 | 8.559 | .0067 |
| Outcome * Result * Subject | 28 | 11569.134 | 413.183 | | |
| Outcome * Level | 1 | 32.606 | 32.606 | .422 | .5213 |
| Outcome * Level * Subject | 28 | 2163.832 | 77.280 | | |
| Result * Level | 1 | 92.347 | 92.347 | .925 | .3445 |
| Result * Level * Subject | 28 | 2796.591 | 99.878 | | |
| Outcome * Spins | 1 | 200.485 | 200.485 | 1.129 | .2970 |
| Outcome * Spins * Subject | 28 | 4970.203 | 177.507 | | |
| Result * Spins | 1 | 364.054 | 364.054 | 9.040 | .0055 |
| Result * Spins * Subject | 28 | 1127.634 | 40.273 | | |
| Level * Spins | 1 | 32.606 | 32.606 | .276 | .6035 |
| Level * Spins * Subject | 28 | 3307.832 | 118.137 | | |
| Outcome * Result * Level | 1 | 2.071 | 2.071 | .011 | .9189 |
| Outcome * Result * Level * Subject | 28 | 5491.366 | 196.120 | | |
| Outcome * Result * Spins | 1 | .175 | .175 | .003 | .9592 |
| Outcome * Result * Spins * Subject | 28 | 1835.013 | 65.536 | | |
| Outcome * Level * Spins | 1 | 177.519 | 177.519 | 2.286 | .1418 |
| Outcome * Level * Spins * Subject | 28 | 2174.418 | 77.658 | | |
| Result * Level * Spins | 1 | 1.140 | 1.140 | .012 | .9153 |
| Result * Level * Spins * Subject | 28 | 2772.797 | 99.028 | | |
| Outcome * Result * Level * Spins | 1 | 140.140 | 140.140 | 1.936 | .1751 |
| Outcome * Result * Level * Spins * Subject | 28 | 2027.297 | 72.403 | | |

Dependent: Quality

**Experiment four - anova table for analysis of subjects recieving initial experimental result and long-term outcome information (all subjects)**

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 39 | 57006.480 | 1461.705 | | |
| Outcome | 1 | 25750.550 | 25750.550 | 23.308 | .0001 |
| Outcome * Subject | 39 | 43087.660 | 1104.812 | | |
| Result | 1 | 49.729 | 49.729 | .098 | .7556 |
| Result * Subject | 39 | 19738.481 | 506.115 | | |
| Level | 1 | 122.150 | 122.150 | .352 | .5566 |
| Level * Subject | 39 | 13544.560 | 347.296 | | |
| Spins | 1 | 11092.230 | 11092.230 | 22.169 | .0001 |
| Spins * Subject | 39 | 19513.730 | 500.352 | | |
| Outcome * Result | 1 | 1205.604 | 1205.604 | 4.697 | .0364 |
| Outcome * Result * Subject | 39 | 10010.356 | 256.676 | | |
| Outcome * Level | 1 | .110 | .110 | .001 | .9729 |
| Outcome * Level * Subject | 39 | 3682.350 | 94.419 | | |
| Result * Level | 1 | 88.209 | 88.209 | .433 | .5144 |
| Result * Level * Subject | 39 | 7945.251 | 203.724 | | |
| Outcome * Spins | 1 | 648.830 | 648.830 | 2.729 | .1066 |
| Outcome * Spins * Subject | 39 | 9273.380 | 237.779 | | |
| Result * Spins | 1 | 8.464 | 8.464 | .054 | .8171 |
| Result * Spins * Subject | 39 | 6085.246 | 156.032 | | |
| Level * Spins | 1 | 41.820 | 41.820 | .318 | .5760 |
| Level * Spins * Subject | 39 | 5128.390 | 131.497 | | |
| Outcome * Result * Level | 1 | 263.169 | 263.169 | 3.060 | .0881 |
| Outcome * Result * Level * Subject | 39 | 3354.541 | 86.014 | | |
| Outcome * Result * Spins | 1 | 104.329 | 104.329 | .688 | .4119 |
| Outcome * Result * Spins * Subject | 39 | 5915.131 | 151.670 | | |
| Outcome * Level * Spins | 1 | 181.050 | 181.050 | 2.114 | .1539 |
| Outcome * Level * Spins * Subject | 39 | 3339.910 | 85.639 | | |
| Result * Level * Spins | 1 | 142.129 | 142.129 | 1.030 | .3164 |
| Result * Level * Spins * Subject | 39 | 5380.331 | 137.957 | | |
| Outcome * Result * Level * Spins | 1 | 23.409 | 23.409 | .109 | .7436 |
| Outcome * Result * Level * Spins * S... | 39 | 8411.301 | 215.674 | | |

Dependent: Quality

## Experiment four - means table for overall analysis

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Correct, Bias found, %56, Spin300 | 40 | 58.150 | 16.562 | 2.619 |
| Correct, Bias found, %56, Spin700 | 40 | 69.025 | 17.035 | 2.693 |
| Correct, Bias found, %54, Spin300 | 40 | 57.875 | 18.897 | 2.988 |
| Correct, Bias found, %54, Spin700 | 40 | 66.525 | 16.486 | 2.607 |
| Correct, No Bias found, %56, Spin300 | 40 | 54.290 | 18.113 | 2.864 |
| Correct, No Bias found, %56, Spin700 | 40 | 65.200 | 15.590 | 2.465 |
| Correct, No Bias found, %54, Spin300 | 40 | 53.975 | 21.497 | 3.399 |
| Correct, No Bias found, %54, Spin700 | 40 | 64.900 | 20.734 | 3.278 |
| Incorrect, Bias found, %56, Spin300 | 40 | 43.350 | 17.095 | 2.703 |
| Incorrect, Bias found, %56, Spin700 | 40 | 50.450 | 22.647 | 3.581 |
| Incorrect, Bias found, %54, Spin300 | 40 | 44.225 | 17.419 | 2.754 |
| Incorrect, Bias found, %54, Spin700 | 40 | 51.825 | 18.062 | 2.856 |
| Incorrect, No Bias found, %56, Spin300 | 40 | 49.925 | 19.818 | 3.133 |
| Incorrect, No Bias found, %56, Spin700 | 40 | 52.300 | 20.962 | 3.314 |
| Incorrect, No Bias found, %54, Spin300 | 40 | 44.100 | 19.956 | 3.155 |
| Incorrect, No Bias found, %54, Spin700 | 40 | 52.275 | 18.861 | 2.982 |

**Experiments three and four - anova table for comparative between subjects analysis of long-term outcome groups**

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Training | 1 | 310.616 | 310.616 | .187 | .6672 |
| Subject(Group) | 67 | 111563.781 | 1665.131 | | |
| Outcome | 1 | 46842.557 | 46842.557 | 42.063 | .0001 |
| Outcome * Training | 1 | 286.496 | 286.496 | .257 | .6137 |
| Outcome * Subject(Group) | 67 | 74612.758 | 1113.623 | | |
| Result | 1 | 142.721 | 142.721 | .282 | .5973 |
| Result * Training | 1 | 807.646 | 807.646 | 1.594 | .2111 |
| Result * Subject(Group) | 67 | 33945.507 | 506.649 | | |
| Level | 1 | 5977.828 | 5977.828 | 14.993 | .0002 |
| Level * Training | 1 | 3097.717 | 3097.717 | 7.770 | .0069 |
| Level * Subject(Group) | 67 | 26712.690 | 398.697 | | |
| Spins | 1 | 11916.540 | 11916.540 | 24.404 | .0001 |
| Spins * Training | 1 | 703.458 | 703.458 | 1.441 | .2343 |
| Spins * Subject(Group) | 67 | 32716.275 | 488.303 | | |
| Outcome * Result | 1 | 3932.611 | 3932.611 | 10.518 | .0018 |
| Outcome * Result * Training | 1 | 625.975 | 625.975 | 1.674 | .2001 |
| Outcome * Result * Subject(Group) | 67 | 25051.924 | 373.909 | | |
| Outcome * Level | 1 | 46.425 | 46.425 | .609 | .4380 |
| Outcome * Level * Training | 1 | .895 | .895 | .012 | .9140 |
| Outcome * Level * Subject(Group) | 67 | 5108.422 | 76.245 | | |
| Result * Level | 1 | 111.087 | 111.087 | .873 | .3534 |
| Result * Level * Training | 1 | 47.905 | 47.905 | .377 | .5415 |
| Result * Level * Subject(Group) | 67 | 8523.062 | 127.210 | | |
| Outcome * Spins | 1 | 849.357 | 849.357 | 4.888 | .0305 |
| Outcome * Spins * Training | 1 | 22.607 | 22.607 | .130 | .7195 |
| Outcome * Spins * Subject(Group) | 67 | 11641.550 | 173.754 | | |
| Result * Spins | 1 | 192.538 | 192.538 | 1.878 | .1751 |
| Result * Spins * Training | 1 | 324.243 | 324.243 | 3.163 | .0799 |
| Result * Spins * Subject(Group) | 67 | 6868.320 | 102.512 | | |
| Level * Spins | 1 | 48.258 | 48.258 | .499 | .4824 |
| Level * Spins * Training | 1 | 1.165 | 1.165 | .012 | .9129 |
| Level * Spins * Subject(Group) | 67 | 6478.670 | 96.697 | | |
| Outcome * Result * Level | 1 | 42.975 | 42.975 | .300 | .5857 |
| Outcome * Result * Level * Training | 1 | 35.038 | 35.038 | .245 | .6225 |
| Outcome * Result * Level * Subject(Group) | 67 | 9598.592 | 143.263 | | |
| Outcome * Result * Spins | 1 | 25.697 | 25.697 | .206 | .6512 |
| Outcome * Result * Spins * Training | 1 | 72.815 | 72.815 | .585 | .4472 |
| Outcome * Result * Spins * Subject(Group) | 67 | 8345.602 | 124.561 | | |
| Outcome * Level * Spins | 1 | 9.087 | 9.087 | .107 | .7450 |
| Outcome * Level * Spins * Training | 1 | 404.094 | 404.094 | 4.741 | .0330 |
| Outcome * Level * Spins * Subject(Group) | 67 | 5710.204 | 85.227 | | |
| Result * Level * Spins | 1 | 62.317 | 62.317 | .525 | .4710 |
| Result * Level * Spins * Training | 1 | 44.787 | 44.787 | .378 | .5409 |
| Result * Level * Spins * Subject(Group) | 67 | 7945.404 | 118.588 | | |
| Outcome * Result * Level * Spins | 1 | 172.350 | 172.350 | 1.240 | .2695 |
| Outcome * Result * Level * Spins * Training | 1 | 59.694 | 59.694 | .429 | .5146 |
| Outcome * Result * Level * Spins * Subject(Group) | 67 | 9316.173 | 139.047 | | |

## APPENDIX 4 - EXPERIMENT FIVE ANALYSES

**Experiment five - anova table for complete analysis of all subjects**

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Perspective | 1 | 8.939 | 8.939 | .003 | .9537 |
| Subject(Group) | 59 | 154812.783 | 2623.945 | | |
| Outcome | 1 | 46850.936 | 46850.936 | 36.767 | .0001 |
| Outcome * Perspective | 1 | 1297.852 | 1297.852 | 1.019 | .3170 |
| Outcome * Subject(Group) | 59 | 75181.051 | 1274.255 | | |
| Result | 1 | 15.513 | 15.513 | .031 | .8614 |
| Result * Perspective | 1 | 1.220 | 1.220 | .002 | .9609 |
| Result * Subject(Group) | 59 | 29773.545 | 504.636 | | |
| Level | 1 | 2619.195 | 2619.195 | 11.140 | .0015 |
| Level * Perspective | 1 | 210.464 | 210.464 | .895 | .3479 |
| Level * Subject(Group) | 59 | 13871.496 | 235.110 | | |
| Spins | 1 | 11508.872 | 11508.872 | 14.826 | .0003 |
| Spins * Perspective | 1 | 5639.702 | 5639.702 | 7.265 | .0091 |
| Spins * Subject(Group) | 59 | 45799.924 | 776.270 | | |
| Outcome * Result | 1 | 4432.511 | 4432.511 | 11.016 | .0016 |
| Outcome * Result * Perspective | 1 | 11649.603 | 11649.603 | 28.953 | .0001 |
| Outcome * Result * Subject(Group) | 59 | 23739.162 | 402.359 | | |
| Outcome * Level | 1 | 2.776 | 2.776 | .026 | .8722 |
| Outcome * Level * Perspective | 1 | 210.799 | 210.799 | 1.981 | .1645 |
| Outcome * Level * Subject(Group) | 59 | 6277.464 | 106.398 | | |
| Result * Level | 1 | 133.174 | 133.174 | .701 | .4057 |
| Result * Level * Perspective | 1 | 555.487 | 555.487 | 2.925 | .0925 |
| Result * Level * Subject(Group) | 59 | 11203.590 | 189.891 | | |
| Outcome * Spins | 1 | 228.625 | 228.625 | 1.083 | .3024 |
| Outcome * Spins * Perspective | 1 | 3.357 | 3.357 | .016 | .9001 |
| Outcome * Spins * Subject(Group) | 59 | 12459.380 | 211.176 | | |
| Result * Spins | 1 | 76.346 | 76.346 | .861 | .3573 |
| Result * Spins * Perspective | 1 | 213.950 | 213.950 | 2.413 | .1257 |
| Result * Spins * Subject(Group) | 59 | 5231.826 | 88.675 | | |
| Level * Spins | 1 | 5.243 | 5.243 | .056 | .8130 |
| Level * Spins * Perspective | 1 | 59.655 | 59.655 | .643 | .4260 |
| Level * Spins * Subject(Group) | 59 | 5477.270 | 92.835 | | |
| Outcome * Result * Level | 1 | 35.338 | 35.338 | .229 | .6343 |
| Outcome * Result * Level * Perspec... | 1 | .217 | .217 | .001 | .9703 |
| Outcome * Result * Level * Subject... | 59 | 9117.418 | 154.533 | | |
| Outcome * Result * Spins | 1 | 26.652 | 26.652 | .184 | .6699 |
| Outcome * Result * Spins * Perspec... | 1 | 30.003 | 30.003 | .207 | .6511 |
| Outcome * Result * Spins * Subject... | 59 | 8566.523 | 145.195 | | |
| Outcome * Level * Spins | 1 | 9.912 | 9.912 | .064 | .8004 |
| Outcome * Level * Spins * Perspect... | 1 | 90.422 | 90.422 | .588 | .4462 |
| Outcome * Level * Spins * Subject(... | 59 | 9070.261 | 153.733 | | |
| Result * Level * Spins | 1 | 3.595 | 3.595 | .019 | .8914 |
| Result * Level * Spins * Perspective | 1 | 2.118 | 2.118 | .011 | .9165 |
| Result * Level * Spins * Subject(Gr... | 59 | 11275.391 | 191.108 | | |
| Outcome * Result * Level * Spins | 1 | 549.812 | 549.812 | 1.230 | .2720 |
| Outcome * Result * Level * Spins *... | 1 | 841.703 | 841.703 | 1.883 | .1752 |
| Outcome * Result * Level * Spins *... | 59 | 26379.441 | 447.109 | | |

## Experiment five - means table for complete analysis of all subjects

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Correct, Bias found, %56, Spin300, Manufacturer | 30 | 49.233 | 17.754 | 3.241 |
| Correct, Bias found, %56, Spin300, Fraudster | 31 | 64.581 | 23.609 | 4.240 |
| Correct, Bias found, %56, Spin700, Manufacturer | 30 | 61.800 | 19.072 | 3.482 |
| Correct, Bias found, %56, Spin700, Fraudster | 31 | 62.452 | 27.281 | 4.900 |
| Correct, Bias found, %54, Spin300, Manufacturer | 30 | 44.333 | 19.987 | 3.649 |
| Correct, Bias found, %54, Spin300, Fraudster | 31 | 58.968 | 27.315 | 4.906 |
| Correct, Bias found, %54, Spin700, Manufacturer | 30 | 56.683 | 18.685 | 3.411 |
| Correct, Bias found, %54, Spin700, Fraudster | 31 | 63.968 | 24.153 | 4.338 |
| Correct, No Bias, %56, Spin300, Manufacturer | 30 | 50.633 | 19.593 | 3.577 |
| Correct, No Bias, %56, Spin300, Fraudster | 31 | 49.903 | 25.350 | 4.553 |
| Correct, No Bias, %56, Spin700, Manufacturer | 30 | 62.667 | 21.619 | 3.947 |
| Correct, No Bias, %56, Spin700, Fraudster | 31 | 57.387 | 25.474 | 4.575 |
| Correct, No Bias, %54, Spin300, Manufacturer | 30 | 48.267 | 19.943 | 3.641 |
| Correct, No Bias, %54, Spin300, Fraudster | 31 | 48.710 | 26.041 | 4.677 |
| Correct, No Bias, %54, Spin700, Manufacturer | 30 | 62.367 | 19.188 | 3.503 |
| Correct, No Bias, %54, Spin700, Fraudster | 31 | 50.000 | 26.294 | 4.723 |
| Incorrect, Bias found, %56, Spin300, Manufacturer | 30 | 39.867 | 19.917 | 3.636 |
| Incorrect, Bias found, %56, Spin300, Fraudster | 31 | 36.548 | 18.308 | 3.288 |
| Incorrect, Bias found, %56, Spin700, Manufacturer | 30 | 52.533 | 21.061 | 3.845 |
| Incorrect, Bias found, %56, Spin700, Fraudster | 31 | 38.581 | 22.503 | 4.042 |
| Incorrect, Bias found, %54, Spin300, Manufacturer | 30 | 36.700 | 21.549 | 3.934 |
| Incorrect, Bias found, %54, Spin300, Fraudster | 31 | 34.065 | 18.204 | 3.270 |
| Incorrect, Bias found, %54, Spin700, Manufacturer | 30 | 47.333 | 22.655 | 4.136 |
| Incorrect, Bias found, %54, Spin700, Fraudster | 31 | 31.419 | 17.536 | 3.150 |
| Incorrect, No Bias, %56, Spin300, Manufacturer | 30 | 35.800 | 15.307 | 2.795 |
| Incorrect, No Bias, %56, Spin300, Fraudster | 31 | 50.194 | 21.668 | 3.892 |
| Incorrect, No Bias, %56, Spin700, Manufacturer | 30 | 46.633 | 21.306 | 3.890 |
| Incorrect, No Bias, %56, Spin700, Fraudster | 31 | 48.484 | 23.176 | 4.163 |
| Incorrect, No Bias, %54, Spin300, Manufacturer | 30 | 38.233 | 20.216 | 3.691 |
| Incorrect, No Bias, %54, Spin300, Fraudster | 31 | 40.097 | 18.752 | 3.368 |
| Incorrect, No Bias, %54, Spin700, Manufacturer | 30 | 46.467 | 20.375 | 3.720 |
| Incorrect, No Bias, %54, Spin700, Fraudster | 31 | 47.258 | 25.457 | 4.572 |

## Experiment five - anova table for analysis of subjects given the fraudster perspective

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 30 | 65288.335 | 2176.278 | | |
| Outcome | 1 | 32403.389 | 32403.389 | 19.799 | .0001 |
| Outcome * Subject | 30 | 49099.673 | 1636.656 | | |
| Result | 1 | 4.083 | 4.083 | .006 | .9389 |
| Result * Subject | 30 | 20509.730 | 683.658 | | |
| Level | 1 | 2193.244 | 2193.244 | 6.565 | .0157 |
| Level * Subject | 30 | 10023.069 | 334.102 | | |
| Spins | 1 | 526.454 | 526.454 | .484 | .4918 |
| Spins * Subject | 30 | 32609.609 | 1086.987 | | |
| Outcome * Result | 1 | 15480.728 | 15480.728 | 24.941 | .0001 |
| Outcome * Result * Subject | 30 | 18621.085 | 620.703 | | |
| Outcome * Level | 1 | 133.163 | 133.163 | 1.072 | .3088 |
| Outcome * Level * Subject | 30 | 3727.149 | 124.238 | | |
| Result * Level | 1 | 73.550 | 73.550 | .296 | .5902 |
| Result * Level * Subject | 30 | 7444.012 | 248.134 | | |
| Outcome * Spins | 1 | 89.760 | 89.760 | .330 | .5701 |
| Outcome * Spins * Subject | 30 | 8168.802 | 272.293 | | |
| Result * Spins | 1 | 277.502 | 277.502 | 2.850 | .1018 |
| Result * Spins * Subject | 30 | 2921.310 | 97.377 | | |
| Level * Spins | 1 | 50.970 | 50.970 | .452 | .5067 |
| Level * Spins * Subject | 30 | 3384.843 | 112.828 | | |
| Outcome * Result * Level | 1 | 15.260 | 15.260 | .142 | .7089 |
| Outcome * Result * Level * Subject | 30 | 3222.302 | 107.410 | | |
| Outcome * Result * Spins | 1 | .050 | .0504.851E-4 | | .9826 |
| Outcome * Result * Spins * Subject | 30 | 3117.262 | 103.909 | | |
| Outcome * Level * Spins | 1 | 20.567 | 20.567 | .125 | .7264 |
| Outcome * Level * Spins * Subject | 30 | 4945.746 | 164.858 | | |
| Result * Level * Spins | 1 | .099 | .0993.643E-4 | | .9849 |
| Result * Level * Spins * Subject | 30 | 8135.964 | 271.199 | | |
| Outcome * Result * Level * Spins | 1 | 1398.970 | 1398.970 | 1.895 | .1789 |
| Outcome * Result * Level * Spins * Subject | 30 | 22150.593 | 738.353 | | |

Dependent: Quality

## Experiment five - means table for analysis of subjects given the fraudster perspective

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Correct, Bias found, %56, Spin300 | 31 | 64.581 | 23.609 | 4.240 |
| Correct, Bias found, %56, Spin700 | 31 | 62.452 | 27.281 | 4.900 |
| Correct, Bias found, %54, Spin300 | 31 | 58.968 | 27.315 | 4.906 |
| Correct, Bias found, %54, Spin700 | 31 | 63.968 | 24.153 | 4.338 |
| Correct, No Bias, %56, Spin300 | 31 | 49.903 | 25.350 | 4.553 |
| Correct, No Bias, %56, Spin700 | 31 | 57.387 | 25.474 | 4.575 |
| Correct, No Bias, %54, Spin300 | 31 | 48.710 | 26.041 | 4.677 |
| Correct, No Bias, %54, Spin700 | 31 | 50.000 | 26.294 | 4.723 |
| Incorrect, Bias found, %56, Spin300 | 31 | 36.548 | 18.308 | 3.288 |
| Incorrect, Bias found, %56, Spin700 | 31 | 38.581 | 22.503 | 4.042 |
| Incorrect, Bias found, %54, Spin300 | 31 | 34.065 | 18.204 | 3.270 |
| Incorrect, Bias found, %54, Spin700 | 31 | 31.419 | 17.536 | 3.150 |
| Incorrect, No Bias, %56, Spin300 | 31 | 50.194 | 21.668 | 3.892 |
| Incorrect, No Bias, %56, Spin700 | 31 | 48.484 | 23.176 | 4.163 |
| Incorrect, No Bias, %54, Spin300 | 31 | 40.097 | 18.752 | 3.368 |
| Incorrect, No Bias, %54, Spin700 | 31 | 47.258 | 25.457 | 4.572 |

## Experiment five - anova table for analysis of subjects given the manufacturer perspective

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 29 | 89524.448 | 3087.050 | | |
| Outcome | 1 | 16014.076 | 16014.076 | 17.806 | .0002 |
| Outcome * Subject | 29 | 26081.378 | 899.358 | | |
| Result | 1 | 12.513 | 12.513 | .039 | .8445 |
| Result * Subject | 29 | 9263.815 | 319.442 | | |
| Level | 1 | 661.526 | 661.526 | 4.985 | .0335 |
| Level * Subject | 29 | 3848.428 | 132.704 | | |
| Spins | 1 | 16362.513 | 16362.513 | 35.974 | .0001 |
| Spins * Subject | 29 | 13190.315 | 454.838 | | |
| Outcome * Result | 1 | 841.376 | 841.376 | 4.767 | .0372 |
| Outcome * Result * Subject | 29 | 5118.078 | 176.485 | | |
| Outcome * Level | 1 | 81.263 | 81.263 | .924 | .3444 |
| Outcome * Level * Subject | 29 | 2550.315 | 87.942 | | |
| Result * Level | 1 | 606.376 | 606.376 | 4.677 | .0389 |
| Result * Level * Subject | 29 | 3759.578 | 129.641 | | |
| Outcome * Spins | 1 | 141.376 | 141.376 | .956 | .3364 |
| Outcome * Spins * Subject | 29 | 4290.578 | 147.951 | | |
| Result * Spins | 1 | 17.063 | 17.063 | .214 | .6470 |
| Result * Spins * Subject | 29 | 2310.515 | 79.673 | | |
| Level * Spins | 1 | 14.526 | 14.526 | .201 | .6570 |
| Level * Spins * Subject | 29 | 2092.428 | 72.153 | | |
| Outcome * Result * Level | 1 | 20.213 | 20.213 | .099 | .7548 |
| Outcome * Result * Level * Subject | 29 | 5895.115 | 203.280 | | |
| Outcome * Result * Spins | 1 | 55.692 | 55.692 | .296 | .5903 |
| Outcome * Result * Spins * Subject | 29 | 5449.261 | 187.906 | | |
| Outcome * Level * Spins | 1 | 78.813 | 78.813 | .554 | .4626 |
| Outcome * Level * Spins * Subject | 29 | 4124.515 | 142.225 | | |
| Result * Level * Spins | 1 | 5.526 | 5.526 | .051 | .8228 |
| Result * Level * Spins * Subject | 29 | 3139.428 | 108.256 | | |
| Outcome * Result * Level * Spins | 1 | 15.230 | 15.230 | .104 | .7489 |
| Outcome * Result * Level * Spins * Subject | 29 | 4228.848 | 145.822 | | |

Dependent: Quality

## Experiment five - means table for analysis of subjects given the manufacturer perspective

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Correct, Bias found, %56, Spin300 | 30 | 49.233 | 17.754 | 3.241 |
| Correct, Bias found, %56, Spin700 | 30 | 61.800 | 19.072 | 3.482 |
| Correct, Bias found, %54, Spin300 | 30 | 44.333 | 19.987 | 3.649 |
| Correct, Bias found, %54, Spin700 | 30 | 56.683 | 18.685 | 3.411 |
| Correct, No Bias, %56, Spin300 | 30 | 50.633 | 19.593 | 3.577 |
| Correct, No Bias, %56, Spin700 | 30 | 62.667 | 21.619 | 3.947 |
| Correct, No Bias, %54, Spin300 | 30 | 48.267 | 19.943 | 3.641 |
| Correct, No Bias, %54, Spin700 | 30 | 62.367 | 19.188 | 3.503 |
| Incorrect, Bias found, %56, Spin300 | 30 | 39.867 | 19.917 | 3.636 |
| Incorrect, Bias found, %56, Spin700 | 30 | 52.533 | 21.061 | 3.845 |
| Incorrect, Bias found, %54, Spin300 | 30 | 36.700 | 21.549 | 3.934 |
| Incorrect, Bias found, %54, Spin700 | 30 | 47.333 | 22.655 | 4.136 |
| Incorrect, No Bias, %56, Spin300 | 30 | 35.800 | 15.307 | 2.795 |
| Incorrect, No Bias, %56, Spin700 | 30 | 46.633 | 21.306 | 3.890 |
| Incorrect, No Bias, %54, Spin300 | 30 | 38.233 | 20.216 | 3.691 |
| Incorrect, No Bias, %54, Spin700 | 30 | 46.467 | 20.375 | 3.720 |

## Experiment five - Planned contrasts for the effects of outcome for both perspectives

### Comparison of outcome interaction means - Fraudster perspective

**Comparison 1**
**Effect: Outcome * Result**
**Dependent: Quality**

| | Cell Weight |
|---|---|
| Correct, Bias found | .900 |
| Correct, No Bias | .100 |
| Incorrect, Bias found | -.700 |
| Incorrect, No Bias | -.300 |

| | |
|---|---|
| df | 1 |
| Sum of Squares | 46176.611 |
| Mean Square | 46176.611 |
| F-Value | 74.394 |
| P-Value | .0001 |
| G-G | .0001 |
| H-F | .0001 |

### Comparison of outcome interaction means - Manufacturer perspective

**Comparison 1**
**Effect: Outcome * Result**
**Dependent: Quality**

| | Cell Weight |
|---|---|
| Correct, Bias found | .300 |
| Correct, No Bias | .700 |
| Incorrect, Bias found | -.100 |
| Incorrect, No Bias | -.900 |

| | |
|---|---|
| df | 1 |
| Sum of Squares | 14656.141 |
| Mean Square | 14656.141 |
| F-Value | 83.044 |
| P-Value | .0001 |
| G-G | .0001 |
| H-F | .0001 |

# APPENDIX 5
# EXPERIMENT SIX ANALYSIS

## Experiment six - Complete anova table

**Type III Sums of Squares**

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value | G-G | H-F |
|---|---|---|---|---|---|---|---|
| Perspective | 1 | 25.352 | 25.352 | 4.552 | .0402 | | |
| Subject(Group) | 34 | 189.352 | 5.569 | | | | |
| Rating | 5 | 475.537 | 95.107 | 25.133 | .0001 | .0001 | .0001 |
| Rating * Perspective | 5 | 74.148 | 14.830 | 3.919 | .0022 | .0061 | .0036 |
| Rating * Subject(Group) | 170 | 643.315 | 3.784 | | | | |

Dependent: Ratings

**Table of Epsilon Factors for df Adjustment**
**Dependent: Ratings**

| | G-G Epsilon | H-F Epsilon |
|---|---|---|
| Rating | .743 | .870 |

## Experiment six - Overall means table

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Subjects, You review | 18 | 7.333 | 1.414 | .333 |
| Subjects, They review | 18 | 6.889 | 1.023 | .241 |
| Level, You review | 18 | 5.833 | 2.036 | .480 |
| Level, They review | 18 | 7.167 | 1.295 | .305 |
| Result, You review | 18 | 6.167 | 2.176 | .513 |
| Result, They review | 18 | 4.722 | 2.803 | .661 |
| Effect, You review | 18 | 5.944 | 1.211 | .286 |
| Effect, They review | 18 | 5.722 | 2.218 | .523 |
| Outcome, You review | 18 | 5.889 | 1.937 | .457 |
| Outcome, They review | 18 | 5.056 | 2.733 | .644 |
| Money, You review | 18 | 3.667 | 2.849 | .672 |
| Money, They review | 18 | 1.167 | 1.339 | .316 |

## Experiment six - Follow up analysis for ratings

| | Vs. | Diff. | Std. Error | t-Test | P-Value |
|---|---|---|---|---|---|
| Subjects | Level | .611 | .459 | 1.333 | .1844 |
| | Result | 1.667 | .459 | 3.635 | .0004 |
| | Effect | 1.278 | .459 | 2.787 | .0059 |
| | Outcome | 1.639 | .459 | 3.574 | .0005 |
| | Money | 4.694 | .459 | 10.238 | .0001 |
| Level | Result | 1.056 | .459 | 2.302 | .0225 |
| | Effect | .667 | .459 | 1.454 | .1478 |
| | Outcome | 1.028 | .459 | 2.242 | .0263 |
| | Money | 4.083 | .459 | 8.906 | .0001 |
| Result | Effect | -.389 | .459 | -.848 | .3975 |
| | Outcome | -.028 | .459 | -.061 | .9518 |
| | Money | 3.028 | .459 | 6.603 | .0001 |
| Effect | Outcome | .361 | .459 | .788 | .4320 |
| | Money | 3.417 | .459 | 7.452 | .0001 |
| Outcome | Money | 3.056 | .459 | 6.664 | .0001 |

# APPENDIX 6
# EXPERIMENT SEVEN ANALYSES

**Experiment seven - anova table and means tables for analysis of initial quality judgements**

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Level | 1 | 2.6 | 2.6 | 0.02 | 0.898 |
| Snos | 3 | 1014.2 | 338.1 | 2.16 | 0.094 |
| Result | 1 | 1515.0 | 1515.0 | 9.66 | 0.002 |
| Level*Snos | 3 | 1120.7 | 373.6 | 2.38 | 0.070 |
| Level*Result | 1 | 48.6 | 48.6 | 0.31 | 0.578 |
| Snos*Result | 3 | 1192.3 | 397.4 | 2.53 | 0.058 |
| Level*Snos*Result | 3 | 86.3 | 28.8 | 0.18 | 0.908 |
| Error | 224 | 35134.3 | 156.8 | | |
| Total | 239 | 40113.9 | | | |

MEANS

| Level | N | Quality |
|---|---|---|
| 1 | 120 | 50.122 |
| 2 | 120 | 51.742 |

| Snos | N | Quality |
|---|---|---|
| 1 | 60 | 48.116 |
| 2 | 60 | 55.721 |
| 3 | 60 | 55.500 |
| 4 | 60 | 46.784 |

| Result | N | Quality |
|---|---|---|
| 1 | 120 | 56.638 |
| 2 | 120 | 46.543 |

## Experiment seven - anova table for analysis of quality judgements

**Type III Sums of Squares**

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 59 | 1910.536 | 32.382 | | |
| Outcome | 1 | 233.051 | 233.051 | 7.554 | .0079 |
| Outcome * Subject | 59 | 1820.136 | 30.850 | | |
| Result | 1 | .051 | .051 | .001 | .9748 |
| Result * Subject | 59 | 2984.886 | 50.591 | | |
| Outcome * Result | 1 | 18.984 | 18.984 | .548 | .4619 |
| Outcome * Result * Subject | 59 | 2042.703 | 34.622 | | |

Dependent: Memory for quality

## Experiment seven - means table for analysis of quality judgements

**Means Table**
**Effect: Outcome**
**Dependent: Memory for quality**

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Replicated | 120 | 1.350 | 5.127 | .468 |
| Not Replicated | 120 | -.621 | 6.890 | .629 |

## Experiment seven - anova table for analysis of memory for subject numbers

**Type III Sums of Squares**

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 59 | 19975.733 | 338.572 | | |
| Outcome | 1 | 858.817 | 858.817 | 4.209 | .0447 |
| Outcome * Subject | 59 | 12038.183 | 204.037 | | |
| Result | 1 | 135.000 | 135.000 | .818 | .3694 |
| Result * Subject | 59 | 9736.000 | 165.017 | | |
| Outcome * Result | 1 | 114.817 | 114.817 | .416 | .5215 |
| Outcome * Result * Subject | 59 | 16290.183 | 276.105 | | |

Dependent: Memory for subjects

## Experiment seven - means table for analysis of memory for subject numbers

**Means Table**
**Effect: Outcome**
**Dependent: Memory for subjects**

| | Count | Mean | Std. Dev. | Std. Error |
|---|---|---|---|---|
| Replicated | 120 | 4.508 | 16.202 | 1.479 |
| Not Replicated | 120 | .725 | 15.077 | 1.376 |

## Experiment seven - anova table for analysis of memory for significance levels

**Type III Sums of Squares**

| Source | df | Sum of Squares | Mean Square | F-Value | P-Value |
|---|---|---|---|---|---|
| Subject | 59 | 26.412 | .448 | | |
| Outcome | 1 | .267 | .267 | 1.000 | .3214 |
| Outcome * Subject | 59 | 15.733 | .267 | | |
| Result | 1 | .150 | .150 | .330 | .5681 |
| Result * Subject | 59 | 26.850 | .455 | | |
| Outcome * Result | 1 | .204 | .204 | .221 | .6401 |
| Outcome * Result * Subject | 59 | 54.546 | .925 | | |

Dependent: Memory for sig. level

## APPENDIX 7

## CALCULATION OF EFFECT SIZES

The use of a product moment as a standard measure of effect size is recommended by Cohen (1977), and was used by Christensen-Szalanski and Willham (1991) in their meta-analysis of hindsight studies. The same measure of effect size was used throughout this study to enable direct comparison. This method of determining effect size is also described by Lipsey (1990).

In within subject designs, where there is a dichotomous independent variable (e.g. long-term outcome) and a graduated dependent variable (e.g. quality), point biserial correlation (r) can be used to gain a measure of the proportion of variance in the dependent variable which can be directly attributed to the effect of the independent variable. This method involves assigning numeric values to the two levels of the independent variable (usually 0 and 1) in line with the experimental hypothesis such that the higher value is given to the level which is expected to be associated with higher values of the dependent variable. These arbitary values are then correlated with the resulting values of the dependent variable producing a coefficient r. This method effectively regresses the levels of the independent variable onto the values of the dependent variable. As in linear regression the square of this correlation coefficient ($r^2$) reflects the proportion of variance in a given measure which can be directly attributed to the independent variable in question.

It should be noted that the actual numeric values assigned to the levels of the independent variable have no effect on the resulting r value where this variable is dichotomous. For example values of 7 and 139 will produce the same result as values of 0 and 1. If these values have been assigned in line with the predictions of the experimental hypothesis an effect in the opposite direction to the experimental hypothesis will result in a negative r value.

It should also be noted that in the case of between subject designs this method of calculation of effect sizes must be adjusted to allow for unequal subject group sizes. This adjustment was not required in the present research, however the appropriate formula may be found in Lipsey (1990) (Page 84)