OVERCONFIDENCE IN HUMAN JUDGEMENT

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Briony D. Pulford B.Sc. (Leicester)

Department of Psychology

University of Leicester

1996

To Mum, Dad and Rowan

He who knows and knows that he knows,
He is wise, follow him.

He who knows and knows not that he knows,
He is asleep, awaken him.

He who knows not and knows not that he knows not,
He is a fool, shun him.

He who knows not and knows that he knows not,
He is a child, teach him.

Arabian proverb

Acknowledgements

I should like to thank my parents, Brenda and Stuart Pulford, for working so hard to provide for my education and for their unconditional love, support and encouragement. Thanks also to my friends, Alison, James, Joanna, John, May, Rochelle, and especially Fatima, who were always there when I needed them. To Andrew Colman, my friend and mentor, who deserves a medal for patience throughout the supervision of this thesis, thank you.

Thanks are also due to the British Psychological Society for employing me in their office part-time for the first two years of my PhD, which was in itself a valuable experience and without which this research could not have been begun. The rest of the thesis was completed with the financial assistance of an Economic and Social Research Council studentship, which gave me the freedom to pursue and enjoy full-time research.

Finally, I would like to thank Brian Parkinson and Peter Harris for examining this thesis, and for their many helpful contributions.

OVERCONFIDENCE IN HUMAN JUDGEMENT

Briony D Pulford

This thesis examines the relationship (termed overconfidence) between subjective confidence in judgements and objective reality, to determine and clarify the nature of influencing factors. Chapter 1 examines whether predictions of future behaviour for oneself are less overconfident than those for a partner. Chapter 2 studies the influence of base rates and the desirability of the outcome on overconfidence in behavioural predictions, and whether people can accurately estimate population base rates. Chapters 3 and 4 then progress to ask whether individual characteristics (such as field dependence, probabilistic reasoning ability, and self-esteem) of the person making the judgements are related to overconfidence in general knowledge tasks and behavioural predictions. Chapter 5 investigates whether feedback can affect overconfidence. A questionnaire based methodology was used. The results indicated that overconfidence is strongly related to task difficulty, which varies according to the person being predicted for and the base rate. It was found that predictions about another person were more overconfident than predictions for oneself. It was also shown that maximum overconfidence occurred at mid base rates with underconfidence at extreme base rates. Higher overconfidence in predictions was found for events which were positive in outcome than those with negative outcomes. Large variations in individuals' overall overconfidence occurred, but no evidence was found to link this with the individual differences studied. Finally, feedback did not reduce overconfidence on average, but did reduce overconfidence over time for consistently hard questions, but not for easy questions. Issues raised by this thesis are examined in the Discussion.

CONTENTS

CHAPTER 5 - Feedback and item difficulty effects on overconfidence

CHAPTER 6 - Discussion

Table of Contents: Figures

Table of Contents: Tables

Table of Contents: Appendices

APPENDIX K

Experiment 4: Variable Labels

Experiment 4: Data

APPENDIX L

Experiment 4: Data Analysis

APPENDIX M

Experiment 5: Self-Esteem Questionnaire

APPENDIX N

Experiment 5: Data

Experiment 5: Variable Labels

APPENDIX O

Experiment 5: Data Analysis

APPENDIX P

Experiment 6: Pilot Study Questionnaire

Experiment 6: Main Study Questions and Answers

APPENDIX Q

Experiment 6: Pilot Question Means

Experiment 6: Main Study Means

Experiment 6: Variable Labels

Experiment 6: Means for Questions

APPENDIX R

Experiment 6: Data Analysis

{{ NOTE: Materials can be obtained from the author by email request }}

Publications Arising From This Thesis

Pulford, B. D. (1993). Base rates and the overconfidence effect. *Proceedings of the British Psychological Society, 1*, 53. [Abstract].

Pulford, B. D. (1994). Overconfidence: Some of the people, some of the time. *The Psychology Postgraduate Affairs Group Newsletter, 12*, 35-39.

Pulford, B. D. (1995). The effects of feedback and gender on overconfidence. *Proceedings of the British Psychological Society, 3,* 28 . [Abstract].

Pulford, B. D., & Colman, A. M. (1996). Overconfidence, base rates, and outcome positivity/negativity of predicted events. *British Journal of Psychology, 87,* 431-445.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences, 23*(1), 125-133.

CHAPTER 0
Overconfidence in Human Judgement

Confidence

This thesis aims to ask and to answer some questions about people's confidence in their judgements. Confidence is a commonly used term and people seem to have an intuitive understanding of what it is, but its meaning is rarely put into words or explained. Confident people are generally admired by our society and it is seen as disadvantageous to lack confidence, which is usually viewed very positively. But what is confidence? A dictionary definition of confidence states that it is "**1** firm trust (*have confidence in his ability*). **2 a** a feeling of reliance or certainty. **b** a sense of self-reliance; boldness. **3 a** some-thing told confidentially. **b** the telling of private matters with mutual trust." (Hawkins & Allen, Eds, *The Oxford Encyclopedic English Dictionary*, 1991, p. 306). Of these definitions number 2a seems to get to the heart of the matter, as a feeling of certainty can explain definition 1, because trust implies a high degree of certainty in something or someone. Definition 2a also covers definition 2b as having certainty about one's abilities and attributes. Thus it seems that confidence is essentially a feeling of certainty, the strength of this feeling being the level of confidence.

In everyday life when we talk of confidence we are referring to a degree of certainty that we hold in the validity, or correctness, of our mental states: beliefs, knowledge, perceptions, predictions, judgements, or decisions. When-ever someone makes a prediction or gives an answer, they have an associated level of confidence in that answer/decision, because "a concomitant feeling of confidence is typical of almost any type of mental activity" (Zakay & Tsal, 1993, p. 53). The amount of confidence varies, as it is a continuous variable, ranging from low (a complete absence of certainty) to high (absolute certainty). For example, when we make a decision we have an associated level of confidence in it, that is, a belief that the decision is the right, or the best one. Throughout this thesis the term *judgement* is mainly used because it implies making distinctions and estimations, rather than the term *decision* which implies making a choice between several alternatives based on previous judgements. People also have confidence in their decisions, and when a choice between alternatives has been made then this term may be used instead of the term judgement.

A feeling of certainty can remain within a person or be translated from a feeling to a more concrete form and communicated to other people via verbal expressions such as "I'm *certain* that Rome is the capital of Italy" or "I *think* that Rome is *probably* the capital of Italy". More numerical forms of language can also be used such as "I'm 99 percent sure that Paris is the capital of France". The latter form of statement may seem to be a more accurate way of conveying information to the listener about the strength of belief or likely outcomes, as people may have different interpretations of words such as "certain" or "probably". Alternatively, numerical statements are sometimes not always appropriate and can also lead to misuse or misunderstandings.

In common use is the term "self-confidence" which relates to beliefs about ourselves, for example, how certain we are about our own abilities in different situations. Essentially self-confidence reflects our beliefs about our own abilities, self-worth and value, and as beliefs change so can self-confidence; it need not remain stable. We have beliefs about ourselves, our abilities, our resources, the situation, and other people, all of which we have different levels of certainty in and thus to some extent our confidence varies according to circumstances. Overall self-confidence results from the combination of the many beliefs we have about ourselves which are present at any one time. As such, self-confidence can vary from situation to situation because the level of certainty in our ability to handle different situations varies with experience. Sometimes situations arise where the confidence a person exhibits to other people may not be truthfully communicated, as people are influenced by factors which may motivate them to give a false impression of their level of certainty. Motivational factors can affect how we portray our level of confidence to other people, but do they also distort the true confidence within ourselves? Can we mislead ourselves, and if so, why?

Confidence, then, is a feeling of certainty about the state of reality, but this feeling is internal to the person and may not actually correspond with external reality. It may be the case that a person has high confidence in a fact which turns out not to be true, or high confidence in a prediction which then does not occur. This thesis will look at the relationship between subjective confidence (certainty) in judgements and reality, to see how well they correspond, and if they don't, then what factors influence any discrepancy. The term for the discrepancy between confidence and objective reality is called *calibration,* and overconfidence is a measure of calibration.

Confidence as a Subjective Probability

Kahneman and Tversky (1982) stated that "confidence is the subjective probability or degree of belief associated with what we 'think' will happen" (p. 515). Correspondingly, subjective probability is the subject's degree of belief in the correctness of a judgement or decision. If, as Gigerenzer, Hoffrage, and Kleinbölting (1991) proposed, a choice between alternatives and the confidence in that judgement are expressions of the same conditional probability, and are activated at the same time, then as the subjective probability of a specific outcome increases so will the associated confidence in it.

To a degree at least, subjective probability may be inferred from behaviour, such as gambling bets, although beliefs and behaviour may not always correspond, for example, people may know that smoking causes cancer and not approve of smoking, but still smoke. In psychological research, confidence has typically been measured by asking people for a rating of the likelihood or probability that an answer or prediction is correct. Confidence estimates have been used in research to indicate the degree of subjective probability in judgements.

It is when the subjective probability of a judgement does not match its objective probability that poor calibration (over- or underconfidence) is said to occur, as people over- or underestimate the validity of their beliefs. The causes of poor calibration seem to occur when a person cannot accurately assess the subjective probability of a judgement being true, so that it is in line with objectively measured probabilities. Motivational factors, people's beliefs, cognitive problems of recall and processing, and problems of quantifying and communicating confidence combine and leave ample room for biases to appear. Some judgements lead to very poor calibration as the objective probability is much lower or higher than the subject believes and expected, thus their subjective probability is discrepant from reality.

It is sometimes possible to evaluate whether a person's belief is accurate or not, but evaluating the validity of a level of confidence is more problematic. Much previous research has looked at how confidence is expressed in terms of probability distributions. A classical interpretation of probability is that it is a relative frequency, and as such it is inapplicable to single events. But in Bayesian statistics, probability is defined as a degree of belief, so is applicable to single events. Since a probability estimate is a measure of a person's subjective belief in the truth of a proposition it could be argued that it is not open to external validation because it is subjective. This personalistic view of confidence regards subjective probability as a statement of degree of belief, and as such this perspective argues that it does not have to equal the proportion of correct outcomes. DeFinetti (1962) took this stance, believing that a person's subjective conviction of the truth of a hypothesis is not capable of being proved or disproved because it is internal and subjective. However, if people's subjective beliefs do not tally with reality, then it can be argued that biases are present, and this thesis accepts this principle and adopts a Bayesian perspective. The discrepancy between subjective confidence and objective accuracy, or *calibration* of the two, has received much experimental study. A general bias has been found, as confidence typically exceeds accuracy, the so called *overconfidence effect*. With these theoretical debates in mind, a review of the research into calibration will follow.

Reasons To Be Well Calibrated

Adams and Adams (1961) stated that "realism of confidence" (as they called it before the term calibration came into use) may be more important in everyday life than actual performance, for example they stated that "it might be more important, in terms of his future work, social interactions, confidence in himself, etc., for a student to be able to discriminate realistically between what he knows and what he does not know than it would be for him to know considerably more than he does know without such discrimination" (p. 36). If one knows what one knows and what one doesn't know, then one can make better judgements than when one doesn't realise that one's facts are wrong.

It could be argued that confidence need not accurately equal the probability of the chosen answer being correct, but if confidence does not reflect degrees of certainty in an answer, then it is not a useful measure of the degree of certainty we have in that answer. If confidence is unrelated to accuracy then people will be confident when they are wrong and unconfident when they are right. It is necessary for confidence to reflect accuracy and be unbiased, or incorrect decisions will be made and if certainty cannot be assessed then unexpected errors will occur. If confidence and accuracy do not correspond to each other then a bias is said to be present.

A bias is a deviation from a correct or accurate standard, which is consistent and systematic, rather than the odd mistake. The defined standard is the rational or most suitable solution which people's responses can be compared with to measure accuracy. It can, however, be difficult to define a standard in some situations where the correct answer is subjective. Around the 1970's the consensus was that human judgement is full of errors, but during the 1980's several researchers challenged this view (Einhorn & Hogarth, 1985; Funder, 1987), and currently the debate is on the external conditions necessary for accuracy and the internal processes within people. Cognitive biases may have costs attached to them and thus it may be beneficial to help people avoid them, alternatively they may be beneficial and the inaccuracy produced may be less than the mental energy saved. Determining whether cognitive biases are a help or a hindrance is an aim of this field of research.

Having appropriate or well calibrated probability assessments is important. For example, doctors giving diagnoses need to accurately assess the probability of certain diseases, and the certainty (confidence) in those judgements is important because it conveys the degree of certainty in that judgement and transmits information to other people about probabilities. By speaking confidently when we are only partly sure of something, or speaking without confidence when we do have belief in what we say, we are influencing how other people will interpret the validity of the information we are communicating to them.

Some situations may occur where the subjective probability (degree of belief in the statement being true) is biased. Comparing subjective confidence and objective reality has revealed biases, and it is important to know what factors influence these biases, as poor judgements may result which may be avoidable. Situations may, however, arise where a person cannot admit lack of confidence publicly for fear of loss of face, even though they know that they are wrong. People can purposefully deceive others about their confidence, but this is not the area this thesis will concentrate upon. Instead the true failure of the person to be internally well calibrated will be studied, to find out what factors influence this discrepancy.

Defining Overconfidence

The *overconfidence effect* has been defined as occurring “when the confidence judgements are larger than the relative frequencies of the correct answers” (Gigerenzer, Hoffrage, & Kleinbölting, 1991, p. 506). The most consistent finding is that people are not well calibrated, they tend to have too strong a belief in the correctness of their judgements, that is, they are too confident. When people are overconfident they believe that they know more than they in fact do know, or believe their accuracy to be higher than it in fact is. Over-confidence was referred to as a “cognitive conceit” by Block and Harper (1991).

The terms perceived likelihood/personal probabilities and confidence are often used interchangeably as expressions of uncertainty. Koriat, Lichtenstein, and Fischhoff (1980) stated that “an individual is well calibrated if, over the long run, for all answers assigned a given probability, the proportion correct equals the probability assigned” (p. 108). In their often cited review of calibration of probabilities, Lichtenstein, Fischhoff, and Phillips (1982) stated that:

> if a person assesses the probability of a proposition being true as .7 and later finds that the proposition is false, that in itself does not invalidate the assessment. However, if a judge assigns .7 to 10,000 independent propositions, only 25 of which subsequently are found to be true, there is something wrong with these assessments. The attribute that they lack is called calibration; it has also been called realism (Brown & Shuford, 1973), external validity (Brown & Shuford, 1973), realism of confidence (Adams & Adams, 1961), appropriateness of confidence (Oskamp, 1962), secondary validity (Murphy & Winkler, 1971), and reliability (Murphy, 1973). Formally, a judge is calibrated if, over the long run, for all propositions assigned a given probability, the proportion that is true equals the probability assigned. (pp. 306-307).

There has been a lot of research on the overconfidence effect in decision making and judgements (see Lichtenstein, Fischhoff, & Phillips, 1982, for a review). People are generally unaware of the "shortcomings of human inference" (Nisbett & Ross, 1980; Ross, 1977) thus they rely on biased samples of data, over-rely on trait theories and use nonregressive prediction strategies. As people are generally unaware of these inferential errors they tend not to adjust their subjective confidence and thus are overconfident. Fischhoff (1982) reviewed the literature up to that date and concluded that overconfidence is a robust phenomenon and is not eliminated by warnings or rewards.

Peterson and Pitz (1988) drew a distinction between uncertainty and confidence, which are usually seen as interchangeable terms, but refer to different aspects of a judgement. They “use the term *confidence* to describe a person’s belief that his or her prediction is correct” and “the term *uncertainty* whenever the task requires the decision maker to consider the variability of possible values for the quantity” (p. 91) and is therefore the belief about possible values of a quantity. They stated that the lack of agreement between uncertainty estimates and reality, which they call *hyperprecision*, is a widespread bias whereas overconfidence is more dependent upon the conditions of the task.

Along the same lines, Liberman and Tversky (1993) subsequently distinguished between two types of over-confidence, *specific* and *generic*. Specific overconfidence is when a person “overestimates the probability of a specific hypothesis or a designated outcome” (p. 166), for example, I think that there is a 75% chance of rain tomorrow. Generic overconfidence is when “he or she overestimates the probability of the hypothesis that he or she considers most likely” (p. 166), such as, I am 95% sure that it will rain tomorrow. This is an important distinction between the likelihood of the judgemental answer and the likelihood that one’s answer is correct.

Measurement of Overconfidence

To measure overconfidence one must have a judgement where the answer is verifiable as correct or not and a measure of the degree of confidence in that judgement. Confidence in a judgement or decision can be measured either in descriptive terms or in more quantifiable ways such as converting the degree of strength of belief/certainty into an estimate of the subjective probability that the judgement is correct. Confidence is therefore often measured as an estimate of the subjective probability in the correctness of a judgement.

The subjective probability is a person's estimate of how probable a judgement is of being correct, and it can be argued that this is the same as a feeling of certainty which can be expressed in more everyday terminology as confidence. For example, saying "I'm 95% sure that X is the correct answer" indicates a degree of belief paralleled by saying " I am almost totally sure that X is the correct answer". Both of these statements leave a small margin of room for error. On the other hand, a statement of certainty such as "I am 100% sure X is correct" or "X is certainly the correct answer" leaves no doubt that the person believes that the answer is correct.

Two types of measurement scales are typically used to measure confidence, the full- and half-range scales. When the judgement task requires a subject to generate their own answer and then evaluate their confidence that the answer is true, or when a large selection of answers to a question is possible, then a full-range scale, 0-100%, can be used. When a subject has a choice between two mutually exclusive answers, such as a yes/no option, then a half-range scale of 50-100% is appropriate, because, for example, 55% confidence in answer X would equal 45% confidence in Y. This principle can be generalised as follows. If K is the number of answers/options available to the subject for a decision then the measurement scale of confidence should extend from $100/K$ to 100% (Adams & Adams, 1961). The full-range scale can, however, also be used in limited option decisions if subjects are rating their confidence that the answer is true.

In many previous experiments subjects have been asked to make a decision about which of several different given answers is correct. Under such circumstances, however, subjects can rely upon recognition of the answer rather than direct recall from memory, and it also allows for the deduction of the correct answer if other answers cannot logically be correct. It would, however, seem preferable to measure confidence in the correctness (validity) of answers generated by the person rather than the experimenter because this is more true to the processes that occur in real life, where answers are rarely pre-provided, and this is the methodology which will be adopted in this thesis.

Before Adams and Adams (1961) it was generally believed that there was no way of telling whether the level of confidence was appropriate to the level of accuracy, except at the lowest point of the scale, where accuracy should be at chance levels, and at full confidence where total accuracy should be achieved. Adams and Adams instructed subjects that "of all those decisions made with confidence *p*, about *p*% should be correct" (p. 37). This technique reduces subjects' uncertainty about what the scale represents and confidence can be compared with accuracy across the whole scale. This method provided the conceptual basis for research into the calibration of subjective probability (confidence) and accuracy since that date. Adams and Adams did, however, point out three limitations with this method: a measure of accuracy must be available, one confidence judgement on its own cannot be assessed as realistic or not unless it is 0 or 100% paired with an incorrect or correct decision, and large numbers of judgements are needed for precise measures of calibration.

The continuous variable of subjective confidence can be divided into discrete ranges and the objective proportion of answers which are correct (accuracy) can be plotted on a graph against the mean confidence for each range; this is called a *calibration curve* (see Figure 0.1). Figure 0.1 shows a hypothetical calibration curve where a half-range scale (.5 to 1.0) has been used, with only two answers available to the subject. On this graph an *identity line* has been drawn, where confidence exactly equals accuracy and perfect calibration and no over- or underconfidence exists. When the proportion correct is less than the subjective probability (confidence) *overconfidence* is said to occur, and when it is higher there is *underconfidence*.

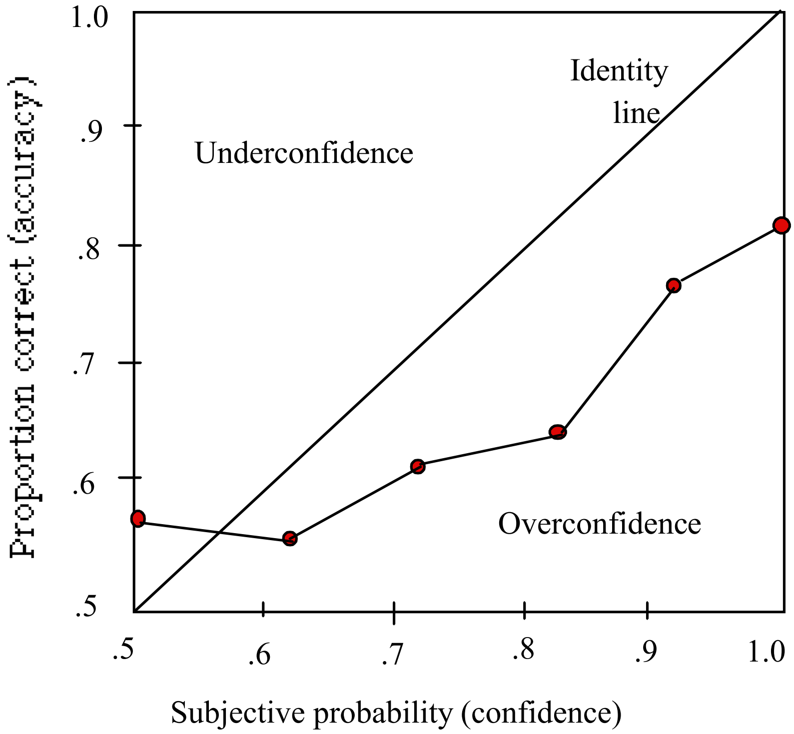

Figure 0.1. Example of a calibration curve with an identity line.

Using the concept that when mean confidence equals, for example, 70%, the relative accuracy should be about 70%, statistical measurements of the discrepancy between confidence and accuracy can be computed. One such measure of this discrepancy is the *calibration score*, which is computed by the division of the confidence variable into discrete categories (ranges), such as .60 to .69, and then the discrepancy between confidence and the proportion of correct answers is calculated for all of the categories. "The calibration score is the weighted mean of the squared differences between the data points in a Figure and the identity line" (Koriat, Lichtenstein, & Fischhoff, 1980, p. 110). A perfectly calibrated person would score zero on this measure. Thus the standard measure of calibration (C) is defined as the variance of the observed proportions correct c_t around the corresponding levels of confidence r_t, where n_t is the number of responses in confidence category t , T is the total number of response categories used, and N is the total number of responses (see Murphy, 1973):

$$\mathrm{C} = \frac{1}{N} \sum_{t=1}^{T} n_t (r_t - c_t)^2 .$$

Overconfidence, however, is a slightly different measure of the confidence-accuracy discrepancy, which retains the positive or negative sign to indicate whether the bias is over- or underconfident. Overconfidence is calculated using the following formula:

$$\text{Over / underconfidence} = \frac{1}{N}\sum_{t=1}^{T} n_t\left(r_t - c_t\right)$$

where T = total number of response categories used, n_t = the number of times the response r_t was used and c_t is the proportion correct for all items assigned probability r_t, which can, however, be simplified as it equals the difference between the mean confidence score x and the mean proportion correct c (accuracy).

$$\text{Over / underconfidence} = x - c$$

Other statistical measures of calibration include Brier scores, resolution scores, and eta squared (see Yates, 1982; Sharp, Cutler, & Penrod, 1988, for detailed descriptions of these calculations). Resolution is a measure of how well an individual can tell correct from incorrect answers via their confidence score, regardless of their absolute level of confidence. Resolution reflects the slope of a calibration curve on a figure in relation to the ideal of the identity line.

Factors Influencing Overconfidence

Overconfidence has been repeatedly shown to occur in many different samples of the population, for example students (Fischhoff, Slovic, & Lichtenstein, 1977; Koriat et al., 1980), members of the armed forces (Hazard & Peterson, 1973), CIA analysts (Cambridge & Shreckengost, 1978), clinical psychologists (Oskamp, 1962), bankers (Staël von Holstein, 1972), executives (Moore, 1977), negotiators (Neale & Bazerman, 1990), lawyers (Wagenaar & Keren, 1986), and civil engineers (Hynes & Vanmarcke, 1976). The predictive validity of confidence in resulting accuracy is also weak for most types of judgements studied (Glenberg & Epstein, 1985). Overconfidence is a pervasive finding, however, in some situations such as when the questions are very easy *under*confidence is found. Four factors were identified by Keasey and Watson (1989) as influences on the accuracy-confidence relationship: the complexity of the task, amount of feedback given, motivation level of the subjects, and skill of the subjects. Thus the causes and situations which determine the appropriateness of confidence need expanding here.

Task Difficulty and the Hard-Easy Effect

Many studies have shown that people are especially overconfident in their general-knowledge judgements that are moderate or hard in difficulty (Arkes, Christensen, Lai, & Blumer, 1987; Fischhoff, Slovic, & Lichtenstein, 1977; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977, 1980; Phillips & Wright, 1977; Ronis & Yates, 1987; Sniezek, Paese, & Switzer, 1990). Over-confidence is generally found to be much more prevalent when the items being judged are difficult, and less so when the task is easy, as defined by the number of subjects who answer the question correctly. "The *hard-easy effect* occurs when the degree of overconfidence increases with the difficulty of the questions, where the difficulty is measured by the percentage of correct answers" (Gigerenzer, Hoffrage, & Kleinbölting, 1991, p. 506).

An analysis by Lichtenstein and Fischhoff (1977) using one hard and one easy set of questions found pervasive overconfidence for hard questions but only found overconfidence for easy questions when confidence was .9 or 1.0 and underconfidence for less confident judgements. For these general knowledge questions, overconfidence declined as the percentage of correct answers increased, showing the hard-easy effect. Judgements which were made with a certainty of 1.0 proved to be overconfident, as accuracy was only .72 to .83 (Fischhoff, Slovic, & Lichtenstein, 1977).

In much of the reported research the difficulty of the task is defined by how well the subjects perform, that is, a difficult judgement would be one where few subjects answer it correctly and an easy one where most subjects answer it correctly. Juslin (1993) criticised this definition though and said that the hard-easy effect is due to the *post hoc* division of questions into these categories on the basis of the proportion correct, and that when item difficulty is defined in terms of low familiarity/high familiarity to the subjects, the hard-easy effect disappears. An alternative method of determining item difficulty was reported by Lichtenstein, Fischhoff, and Phillips (1982), by only using questions which required a numerical evaluation of information. Three types of questions were used: which city/state/country has the largest population (A or B); which place (A or B) is furthest from C; and which historical event occurred first (A or B)? Using this technique, questions could be defined as easy or hard via the ratio of the numbers within the question, with the largest ratios being easy questions and the smallest ratios being hard. For example, a question such as " which is furthest from London; Southampton or Edinburgh" is easier to answer than "which is furthest from London; Oxford or Cambridge", because the ratio of distances is much larger in the first instance. Large ratios presumably require less precise knowledge and are easier to answer. Their results found that when using this type of classification of difficulty, the hard-easy effect of overconfidence in hard questions and underconfidence in easy questions still occurred.

The hard-easy effect seems to arise because subjects do not realise just how easy or hard a task is in reality. The hard-easy effect is not an artefact of the test situation because if a person was well calibrated they would correctly estimate and allow for the level of difficulty and thus no relationship between accuracy and overconfidence would occur. It appears that there are probably two types of difficult questions, those which are difficult and the subjects recognise them as difficult, and those which are difficult but appear easy for some reason and thus subjects express too high a level of confidence in them. Perhaps some measure could also be made of how difficult a subject perceives the question to be, for example by using a simple rating scale, and this information compared with overconfidence. On balance, if there are deceptively hard questions included in a test then the hard-easy effect in overconfidence will be found. The question which needs addressing is what makes some questions deceptively difficult?

Information

Studies have found that in decision making tasks, confidence increases with the amount of information given (Oskamp, 1965) and the amount of observation time allowed (Ryback, 1967), but that accuracy did not improve in either of these experiments. The amount of information and the strength of that information influences people's confidence in their decisions (Koriat, Lichtenstein, & Fischhoff, 1980). Sometimes information which is uninformative is, however, not seen as useless, and confidence, but not accuracy, increases without justification (Oskamp, 1965).

In an experiment which involved predicting baseball team wins, Peterson and Pitz (1986) found that increasing the amount of information given to subjects reduced overconfidence, because it increased accuracy. They theorised that when one piece of information is given, judgements become extreme and confident, whereas when several pieces of useful information are given they conflict with each other and the resulting prediction is close to the average but with high uncertainty (low confidence), reducing overconfidence. This result conflicts with Oskamp's (1965) previously mentioned study which found that more information increased overconfidence. Peterson and Pitz (1988) explained this discrepancy by saying that their study did not measure confidence, as Oskamp's study did, but measured uncertainty - which they defined as being the person's beliefs about the variability of possible outcomes. They then showed experimentally that increasing the amount of information given to people can increase their confidence, as Oskamp found, but also leads to more alternative hypotheses being thought of and thus greater uncertainty as well.

Another important difference between these two studies was the value of the information given, as in Oskamp's (1965) study the information was useless and did not improve accuracy whereas in Peterson and Pitz's (1986) experiment the information was useful and did increase accuracy. This would also account for the increase in overconfidence in the first but not the second case, as confidence rises with the amount of information available, regardless of its usefulness. The experimental situation that was created increased confidence, but not accuracy, in Oskamp's study. Subjects might have believed that if an experimenter gives them information then it must be useful and they should do something with it, rather than just ignore it. Experimental situations may induce subjects to show more confidence because they do not dismiss information in the same way as they would in everyday life. This would then raise overconfidence if the information is not helpful, but the subject believes that it must be. Problems of measurement of overconfidence will be discussed in more depth later in this introduction.

Practice and Expertise

Experience in a task can improve calibration, which does show that the overconfidence bias can be reduced. For example Murphy and Winkler (1977) showed that weather forecasters, who were experienced at estimating confidence intervals, were well calibrated. There is also obviously great incentive for weather forecasters to be well calibrated, as it is their job to inform people of the probability of weather patterns. Similarly, Garb (1986) found less overconfidence in experienced clinicians than inexperienced judges. However, in another study, *greater* overconfidence was found for tasks which subjects considered they had more expertise in (Heath & Tversky, 1991).

Research by Armelius and Armelius (1976) indicated that how well a person believes they perform is more influential on confidence than how well they actually perform, thus if the person is more experienced and more proficient at a task then accuracy is more likely to match the high confidence people tend to exhibit. When assessing their confidence, people should consider the amount of practice or experience they have in the type of task they are attempting. For example, a person who has been a car driver for a year may feel confident in their ability to handle different road situations. Another person with twenty years of driving experience may feel similarly confident, but may in actual fact be much more able to handle any difficult road situations. The second driver may also have experienced a crash, or narrow escape, which has lowered their confidence to a level more appropriate to their ability. The influence of confidence on risk taking, and other behaviours, will be examined in the Discussion.

Base Rate

How we evaluate our own behaviour is dependent upon what we perceive are the norms of the group or society that we belong to. By comparing our behaviour or beliefs with those of the larger group we can tell if we are in-line or inconsistent with the norm. This may influence the resulting confidence in those beliefs or decisions, because being in accord with the group norm seems to validate one's confidence. For example, if a person believes that the Prime Minister is honest, but the rest of their friends say the opposite, then that person's confidence in their belief may be somewhat reduced by the knowledge that they are out of line with the group norm. The behaviour of a population, with regard to a specific event, can be expressed as a percentage, known as the *base rate*. This can be defined in both negative and positive terms, according to whether the presence or absence of a behaviour is being looked for, for example 30% of a population eat toast for breakfast or 70% do not eat toast for breakfast. Base rate information is akin to consensus information in attribution theory, i.e. how many other people do this?

Base rate/consensus information can be used to determine confidence in one's own beliefs, and it can also be informative when making predictions about the future. For example, if one knows that over the last few weeks ten people have won a major prize in the National Lottery, then the base rate for such wins is the total number of winners divided by the number of entrants. With this base rate knowledge the future likelihood of winning can be predicted. Previous research has, however, shown that people make their predictions and judgements, and ascribe their confidence in them, without giving enough weight to the relevant base rate of the population that experience the event. Neither can ignorance be blamed, because even when base rate information is given to people they tend to disregard it (Dunning, Griffin, Milojkovic, & Ross, 1990; Vallone, Griffin, Lin, & Ross, 1990).

The usefulness of base rate information varies according to the prediction being made, being more useful for some predictions about other people than for oneself. Osberg and Shrauger (1986) examined different influences on predictive accuracy, and stated that "the variables that might influence the accuracy of *self-predictions* can be categorized into three general areas: (a) properties of specific behaviors being predicted; (b) individual differences in reactions to particular behaviors; and (c) attributes of people making predictions" (p. 1045). They found that 94% of reasons generated to justify judgements fell into five categories; personal base rate 41%; circumstances 29%; personal disposition 18%; intention 5%; population base rate 1%. If people's self-reports of their thinking behind judgements are to be believed, base rate information was only used 1% of the time when making predictions about one's own future behaviour, with 99% of reasons referring to subjects' own behaviour, dispositions and situations.

Osberg and Shrauger (1986) showed quite clearly that subjects did not explain their judgements with reference to population base rates (only 1%). This, however, does not rule out the role of these base rates underlying personal base rate knowledge, that is, how often the event occurs for oneself rather than to the population. They drew an interesting parallel between distinctiveness, consensus, and consistency information used in attribution theory and personal disposition, population base rate and personal base rate respectively. In both Osberg and Shrauger's and Major's (1980) studies, "consensus or population base rate information was used least often, and consistency or personal base rate was the most widely used category" (Osberg & Shrauger, 1986, p. 1047).

Osberg and Shrauger (1986) said that when making self-predictions for events with extreme base rates, using personal base rates results in high accuracy, whereas at mid-range base rates using personal dispositional information is most efficient. The use of population base rates was found to be ineffective for self-predictions at all base rate levels. Personal base rates and dispositional information may be most useful for short term judgements, but they become less useful when making judgements about other people, as information may be lacking and thus population base rates become more useful. Population base rates may also be significantly more useful in long term judgements for example about future illnesses/health.

It seems, therefore, that population base rates are more useful when making predictions and attributions about other people, but it has been found that when people make judgements about other people they tend to ignore base rate information, and tend to rely instead on more individuating behaviour (Kahneman & Tversky, 1973; Bar-Hillel, 1983). High levels of confidence are also often placed in those judgements which are founded upon brief personality descriptions (Kahneman & Tversky, 1973) and neglecting to consider sample size, with high confidence in small samples, is another common judgemental error (Tversky & Kahneman, 1971).

In their study, Dunning, Griffin, Milojkovic, and Ross (1990) found that when subjects predicted that a person would behave differently from their peers, that is, contrary to the base rate, this prediction was more likely to be inaccurate and therefore overconfident. When making predictions of other people's behaviour, or making attributions about them, people tend to ignore the situation, and what is the most normal behaviour for the general population, and instead make predictions based on internal personal factors of the person. Thus by ignoring the situation, and base rate, they are more likely to make errors of judgement. This is made worse by people placing more confidence in judgements based on internal factors, such as personality, which seem more predictable, than on unpredictable external ones. People who are overconfident exhibit a fundamental attribution bias by placing too much emphasis on inferred or known personality traits of the target, and neglecting to consider the influence of the situation and context on the behaviour to be predicted.

The Locus of Uncertainty

Two types of information are available to a decision maker: prior knowledge and external information, both of which have associated uncertainties attached to them and which influence confidence. How accurate we believe we are determines confidence and the factors which affect accuracy have to be assessed by people to determine their confidence. Thus biases arise when people fail to consider or give sufficient weight to relevant variables which do affect accuracy, or, as Kahneman, Slovic, and Tversky (1982) pointed out, they may believe that an irrelevant variable does influence their accuracy, when in fact it doesn't. The roles of different variables are uncertain and unquantifiable in many circumstances.

Kahneman and Tversky (1982) pointed out that the causes of uncertainty can be either internal, in as much as the state of one's own mind is at question ("I think that I have read that book"), or uncertainty can be external to the person ("I think that Sheffield Wednesday will win their next game"). The perceived locus of control over an outcome is also important, for example Howell (1971) found that subjects were more confident about judgements where they had some control over the outcome (throwing a dart rather than the spin of a roulette-type wheel). We have more control over internal uncertainty than over external uncertainty (Howell & Burnett, 1978) and comparisons of overconfidence for both will be made in this thesis.

Motivation and Self-Presentation Biases

How certain or confident we are about something depends then upon many factors involved in cognitively assessing our uncertainty. Problems of recall or weighting of information may bias internal confidence. Apart from such cognitive factors other factors come into play. Motivational influences can be responsible for overestimating the probability of events in some situations, as for example weather forecasters who overestimate the likelihood of rain (Murphy & Winkler, 1974) because people would rather be prepared for the worst than caught out. Other professional judgements may be overconfident if those people are expected to inspire confidence in other people. For example, we do not expect experts, such as doctors, to appear unconfident in their judgements, and we may feel anxious about accepting their advice if they did not exude confidence. According to Lichtenstein, Fischhoff, and Phillips (1982), there may be times when people are not motivated to be honest in their assessments of confidence, as situations may reward or punish honesty differentially. Thus subtle pressures to conform, impress, or deny may be strong reasons to be mis-calibrated in one's judgements.

The costs and rewards of publicly appearing confident must be considered. Being accountable for one's judgements can depress confidence ratings, as people become unwilling to express high confidence in their judgements, to avoid embarrassment if they are proved wrong (Tetlock & Kim, 1987). Tetlock and Kim examined whether self-presentation biases alone were operating, but found that much more was occurring. Their research showed that accountability for a judgement changed the cognitive processes involved in impression formation, in the task they used. Subjects who were told they were accountable before the task, were more accurate and less overconfident than those told after the task, or those not accountable. Two conclusions became apparent: first, that accountability motivates people to cognitively process social information in more complex ways, leading to a reduction in overconfidence; and second, that accountability can to some degree prevent the overconfidence bias but not reverse it; as being informed of one's accountability after the initial processing has occurred is not effective in causing reinterpretation of the information and does not reduce overconfidence. They postulated that accountability increases self-consciousness which promotes more analytical processing of information rather than automatic shallow processing, which may result in people considering the opposite and considering that their answers may be wrong. If people have increased motivation, caused by being accountable for their actions and decisions, better calibration occurs.

Two interesting possibilities were raised by Mayseless and Kruglanski (1987) in their study of recognition of tachistoscopically presented digits under conditions of high or low fear of invalidity, in which some subjects wanted to be correct more than others. Subjects in their high need-for-validity condition displayed behaviour consistent with more processing of the visual information, and higher resulting confidence. This could either be due to the increased level of processing, resulting in more confidence, or the desire to be correct causing motivated subjects to believe in their answers more strongly, due to wishful thinking. Whether deeper processing or motivation raised confidence is unknown, but the fact remains that situational pressures in the experiment resulted in different approaches to the task and different resulting confidence. The obtained levels of accuracy for the different conditions were not reported though, so overconfidence cannot be assessed.

This issue is, however, not so straightforward, as Sieber (1974) found that increased motivation, and thus arousal (in students who believed they were taking a mid-term test rather than a preparatory test), resulted in similar accuracy scores to those less motivated students, but greater overconfidence. Tetlock and Kim (1987) pointed out that there may be an optimal level of arousal, because in Sieber's study high arousal, linked to increased importance of a decision task, produced more overconfidence, and thus they concluded that "arousal is related in a curvilinear fashion to integrative complexity, with moderate levels most conducive to complex functioning" (p. 707). Arousal changes as the stakes involved in the outcomes of a task vary. The effect of increasing the stakes in judgement tasks has been shown to have mixed results, Fischhoff (1982) found no improvement in judgement, but other researcher have found changes in both directions (cf. Kruglanski & Freund, 1983), and this is an obvious source of motivation which needs more research.

Making Decisions / Choices

Motivation seems, therefore, to influence overconfidence, but how also does overconfidence change when a decision has to be made? Sniezek, Paese, and Switzer (1990) found that judges who chose an answer publicly on a general knowledge test were less overconfident than those who did not have to state their choice of answer, and both of these groups were less confident and more overconfident than a third group who were given a pre-selected answer in which to rate their confidence. Similar results were found by Ronis and Yates (1987), in their comparison of subjects who either did or did not make a choice on general knowledge tests and basketball game predictions. Paese and Sniezek (1991) proposed that if subjects make a decision based on their judgements, their subsequent confidence would be higher due to increased commitment in their answer. The reverse was found, however, overconfidence dropped significantly when a decision preceded the confidence rating of the judgement. Thus committing oneself to a decision seemed to temper high confidence, possibly through the previously mentioned factor of accountability.

Individual Differences

Cognitive factors alone probably do not account for the huge individual differences observed between subjects in the calibration research literature; personality traits should also be studied for their influence on confidence. Wright and Phillips (1976) took seven measures of calibration and found two correlations with authoritarianism (.41 and .34) but not with any measures of conservatism, dogmatism, intolerance of ambiguity or verbal expressions of uncertainty used in answering questions.

Because confidence is an expression of a subjective feeling of certainty it would seem logical to expect that those people who are more introspective will differ in how well calibrated they are from those people who are less introspective and less self-monitoring. In Shrauger and Osberg's (1982) study subjects who were low in private self-consciousness used more personal dispositional explanations (20.2%) than those subjects who were high in private self-consciousness (14.7%). They also found that subjects high in self-confidence used personal base rate information more than low confidence subjects (45.6% vs. 35.7%). In general subjects who showed high public self-consciousness, indicating concern with self-presentation and less introspection, were found to be less accurate in their self-predictions. Similarly subjects high in private self-consciousness (more introspective) were more accurate in their predictions of their own behaviour. An obvious extension of this research would be to relate it to overconfidence as well as to accuracy of self-predictions. Further research into individual differences will be discussed later, in the introductions to Chapters 4 and 5 which will study relationships between overconfidence and individual differences.

Mood / Depression

Mood state may influence confidence or accuracy and thus overconfidence. Theoretically, mood may make certain information more easily accessed from memory and therefore this may bias the judgemental processes affecting accuracy or confidence. Mood may also influence the self-evaluation of the person and alter confidence. In an ecologically valid study of predicting one's future life experiences, Dunning and Story (1991) found depressed students to be more pessimistic than non-depressed students, in line with previous research. Their experiment, however, also measured the actual outcome of the predictions and found that depressed subjects were *more* likely to experience negative events and as such they were *less* accurate and more *overconfident* in their predictions. They concluded that although depressed subjects are more pessimistic they are less realistic in their judgements, contradicting the generally accepted "depressive realism hypothesis" which stated that depressed subjects are more realistic and less prone to biases of judgement.

In a study by Allwood and Björhag (1991) which experimentally altered mood state, no differences in overconfidence were found between sad subjects and control subjects. Dunning and Story (1991), however, found that mildly depressed individuals were less confident but also less accurate in self-predictions of future events and thus were more overconfident. This contradictory evidence may be reconciled if the different natures of the tasks and differing levels of mood are considered. For example, the depressed mood was more severe and naturally occurring in Dunning and Story's study, and also the self-prediction type task may be more affected by mood than a general knowledge type task. Further research in this area is obviously required.

Sex Differences

Maccoby and Jacklin's (1974) review of the literature concluded that there are few sex differences with respect to intellectual and academic ability, achievement motivation and self-esteem. Self-confidence, however, they reported as being lower in women than in men. This may have negative implications, as motivation and initiative may be reduced when poorer performance is predicted, less taxing tasks may be preferred and people may tend to give up more easily (Weiner, Frieze, Kukla, Reed, Rest, & Rosenbaum, 1971). Women are also more likely than men to show low self-confidence in achievement settings (Maccoby & Jacklin, 1974).

The level of confidence expressed by women compared with men appears to be situationally dependent, and the type of task being undertaken is also relevant (Lenney, 1977). Lenney stated that women are less self-confident and devalue their performance as compared with men when feedback about their performance is either ambiguous or absent. Sex differences in self-evaluation of performance disappear, however, if feedback is unambiguous (Feather & Simon, 1971; McMahon, 1973). McCarty (1986) found that women were less self-confident than men, regardless of whether positive, negative or no feedback about performance was given. No measure of accuracy was taken however, so overconfidence could not be assessed, or whether males were in fact performing better and thus if their higher self-confidence was appropriate.

Lenney (1977) drew a distinction between those tasks and situations which people expect to be evaluated in and those which are non-evaluated, and said that in the former, social situations, women are less confident than men. Another important factor which effects expressed confidence in predicting how well a task will be performed is whether the task is perceived as being congruent with one's sex role. Tasks which are presented as inappropriate for one's sex result in lower expectations of success and lower confidence (Stein, Pohly, & Mueller, 1971). Nicholls's work with children (1975) showed that if a task is previously described as being about general smartness then boys felt better about it than girls, who in turn felt better if it was described as a size perception task. Similar results have been found for adults, as women are expected to perform better than men on a task described as social rather than labelled as intellectual (McHugh, Fisher, & Frieze, 1975, cited in Lenney, 1977).

These research findings consistently point towards differences existing between males and females in confidence, but how does this relate to overconfidence? Are men and women equally well calibrated? In a study of children Sieber (1979) found that girls were less overconfident than boys, perhaps because males may be more socially rewarded for bravado and overconfidence and the reverse for females. Less overconfidence was also found for girls than boys, in a task of estimating the number of dots, by Newman (1984). In general either no differences between males and females are found in calibration or overconfidence in adult samples, for example Lichtenstein and Fischhoff (1981) and also Gigerenzer, Hoffrage, and Kleinbölting (1991), or they are not looked for or reported. However, considering the different perspectives of researchers in the field of calibration of confidence and those interested in gender issues, it is plausible that the role of gender in overconfidence has not been adequately assessed.

Overconfidence in Visual Perception

Apart from confidence in social and knowledge based judgements, research has shown that for perceptual tasks, where little processing other than sensory encoding is required, underconfidence is the norm and it is unresponsive to feedback (Björkman, Juslin, & Winman, 1993). This corroborates the early finding that when recognising words presented tachistoscopically with increasing illumination over ten trials, subjects were underconfident in their judgements (Adams, 1957).

Dawes (1980) proposed that people overestimate their intellectual capacities, but are more accurately calibrated when perceptual tasks are concerned. His results were inconclusive but later research by Keren (1988) using a perceptual task with Landolt rings found underconfidence in perceptual judgements, which increased when the task was easy, and overconfidence for the same subjects in a traditional general knowledge type task. In general subjects are less confident when making perceptual judgements, and even underconfident. It may be that people do not trust perceptual information, that motivational factors reduce confidence in perceptual tasks, or that no high-level processing is required at which point errors can occur. Confidence in auditory perception and other senses also needs investigating to see if over- or underconfidence occurs, and what factors influence it.

Winman and Juslin (1993) argued that the nature of confidence assessments are intrinsically different for cognitive and sensory tasks, as shown by the fact that feedback has no influence on calibration in sensory discrimination tasks, but improves calibration in cognitive tasks. It was proposed that underconfidence is impossible to avoid because of the fixed sensitivity of the sensory system (Björkman, Juslin, & Winman, 1993). More recently, Ferrell (1994) has argued that a single model accounts for both cognitive and sensory tasks, proposing that higher perceived difficulty and lack of control over sensory tasks reduces confidence and results in underconfidence, and that the hard-easy effect can also be observed for sensory tasks (Ferrell, 1995).

Problems with Measurement

The way that overconfidence is defined or measured may have an influence on the magnitude or even existence of the cognitive bias. Procedures for measuring overconfidence must be carefully considered, for example, to accurately assess confidence it should be measured before any feedback about the correctness of the answer is given, to avoid inflation of confidence due to the hindsight bias, which occurs when subjects retrospectively overestimate how well they knew the answer (Fischhoff, 1977). May (1991) pointed out several problems with this area of research, such as item selection, forced scale use, differences in singular and frequentistic judgements and normative ambiguity of calibration. These problems affect overconfidence research and these issues and others will now be addressed.

Item Selection

The first problem in overconfidence research is with item selection, in that some items can produce overconfidence in subjects and some items produce underconfidence. Thus in previous studies it may be argued that the level of overconfidence found could be due to the selection of items used. May (1991) argued that the calibration method of measuring self-serving human biases is questionable. She pointed out that there is considerable variability of over- and underconfidence and that many questions produce underconfidence. Experiments which only select deceptive items, where an alternative answer to the correct one seems logically correct but is actually wrong, will lead to overconfidence being found. Fischhoff et al. (1977) also reported extreme overconfidence for deceptive items (73%) and much less (< 9%) for non-deceptive items. Deceptive items have been studied a lot because it is interesting to discover what makes them deceptive. May argued that by eliminating deceptive items, overconfidence would be eliminated too. However, many judgements in life are deceptive and thus eliminating them from experiments would also distort the truth.

It has been argued that overconfidence is largely due to test difficulty (Schraw & Roedel, 1994), and it has been argued that overconfidence is not a cognitive bias at all, but is only due to the biased selection of items used in these types of experiments (Gigerenzer, Hoffrage, & Kleinbölting, 1991). Gigerenzer et al. argued that people are well calibrated in real life and that overconfidence and the hard-easy effect are artefacts of experimentation created by using artificial problems or biased samples of general knowledge questions. To avoid biased samples of questions Gigerenzer et al. asked subjects in their experiment to state which of two randomly selected German cities has the largest population and how confident they were in their answer. Their results showed that compared to a group given non-randomly selected general knowledge questions the random-selection group was well calibrated, whereas the former group was poorly calibrated. However, when the questions were matched for item difficulty no differences in overconfidence between the two groups occurred. Griffin and Tversky (1992) argued that Gigerenzer et al. confounded item generation techniques with item difficulty, and thus incorrectly concluded that overconfidence results from biased item selection. Griffin and Tversky then went on to show experimentally that even with random selection of items overconfidence still occurred.

Taking an ecological approach, Juslin (1993) also argued that in everyday life people are well calibrated and that they use general rules (heuristics) which do have high validity. He argued that people only seem biased because they are cognitively calibrated to their environments and experimental problems are not representative of the reference class that people are adapted to. By randomly selecting questions for a general knowledge test rather than an experimenter selecting questions, Juslin showed almost perfect calibration of subjects rather than the usually found overconfidence, as did Björkman (1994).

It is possible that by trying to determine just where biases in human judgement lie, there has been an overwhelming tendency to only carry out or report experiments which discover biases. The difficulty of the task used is obviously important because of the hard-easy effect, and thus if hard tasks are consistently used in experimentation then overconfidence will obviously be found. Using easy questions, a balance of question difficulties, or controlling for difficulty across conditions, should also be considered when designing experiments. There may have been too much emphasis on looking at where people go wrong in judgements and not enough at where they are right, which the more recent research by Juslin (1993), Griffin and Tversky (1992), and Gigerenzer, Hoffrage, and Kleinbölting (1991), has attempted to address.

Framing

Like previous researchers, such as Kahneman and Tversky (1979), Levin, Chapman, and Johnson (1988) found that the *information frame*, for example, describing a gamble in terms of chances of winning rather than chance of losing, alters subjects' judgements. It may be the case, therefore, that an artificial bias may arise because of the way subjects are tested, as experimenters tend to ask for the probability/certainty or confidence that the answer given is ***correct.*** If the question was rephrased to ask for the subjects' confidence that they are *incorrect* a reduction in the overconfidence bias may be apparent. This problem, however, may look like an artificial testing situation but mirrors real life where in fact people do tend to ask themselves if they are correct rather than if they are incorrect (confirmatory biases). Are those people who are well calibrated asking themselves if they are wrong rather than if they are right, and thus becoming better calibrated as a result?

Is the language used in experiments, contributing to the bias which is found? If the experimenter asks for a person's confidence that they are correct this may bias the subsequent information search and evaluation processes to only look for reasons why the answer is correct. Sniezek, Paese, and Switzer (1990) studied this possibility when measuring overconfidence in general knowledge tasks and found that framing had no effect on overconfidence. They concluded that "the well-known tendency to *overestimate* the likelihood of one's choices being correct is balanced by an equally strong tendency to *underestimate* the likelihood of one's answers being wrong" (p. 273). If the experimental situation tends to frame questions in a certain way then it seems to be framing them in the same way as people do in real life, and is therefore arguably still useful.

Forced Scale Use

Confidence and accuracy are related, which might imply that overconfidence is a problem with translating feelings into probabilities. If the proposition that internal confidence is well calibrated, but that translating it to a number causes biases, then perhaps verbal descriptions of confidence, such as "highly likely", are more accurate and less biased (Zimmer, 1983, 1986). It may be the case that confidence is well calibrated internally but that it becomes distorted when trying to translate feelings into numbers or words. If this is the case, then training people in the use of accurate statements of probability may reduce overconfidence and improve calibration. Translation problems probably do not account for the overconfidence effect as a whole, because previous research has shown little transfer of calibration across tasks, which would be expected if translation problems had been remedied to some extent by training in calibration. Also, research using many different methods of measuring confidence, such as marking points on scales rather than assigning a number, has also consistently found overconfidence, indicating that translation difficulties are not the cause of the bias.

Differences in Singular and Frequentistic Judgements

Overconfidence is reliably found by summing confidence and accuracy across individual items, but underconfidence is found when subjects are asked to state how confident they are for a set of items as a whole. May (1991) asked subjects in her experiment to estimate how many questions they had got correct on the whole test (frequency estimates), after it had been completed. When subjects were questioned via this procedure they underestimated the number of answers that they had got correct, which she argued shows underconfidence in knowledge not overconfidence as shown by calibration curves. Similar results were obtained by Schneider (1995). Gigerenzer, Hoffrage, and Kleinbölting (1991) also found that when people are asked for their estimates of correctness and confidence retrospectively for an aggregate (or whole set) of answers, good calibration occurs. They also criticised the calibration research by arguing that probabilities cannot be said to apply to single events, that comparing relative frequencies correct with confidence or subjective probability is misleading, and that frequency judgements are well calibrated.

Alternatively the reduction in overconfidence for frequentistic judgements may arise because confidence in an individual item is weighed up on the basis of evidence for each answer and overconfidence results, whereas confidence in a set of answers depends upon a judgement of the task difficulty, one's ability and effort put in which may result in underconfidence (Sniezek, Paese, & Switzer, 1990). It may be that subjects do not sufficiently take into account their fallibility for individual judgements but tend to overestimate fallibility when asked for confidence in a group of judgements. This may occur because people can think of reasons why each individual answer may be correct, but for a group of questions they are not asked for individual answers but for the probability that some answers are incorrect, which they can assess fairly accurately. Research should also be carried out to see if overconfidence only appears to be reduced for frequentistic judgements because of demand characteristics of the situation, as people may not like to boast about how many questions they answered correctly.

Frequentistic judgements of accuracy are likely to be different from how many were expected to be correct, as implied from subjective confidence, because of the operation of the anchoring and adjustment heuristic in the first case. When asked how many questions you got right out of 100 people probably estimate from one anchor, either 0 or 100. If the task was difficult it may be easier to think how many you got right, for example 10 definitely correct and 5 possibly correct therefore estimate 15. If the task was easy you may estimate that you got 10 wrong but allow for error and say 15 and thus estimate 85 correct. The tendency may be to overestimate from the anchor 0 for hard questions and 100 for easy questions so producing overconfidence for hard tasks and underconfidence for easy tasks in frequentistic judgements. This may also be linked to thinking of reasons why you got answers wrong which reduces overconfidence and thinking of reasons why you got answers right which tends to lead to over-confidence. Research into the hard-easy effect in relation to frequentistic judgements is an avenue still to be explored.

Ecological Validity

The type of situation being studied must be always be borne in mind, and generalisations to other judgemental tasks and real-life situations must be made carefully. The artificiality of the types of tasks used may give the impression of a biased judge, whereas the task itself may be creating the bias. Experimental situations are in many respects artificial, and they lack the necessary situational rewards and punishments for true validity (Kogan & Wallach, 1964). This leads naturally to the question, are psychology experiments finding overconfidence because the experimental situation does not allow time for sufficient thought? Are subjects producing rushed answers to questions which would be tempered by slower deliberation in real life, and giving a false measure of overconfidence? Experimental research tends to study quick, instantaneous decisions and judgements, because of the nature of the research situation. Longer deliberative processes and the influence of time on confidence should also be studied in future.

Measurement for Individuals

Measuring overconfidence for an individual person has some associated problems. To calculate overconfidence for an individual person, a large number of judgements need to be made before an accurate measure of overconfidence can be calculated. The type of task also affects the level of overconfidence measured for a person, some tasks may result in more overconfidence for people than other tasks. If people were unbiased and precisely calibrated then the type of task and its construction would make no difference, because the perfect judge would correct their confidence score to account for the task type or difficulty. Since people are not perfectly calibrated, some tasks may be more unfair to some people than to others when measuring their overconfidence. There may be an ideal test for each person where they are well calibrated. However, problems with selection of questions seems to be more controversial and relevant for general knowledge type tasks and ability type tasks, and less so for social- and self-prediction type tasks.

Recalling the Past vs. Predictions of the Future

A valuable distinction was made by Wagenaar and Keren (1986) who referred to probability assessments about future events as *predictions* and probability assessments regarding past or present events as *postdictions.* Calibration may be different for future-event predictions than for general knowledge questions. Different judgemental heuristics may operate for tasks which require construction/inference (predictions) rather than straightforward retrieval (postdictions) (Kahneman & Tversky, 1982). If a person shows good calibration on one type of task it does not automatically indicate that they will also have good calibration on another type, as, for example, Ronis and Yates (1987) showed that calibration of predictions of basketball games did not correlate highly with quality of predictions in a general knowledge test.

Research by Fischhoff and Beyth (1975) showed that compared to “almanac” (general knowledge type) questions, predictions for future events showed much better calibration of confidence and accuracy. Wright and Wisudha (1982) also compared future events and “almanac” questions, and found that 100% confidence was used much more frequently for general knowledge questions than for future predictions. No firm comparisons of predictions and postdictions could be made in this experiment, however, because the difficulty levels of the tasks were not matched. Another experiment, by Wright (1982), compared overconfidence for past events and future predictions, by rephrasing questions such as “has at least one member of the British Parliament died within the last fourteen days?” to “will at least one member of the British Parliament die within the next fourteen days?” By using this technique, task difficulty was held constant at around 80%, and overconfidence was found for past events (2%) and underconfidence for future events (-3%).

The difference between assessments of one’s knowledge and future predictions is that the latter requires the person to assess the impact of unforeseen and unknown variables, whereas assessments of one’s knowledge derive from the strength of a memory, inference or a logical deduction from the facts. This difference is important, because when working out an answer there is uncertainty, whereas when recalling an answer there is less uncertainty. This also relates to the source of the uncertainty, which is both internal and external for predictions and social judgements, but primarily internal for confidence in one’s own general knowledge. It may be the case that when people make predictions they know they are making inferences and do allow for uncertainty, but they do not realise they also make inferences in knowledge type tasks, and therefore do not sufficiently allow for error when assessing their confidence (Fischhoff et al., 1977).

Hastie and Park (1986) reviewed the literature on social cognition and drew the distinction between "memory based" and "on-line" tasks. In the former type of task, information is retrieved from long-term memory but on-line tasks require new incoming information to form and re-form the judgement, however, research is equivocal in this field and Hastie and Park drew no conclusions. It would be interesting to compare two tasks -- one requiring knowledge, the other needing subjects to work out the answer, and to compare the resulting confidence, accuracy, and overconfidence. Feedback which encourages thinking and processing of information for predictions results in better accuracy later on and a better calibration of confidence (Swindell & Walls, 1993), and the underlying processes of this improvement need further research to clarify them.

The location of uncertainty, internal for judgements where lack of knowledge is the cause of uncertainty and external when the environment is uncertain, may also shed light on the hard-easy effect. Hard questions may be hard because they require assessing uncertainties, whereas easy questions may require recall and little estimation of uncertainty so resulting in underconfidence. The influence of situational variables may not be given enough weight for hard questions, but this may occur less for easy questions, thus giving lower overconfidence. The source of uncertainty may be influential and it would be wise not to generalise findings from general knowledge type questions to predictions of the future and other real-life decisions where uncertainty is external to the person rather than internal.

Reducing Biases

Many experiments have manipulated experimental conditions to see what factors can reduce overconfidence, and this information may then help in formulating theories as to the causes of the overconfidence bias. Fischhoff (1982) pointed out that practice and reinforcement are necessary prerequisites of learning to be better calibrated. When no feedback is given in a repeated task, confidence tends to increase, but if no increase in accuracy actually occurs with practice then overconfidence will increase. This was found to be the case in a task of baseball predictions by Paese and Sniezek (1991).

Hoch (1985) asked subjects in his experiment to predict their future behaviour for positive events, where unrealistic optimism was expected, his aim being to see if getting subjects to generate con reasons, why an event would not occur, would reduce overconfidence. He found that generating con reasons increased subjects' accuracy in their predictions but generating pro reasons did not. This, however, did not occur for predictions with very high base rates, where accuracy was already very high.

Oskamp (1962) asked subjects in his experiment to decide if a patient described in an MMPI (Minnesota Multiphasic Personality Inventory) was psychiatrically or medically ill. Initially overconfidence in judgements was found, which was reduced by training from the experimenter. What is interesting is that the improved calibration was via improved accuracy for subjects trained in improving their performance, (accuracy increased from 67% to 73% and confidence was stable at .78), and reduced confidence for those subjects trained in calibration (confidence dropped from .78 to .74 and accuracy remained at 68%). Thus, it seems that overconfidence can be reduced either by increasing accuracy, as Oskamp and also Hoch (1985) found, or by reducing confidence through training.

To find out what mental processes occur when answering general knowledge questions Allwood and Montgomery (1987) asked subjects to "think aloud" and four resulting categories were identified: immediate recognition of the answer, inference, intuition and guessing. Of these four strategies the highest level of accuracy was due to immediate recognition and this had the highest associated confidence. Confidence decreased across the categories from inference, intuition and the lowest for guessing. They concluded that if subjects gave more consideration of the type of strategy used in coming up with an answer their confidence judgements could be made more appropriate. They also found low accuracy for answers arrived at by inferences and suggested that warning people not to trust their inferences could be used in debiasing techniques to help reduce overconfidence.

Earlier on it was mentioned that Adams and Adams (1961) instructed subjects that "of all those decisions made with confidence *p*, about *p*% should be correct" (p. 37). This was done to reduce subjects' uncertainty about what the rating scale represented and expectation/confidence could then be then compared with accuracy across the whole scale. This very procedure may, however, have influenced subjects' confidence. By defining confidence in this way, subjects are indeed giving subjective likelihoods, but firstly this may not be the same as confidence as they would usually understand it, and secondly, confidence in judgements may be biased, or more likely de-biased, by this instruction, which instructs them to be well calibrated.

An improvement in calibration over time may be explained in terms of the subject receiving cues from the situation (such as feedback about performance), which alter the internal state of certainty either by increasing doubt and reducing confidence or by raising certainty and confidence. Adams and Adams (1961) proposed that increased experience of a situation may promote better discrimination of internal cues, such as feelings of doubt, resulting in better calibration over time. Experiments have been done which try to de-bias subjects in general knowledge type tasks, but not apparently for social-type judgements. This may be because a time lag would occur, but there may be ways around this, for example using social judgements of other people's behaviour or by using computers. Immediate feedback is used in debiasing studies, and applying this to social judgements may be harder than for memory based or reasoning tasks, but may be achievable. Studying debiasing of social type judgements would obviously have many implications for real-life situations, as people judge other people so frequently.

The beliefs that we hold influence how we perceive and process information, and can bias our judgements (Allport, 1954; Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980). Current attitudes and beliefs influence how we perceive new information. People tend to cling to their established impressions, beliefs and theories and to do this they attend to incoming information in a biased manner, to find confirming evidence and criticise disconfirmatory evidence (Lord, Ross, & Lepper, 1979).

When a task is new, or judgement unfamiliar, confidence is low and feedback is sought and evaluated to improve performance. The initial attempts at a task when confidence is low may be receptive to feedback, resulting in greater attention to calibration and modification of confidence. However, a point may be reached beyond which the person believes that a sufficient level of accuracy has been reached so that feedback can be increasingly disregarded, and as a result neither reduction in overconfidence nor better calibration is achieved. Little has been found out about how feedback affects underconfidence. With respect to individual differences it may be that individuals who are more open-minded to feedback are the ones who end up most accurately calibrated. The final level of confidence in a judgement may be influenced by the amount of conflict present in the decision making process (Janis & Mann, 1968) or the level of cognitive dissonance caused (Festinger, 1957).

People like to be able to explain the world around them, and tend to spontaneously search for explanations (Weiner, 1985) but once an explanation is arrived at the search tends to stop and other alternatives never considered (Shaklee & Fischhoff, 1982). The training of people to be more accurate judges could involve trying to change the system of information searching, but this is probably so ingrained and natural that attempts to alter it may not work. Alternatively changing the way people reassess facts and decisions may be more amenable to training. Perhaps by universally reducing confidence in one's judgements a person may be more likely to keep searching for more information and thus become a better judge. Perhaps people who are initially unsure, and have low confidence, keep searching and thinking until they have assessed many possibilities and weighed them up and subsequently have high confidence in those answers. Other people may, however, have high confidence in the first answer they arrive at without looking at other possibilities. Experiments which study confidence throughout the decision making process would be valuable.

Theories of Overconfidence

To conclude: the accuracy-confidence relationship is situation specific and can depend upon the characteristics of the problem being judged and the person making the judgement. Mis-calibration or overconfidence has been shown to occur. There are two possibilities; either people really are incorrect in their level of certainty or they are not, in which case overconfidence is an artefact of experimentation. If a bias exists in reality, and people are not well calibrated, this may occur for several reasons: they may truly be overestimating their certainty; they may not be able to accurately express their level of confidence; or they may have reasons for distorting the confidence that they express because of self-presentation biases. Alternatively people may not be biased, but the tests used by experimenters may be unfair and artificially creating a bias which does not really exist. The overwhelming amount of evidence, however, does support the concept of a psychological bias, and thus the parameters of this bias need to be explored, which this thesis aims to do.

It seems that the cause of the overconfidence bias is either that people truly believe that their accuracy is going to be higher than it really turns out to be (they fail to take into account all the factors which then reduce accuracy), or that they artificially inflate their perceived level of accuracy for a variety of reasons, such as to delude themselves or other people (thus protecting self-esteem or giving favourable impressions to other people). The phenomenon of under-confidence also exists where confidence is too low. This is due to accuracy being higher than expected, although some situations may demand low confidence for a person to appear modest. Alternatively if underconfidence appears to be a pervasive trait in a person it could be because they are using inaccurate methods for assessing future outcomes, or that too much information is overwhelming and confusing them, thus reducing certainty and confidence.

Wagenaar (1988) proposed that overconfidence arises because of problems with inferential processes, and that people do not realise the reconstructive nature of memory. Gigerenzer et al. (1991) proposed a model of the processes involved in confidence. Initially, they said, local *mental models* (MM) are constructed which are solutions using information from memory and also elementary logical operations. If this fails to produce an answer then *probabilistic mental models* (PMM) are used to come up with a solution, using probabilistic information from the environment. Gigerenzer et al. stated that the choice of answer and confidence in it are both expressions of the same conditional probability, and that "PMM theory predicts that the distinction between confidence and relative frequency is psychologically real, in the sense that subjects do not believe that a confidence judgement of X% implies a relative frequency of X%, and vice versa" (p. 512). Their model explains the choice of answer quite well. They also assume that confidence and choice are one process rather than a choice followed by information search then confidence judgement, as proposed by Koriat, Lichtenstein, and Fischhoff (1980).

Koriat et al. (1980) proposed that overconfidence results from biases in information processing, such that recall from memory, either during or after the decision making process, is biased towards evidence supporting a tentative answer and not against it. To test this hypothesis, attempts were made to de-bias subjects (reduce their overconfidence) by asking them to write down all the reasons they could think of why each of the multiple choice answers was right or wrong before they chose an answer and stated a probability that their answer was correct. This method resulted in subjects becoming very well calibrated for all levels of confidence, except the very highest, and significantly reduced overconfidence. The generation of pro/con reasons had no influence on subjects' accuracy levels, but did influence confidence, resulting in reduced overconfidence. Further research by these investigators, reported in the same paper, revealed that the important factor was the production of reasons that contradicted the chosen answer. In their experiment subjects were required to generate either one piece of supporting evidence, or one supporting and one contradicting, and neither of these procedures had any effect on calibration. This finding suggests that it is only generating reasons why the answer may be wrong which reduces overconfidence.

Koriat et al. (1980) proposed that people automatically tend to produce supporting reasons for their answers, and therefore when required to generate additional supporting reasons, this process has no effect on overconfidence. Perhaps this only applies though to overconfidence, whereas when people are underconfident maybe they naturally tend to produce contradicting reasons, and so instructions to generate supporting reasons may have a positive effect on calibration. Koriat, Lichtenstein, and Fischhoff's experiment resulted in global *over*confidence, so there is no way of telling from their results whether *under*confidence can be reduced (calibration improved) by the production of reasons why the answer is correct. There may also be individual differences operating, such that overconfident people tend to produce supporting reasons for their answers and underconfident people produce contradicting evidence for answers. There is no experimental evidence to shed any light on this, as far as I am aware.

Koriat et al. (1980) analysed the reasons given in support of (for) and against each of the possible answers in their experiment. Subjects gave significantly more *for* reasons than *against* reasons irrespective of the chosen answer. There is also strong evidence that the amount of confidence correlates positively with the strength of the reasons *for* the chosen answer ($r = .56$) and with the strength of reason *against* the rejected answer ($r = .37$). The strength of the evidence for the rejected answer and against the chosen answer had no bearing on overconfidence. In conclusion, the level of confidence seems to rely upon the strength of evidence supporting the chosen answer.

According to Koriat et al. (1980), people search their knowledge to come up with an answer and then review the evidence and assess their confidence in that answer. Following this interpretation, there are two points where biases can appear: first, people may generate reasons *for* an answer and favour it over reasons *against* it; and second, they may also disregard evidence contradictory to the chosen answer. Generating contradictory reasons against the chosen answer shifts one's overall overconfidence lower down the scale and improves calibration, but even at this lower level the supporting evidence is still what determines the relative degree of confidence. Asking subjects for both supporting and contradicting evidence did not reduce the bias, the strength of supporting evidence seemed to outweigh any contradicting evidence generated. Thus debiasing techniques should encourage the generation of contradicting evidence and suppress the generation of supporting evidence.

The finding that generating reasons why an answer is correct does not reduce overconfidence, but challenging an answer by asking why it may be wrong does improve calibration, has been found repeatedly with general knowledge tasks (Koriat et al., 1980) and Hoch (1985) also found it with students' predictions of their future jobs. Koehler (1991) reviewed the research on explaining and imagining events and concluded that both of these processes require the person to temporarily believe that the hypothesis/statement is true, and this makes the perceived likelihood and associated confidence increase. Related research has shown that trying to imagine a future event which is difficult to imagine increases mental effort and reduces the person's likelihood estimates and confidence (Sherman, 1985). People tend to look for supporting evidence and are insufficiently critical, so asking them for reasons why they may be wrong provokes thoughts which otherwise tend not to occur. For an excellent review of the literature on how explanation and imagination influence the processes of accuracy and confidence in judgements, see Koehler (1991).

A model to account for the finding that overconfidence drops when people make a decision before expressing their confidence was proposed by Sniezek, Paese, and Switzer (1990). They proposed that when subjects are asked for their confidence in a pre-selected answer they think of evidence to support it and because they do not have to make a decision between alternatives, they do not generate evidence for the alternative answer, thus resulting in more overconfidence because no competing evidence for the alternative is considered. When people have to make a choice they weigh up the evidence for both answers, and the evidence for the rejected answer helps reduce the confidence in the finally chosen answer. Finally, Sniezek, Paese, and Switzer, proposed that very shallow processing of information is carried out when subjects do not actually have to choose any answer and this results in subjects not appreciating the difficulty of the task and thus being overconfident. They conclude that more cognitive processing reduces overconfidence because it increases accuracy and reduces confidence, and this bodes well for better calibration in serious decisions where more thought is expended.

There is contradictory evidence, though, as to whether more complex decisions, which require more thought, do in fact result in better calibration than less complex decisions and judgements. Confidence in complex decisions was found to be lower than for simpler decisions by Zakay (1985), however, Sen and Boe (1991) found overconfidence for complex tasks and underconfidence for simple tasks. Obviously other factors are influencing these results, and need untangling.

Peterson and Pitz (1988) drew a distinction between processes of generating a hypothesis, or thinking of an answer, and then evaluating that hypothesis/answer. In the first process of hypothesis generation more information leads to more uncertainty, but then when the answer has been chosen the level of confidence is assessed in a hypothesis testing process and for this process more information increases confidence.

Sniezek, Paese, and Switzer's (1990) observation that when people are given an answer to a question, less processing goes on and more overconfidence results means that a person may be highly confident in another person's judgement because they are unaware of the contradictory evidence which was dismissed by the original decision maker, and which may have tempered their confidence in their answer. People may have too much faith in other people's decisions and be too confident in their correctness, because they are unaware of all the facts leading to that decision.

When confidence is forming in an answer is it the balance of supporting versus opposing evidence which determines confidence (Koriat, Lichtenstein, & Fischhoff, 1980), or is it determined by the ease with which a potential answer comes to mind? Nelson and Narens (1990) found evidence that, regardless of correctness, faster speeds of recall positively related to confidence in answers, and also that confidence was higher for correct than incorrect answers. However, speed of recall and confidence may just be a consequence of the ease of the answers. Kelley and Lindsay (1993) investigated this possibility by manipulating the ease of recall by pre-testing with irrelevant or relevant information. They found that information which comes to mind quickly is accepted as true, even though it may be familiar but not in fact true. The speed of recall did seem to affect confidence in those answers, and incorrect but rapidly recalled answers also showed high confidence.

This indicates that answers are brought to mind and associated confidence depends upon their ease of recall; which is indicated by speed. After which the answer may be checked or not for its validity, which perhaps only occurs if the associated level of confidence is not high enough for the person to be absolutely sure of the answer. At this stage the reasons for and against the selected answer may be used to determine its validity, and confidence raised to a level where the answer becomes acceptably correct, or reduced and an alternative answer sought.

If, as Fischhoff, Slovic, and Lichtenstein (1977) proposed, people fail to appreciate that memory is a constructive process rather than just total recall, then overconfidence may be explained as due to people's lack of appreciation of the social, cognitive and motivational factors which cause errors in constructing those memories. Fischhoff et al. also pointed out that people try to infer the correct answer when they do not know an answer, and people's levels of confidence do not sufficiently allow for errors in these inferences, resulting in overconfidence.

Failure to Consider the Opposite

A problem which may underlie errors of judgement is a *failure to consider the opposite* (Lord, Lepper, & Preston, 1984) where people fail to realise that the questions that they ask influence the answers that they receive. Instructing subjects to consider opposite possibilities was found to reduce biases of social judgement on two social judgement tasks and increase impartiality. Lord, Lepper, and Preston suggested that the mechanism that may be working is that thinking of the opposite makes the opposite position an anchor and more salient to the person. This seems to be an effective method of retraining people to be less biased in social judgements.

Griffin and Tversky (1992) proposed that "people's confidence is determined by the balance of arguments for and against the competing hypotheses, with insufficient regard for the weight of the evidence" (p. 411). Thus they said that people rely too heavily on the strength or extremity of evidence (e.g. the warmth of a letter) and give insufficient consideration to the weight or credence (e.g. the credibility of the writer). This results in overconfidence when strength is high and weight low, and underconfidence when strength is low and weight high, which is what they found experimentally to be the case. People put too much emphasis on the extremity of a sample and neglect to consider it's size.

Aims of This Thesis

Many questions have been raised in this introduction, some of which this thesis will now try to answer. Chapter 1 will examine if predictions of future behaviour for oneself are better calibrated than those for a partner, the hypothesis being that they will be better calibrated because internal uncertainty is reduced for self-predictions. The literature indicates that people use base rate information insufficiently in their predictions of future behaviour. Chapter 2 studies the influence of base rates on overconfidence, across the entire range of base rates, and whether people can accurately estimate the base rate of the population for each prediction. The outcome of the event being predicted, positive or negative, as a variable to be studied is also introduced in Chapter 3, to see whether the desirability of the outcome causes unrealistic optimism in estimates of likelihood, and thus an increase in overconfidence for positive events.

The first part of this thesis, then, looks at the influence of the type of question being judged, and the person being predicted for, on confidence and overconfidence. Chapters 3 and 4 then progress to ask whether the individual characteristics of the person making the judgements influences overconfidence. Are people who have a better understanding of the rules of probability better calibrated in their judgements? Do people who commit the Fundamental Attribution Error, and fail to recognise the influence of the situation on behaviour, more or less well calibrated than those who are less biased? Are people who have different cognitive styles, such as field dependence, more or less over-confident? These questions will be addressed in relation to overconfidence in knowledge (postdictions) in Chapter 3 and in relation to predictions of one's own future behaviour in Chapter 4. Chapter 4 uses the overconfidence data from Chapter 2 and then measures individual differences and relates these to the measured overconfidence in self-predictions. Chapter 4 also investigates whether a person's self-esteem influences their overconfidence in predictions of their future behaviour.

Chapter 5 then investigates whether the type of feedback available in everyday life can make people better calibrated in their confidence. This experiment will look at confidence in postdictions, according to the level of task difficulty, using a general knowledge task. Increased confidence for consistently easy tasks, and reduced confidence for consistently difficult tasks, should be found if people can adapt their confidence to the situation when feedback is given. Finally, issues raised by the research in this thesis, applications of overconfidence, and ideas for further research will be examined in the Discussion chapter.

CHAPTER 1

Overconfidence in Predictions of Future Behaviour for Oneself and a Partner

When predicting future events, self-awareness of dispositional factors should reduce uncertainty, about one's own future in comparison with predictions for another person who is less well known. To test this hypothesis 70 subjects, in a within-subjects yoked design experiment, predicted their own and a partner's future behaviour over the coming month and the accuracy of these predictions was determined by comparison with reported outcomes one month later. Significant overconfidence was found for both self- and partner-predictions, being significantly higher for one's partner than for oneself. Less overconfidence in predictions for oneself than for one's partner was found at all levels of confidence except at around 75-85%. The difference between overconfidence in predictions for oneself and one's partner was related to the base rate of the predicted event, as predictions for one's partner were more overconfident at mid-range base rates. This supports the hypothesis that at mid base rates, where the base rate is less predictive, incorrect dispositional information about the other person is over relied on.

Overconfidence in Predictions of Future Behaviour for Oneself and a Partner

The study of judgements can ascertain whether judgements which are later found to be erroneous, are made with appropriately low levels of certainty, or if people are confident in their judgements which are incorrect. If people have shortcomings in their inferential processes then these may be reflected by too high a level of certainty in judgements which are not as accurate as expected. The measure of this discrepancy, between confidence/certainty and accuracy, is calibration, which has been extensively studied (see Lichtenstein, Fischhoff, & Phillips, 1982, for a review). To make the calibration literature more relevant to everyday life, Dunning, Griffin, Milojkovic, and Ross (1990) and Griffin, Dunning, and Ross (1990) studied overconfidence in social predictions. Social judgements, such as predicting the behaviour of a peer, which is a task of lay inference, rely upon the everyday experience of judging other people, which we all use, and which may also be prone to errors of judgement.

Dunning, Griffin, Milojkovic, and Ross (1990) asked subjects to give subjective probability estimates of the occurrence of future events for different people. They found that subjects were overconfident in these social judgement predictions, showing confidence of 75% to 80% but the corresponding accuracy of those predictions, across five different social prediction tasks, was only 60% to 68%. Overconfidence was found to be prevalent for behavioural predictions regardless of whether subjects knew the person intimately, had interviewed the person, or only heard a taped interview, and overconfidence was much greater when subjects made judgements about people whose photographs they had seen but with whom they had had no other contact. Their results indicated that people can base their judgements upon even very flimsy evidence such as facial appearance and stereotypes, and this results in the most overconfidence, because accuracy is so low. Over 20 percent of judgements which were made with complete certainty in Dunning, Griffin, Milojkovic, and Ross's experiment turned out to be wrong. Confidence and accuracy did increase together, but differentially, and confidence consistently exceeded accuracy at all levels of confidence in judgements, the gap between the two (overconfidence) was larger when confidence was higher.

Gathering information about oneself and one's personal attributes is useful in predicting one's future behaviour and can help to anticipate events and gain more control over events. However, some situations may be easier to predict than others; people may not be aware of this and so may not modify their confidence in their predictions appropriately. For example, events with high or low base rates, that is those that occur to most people or to very few, are more predictable than those with intermediate base rates, and thus people are more accurate when base rates are extreme.

Dunning, Griffin, Milojkovic, and Ross (1990) found that subjects were most overconfident when they predicted that a peer's behaviour would go against population base rates. When the base rate is very high, going against it results in massive overconfidence in social predictions, but going with it results in underconfidence. They argued that this shows that relying upon dispositional information, for example about the person's character, causes more biased judgements of other people's behaviour. Overconfidence was found even when base rate information was given to subjects, as they seemed disinclined to make use of it when making their predictions. They concluded that predictions which rely on dispositional information and go against base rate norms are especially prone to showing more confidence than accuracy and so caution should be exercised when making these types of predictions. In layman's terms: if most people don't do something but you think that because of their personality someone in particular will do it, then exercise much more caution in expressing high confidence, as it is likely to be unwarranted.

A study by Osberg and Shrauger (1986) assessed accuracy in self-predictions of future behaviour, but it did not look at the relationship between confidence and accuracy, that is, calibration. Their study is, however, very informative about the types of information people use when making predictions about their future behaviour. Five types of information were identified by these authors as important in making predictions. When asked to explain what sort of information subjects based their predictions on, 41% of reported reasons were based on personal base rate information, 29% of reasons were based on circumstances, 18% on personal dispositions, 5% because of a specific intent, and 1% was related to the population base rate. Different types of information are useful for different predictions, for example when making predictions for events with mid-range base rates, where 40-60% of the population experience the event being predicted, attending to personal dispositional information is most useful in promoting accuracy, but this is not the case for extreme base rates, where the most accurate predictions result from predicting in line with the base rate information.

It seems that people do not allow for the uncertainty present in the situation and the uncertainty about how another person views the situation, which is called *situational construal* (Ross, 1987; Ross & Nisbett, 1990). Failure to allow for situational construal may be a cause of overconfidence in social judgement, and its effects can be investigated. When making judgements about oneself, the uncertainty about one's own dispositions is reduced because we know ourselves intimately, thus overconfidence in predictions for ourselves should be less than for predictions about other people, because some of the uncertainty has been removed. Overconfidence should also be less severe for self-predictions as several studies have shown that information about oneself is recalled faster than information about other people and is also more accurate (Bower & Gilligan, 1979; Markus, 1977; Rogers, Kuiper, & Kirker, 1977). Thus people may have better access to relevant information for self-judgements than for other-judgements.

In a follow-up study to Dunning, Griffin, Milojkovic, and Ross (1990), Vallone, Griffin, Lin and Ross (1990) asked students to predict their own and their roommate's behaviour for the following months, to examine uncertainty about the actor and uncertainty about the situation. Confidence in these judgements was then compared to the actual outcomes for the predicted events. Their results backed up the findings of Dunning et al., as predictions which went against the base rate were most overconfident. They also found lower overconfidence for self-predictions (11.3%), than for predictions for a roommate (14.7%). Predictions for oneself were both more confident and more accurate than predictions for a roommate. An analysis of the items being predicted revealed that there were varying levels of overconfidence for different predictions, but that overconfidence was pervasive and not the result of one or two trick questions. Further analysis revealed that predictions that were made with low confidence were well calibrated, irrespective of whom they were made for, (with around 58% confidence and accuracy for both self- and other-predictions), but overconfidence increased as confidence increased to higher levels. High confidence in predictions was given without sufficient justification. Vallone, Griffin, Lin and Ross (1990) believed that even when people know the actor (themselves) intimately, they do not anticipate future changes in the situation, and uncertainty about the future is not weighted sufficiently to reduce confidence to appropriate levels.

Rationale for the Experiment

It seems that dispositional information is associated with higher accuracy in self-predictions (Osberg & Shrauger, 1986), and this increased accuracy is associated with increased confidence which can lead to overconfidence. It may, however, be the case that dispositional information improves accuracy for self-predictions in mid-base rate ranges, but does not increase accuracy when making judgements about other people, because we do not know them as well as we think. Thus the use of base rate information would result in better accuracy and improved calibration, lower overconfidence, for judgements about other people, and also avoid the higher confidence which is associated with the use of personal dispositional information. If this is the case, then when base rates are mid-range and dispositional information is used more than base rates, there should be increased overconfidence for partner-predictions as compared to self-predictions, because only for self-predictions is accuracy much improved by dispositional information.

The hypothesis of the experiment reported below is that the mean level of confidence will be significantly larger than the obtained level of accuracy, and that less overconfidence will be found for self-predictions than for other-person predictions, especially at mid-base-rate ranges. The use of everyday behavioural predictions, which also have been shown to result in overconfidence, counters any argument that overconfidence is only found when the task is unfamiliar. These sorts of predictions also have implications for real life, as overconfidence in personal and social predictions affects how people plan their lives.

Method

Subjects

Seventy first-year psychology undergraduate students participated in the experiment as part of their course requirement. There were 55 females and 15 males, with an average age of 19.5 years (range 17 to 36 years). The subjects volunteered as a pair, with a friend or partner, resulting in 35 pairs (2 male-only pairs, 11 mixed-sex pairs, and 22 female-only pairs).

Design

A within-subjects yoked design was used, the subjects responded to all 20 questions three times, predicting their own behaviour, their partner's behaviour and then, one month later, reporting whether or not the event occurred to themselves. The order of presentation of the first two questionnaires was counterbalanced.

Materials

Three questionnaires were designed. A pilot study was used to generate, select and check the wording of 20 questions, from an initial pool of 30 student-specific items, to be used in the main study. A group of 20 pilot study subjects were each asked to write down a list of events that had happened to them in the last week. A list of 30 activities was then compiled from the pool of events generated by the pilot group. The events were then rephrased into questions to refer to the following month. For example, an event such as missing lectures was changed so that the question asked if the subject would miss more than four lectures in the next month. Other questions reflected social, academic and recreational events, such as: feeling homesick; going to a concert; playing sport; having a hangover.

Each of the 30 questions, based on these events, was read out to the pilot group and they answered the questions using a scale of 0% to 100% ("I am totally sure that I will not do this in the next month", to "I am totally sure that I will do this within the next month"). The pilot group was used to check the clarity of instructions and to find out which questions were not clear or had ambiguous answers. The questions were in some cases revised slightly to amend ambiguities, using the feedback from the subjects, and 20 that showed the least ceiling/floor effects were used in the main study (all questionnaires used in this experiment are presented in Appendix A).

Procedure

The following experiment was based upon the study by Vallone, Griffin, Lin, and Ross (1990), but used a full range scale rather than a two option, half-range scale format. The main study consisted of three versions of the same 20 questions, but with instructions for subjects to predict their own and their partner's behaviour for the following month and then, one month later, to report their own behaviour over the month in question.

Subjects in the main study were recruited in pairs of friends/partners and were tested in group situations ranging from groups of 2 to 12 people. The subjects were required to answer the first two questionnaires, which asked them to predict their own and their partner's behaviour for the following month. All instructions to subjects can be seen in

Appendix A. For each of the twenty events to be predicted, subjects were asked to give a rating using any number between 0 and 100 on the following scale of certainty (see Table 1.1).

Table 1.1

Rating Scale for Levels of Certainty

I am totally sure that **I will not** do this within the next month		(unsure 50:50) Equal probability		I am totally sure that **I will do** this within the next month
0%	- - - - - - - - - -	50%	- - - - - - - - - - -	100%

Questionnaire 2 asked the subject to rate the same questions, but for their partner, instead of for themselves, with the rating scale appropriately modified to say he/she instead of I. Partners' names, age and sex were also recorded. Subjects rated their relationship with their partner using the scale shown in Table 1.2. The third questionnaire was answered a month later and asked subjects to answer yes or no as to whether each of the events had in fact occurred to them since the first two questionnaires were administered.

Table 1.2

Rating Scale for Friendship Level

Unknown	Acquaintance			Friend	Good friend			Intimate	
1	2	3	4	5	6	7	8	9	10

The predictions for self and one's partner were compared with the actual outcomes reported by each individual subject and their partner, to determine their accuracy. Ratings of less than 50% were classed as a negative prediction, that the event would not occur, and above 50% a positive prediction, that it would occur. Predictions which corresponded to the actual outcome were classed as hits and were given a score of one and incorrect predictions scored zero. Ratings of 50% were classed as no prediction, a 50:50 chance of the event occurring or not, thus no comparison with reported answers could be made, and no accuracy score determined. The mean scores for individuals and question totals were converted to percentage scores. The confidence levels were calculated by recoding the scores of 49% and below, so that a rating of 0% (complete certainty that the event would not occur) on the scale equalled 100% confidence and 49% equals 51% confidence (see Table 1.3).

Table 1.3

Recoding Scale to Change Raw Scores to Confidence Levels

Rating	0	10	20	30	40	50	60	70	80	90	100
Confidence	100	90	80	70	60	50	60	70	80	90	100

The mean accuracy scores were compared with the mean confidence levels for each subject and also for each question using t-tests, and correlations between them were also made. Throughout this thesis between-subjects t-tests are used to test for sex differences and within-subjects t-tests are used to compare confidence, accuracy and all other dependent variables, except where otherwise stated. The mean overconfidence for each subject and each question was calculated as mean confidence minus mean accuracy. The base rate for each event was calculated as the mean number of subjects that reported that the event had occurred within the month in question. All subsequently reported means for confidence, accuracy, overconfidence, and base rate are percentages.

Results

The mean friendship rating was 7.3, modal rating 8, indicating that subjects were on average good friends, range from 3 to 10. The friendship variable did not significantly correlate with any other variable. A manipulation check proved that there was no significant effect of the order of completing the first two questionnaires on confidence, accuracy, or overconfidence.

Unfortunately the use of a full-range scale to indicate the degree of certainty that an event will or will not occur, precludes the assessment of an accuracy score for those judgements made with 50% certainty. The use of the 50% rating abdicates making a judgement in either direction, and thus there is no prediction to compare with the outcome. This could be overcome in future experiments by asking for a yes/no decision before using a rating scale to indicate confidence in that answer. Judgements made with 50% certainty were removed from subsequent analyses which require comparisons with accuracy scores.

Analysis by Subjects

Table 1.4

Mean Confidence, Accuracy and Overconfidence for Self and Other Over All Questions

	Self	Other	*t*	*r*
Confidence	89.32	85.65	6.36***	.75***
Accuracy	78.82	71.79	3.80***	.05
Overconfidence	10.51	13.86	1.79*	.18

Note. All statistical comparisons are between confidence and accuracy, over all questions combined, the difference between which equals overconfidence.

$df = 69$ * $p < .05$ ** $p < .01$ *** $p < .001$

Table 1.4 shows that subjects were significantly more confident for predictions about themselves than about their partner ($M = 89.32$ vs. $M = 85.65$), $t\ (69) = 6.36$, $p < .001$ (all t-tests reported in this thesis are two-tailed unless otherwise stated). Correspondingly they were also less accurate in their predictions about their partner than for themselves ($M = 71.79$ vs. $M = 78.82$), $t\ (69) = 3.80$, $p < .001$. Subjects were significantly more confident than they were accurate for both self-predictions ($M = 89.32$ vs. $M = 78.82$), $t\ (69) = 7.47$, $p < .001$, and partner-predictions ($M = 85.65$ vs. $M = 71.79$), $t\ (69) = 9.17$, $p < .001$. Thus significant overconfidence was found for predictions for both oneself and one's partner ($M = 10.51$ vs. $M = 13.86$) (see Table 1.4). Out of 70 subjects 60 (85.7%) were, on average, overconfident in self-predictions and the same percentage were overconfident for predictions for their partner. The predicted main effect of significantly higher overconfidence for one's partner than for oneself was supported, as mean overconfidence for oneself was 10.51% and for one’s partner 13.86%, $t(69) = 1.79$, $p < .05$ (one-tailed). A one-tailed test was used in this case because a clearly directional hypothesis had been stated.

The relationships between overall levels of confidence, accuracy and overconfidence for each subject were examined using Pearson correlation coefficients and these results are shown in Table 1.4. The mean level of confidence for oneself and one’s partner was highly correlated, $r(69) = .75$, $p < .001$, showing that subjects who were more confident about predictions for themselves were also more confident in their predictions for their partner, whereas no significant relationship existed between being accurate for oneself and being accurate in predictions for one’s partner, $r(69) = .05$, $p > .05$, showing that subjects who were good predictors for themselves were not necessarily also good at predicting their partner’s behaviour. It was also found that subjects who were overconfident in self-predictions were not necessarily also overconfident for partner-predictions, as there was no significant correlation between the two, $r(69) = .18$, $p > .05$. Mean confidence and accuracy were not significantly related for oneself, $r(69) = .14$, $p > .05$, and for one’s partner, $r(69) = .15$, $p > .05$, which shows that more confident subjects were not necessarily more or less accurate than less confident subjects.

Other factors influencing overconfidence were also considered. The gender of the subject was found to have a significant effect upon the level of confidence for judgements for both oneself and one’s partner. Male subjects expressed less confidence ($M = 80.92$) than females ($M = 86.95$) when making predictions about their partner, $t(68) = 3.02$, $p < .01$. There was no difference between the sexes in overconfidence though, because accuracy for partner-predictions was also slightly, but non-significantly, higher for females than for males ($M = 72.93$ vs. $M = 67.62$). There was no significant difference between male and female subjects in predictions for oneself, which were made with around 89% confidence. Male subjects also experienced significantly more of the events, the base rate was 45.00% for male subjects and 37.00% for female subjects, $t(68) = 2.40$, $p < .05$.

Analysis by Predictions

An analysis was made to see exactly which predictions were causing the significant levels of overconfidence. Table 1.5 shows a breakdown of how many judgements were made with different levels of confidence, and how accurate they were on average. These data are then shown graphically in Figure 1.2 and Figure 1.3.

Table 1.5

Percentage Accuracy and Overconfidence by Levels of Confidence in Predictions for Oneself and Partner

	Partner-Predictions			Self-Predictions		
Confidence	Accuracy	Over-confidence	Total *N*	Accuracy	Over-confidence	Total *N*
50	-	-	190	-	-	216
55	52.94	2.06	17	57.14	-2.14	7
60	56.41	3.59	117	62.16	-2.16	74
65	55.00	10.00	20	72.73	-7.73	11
70	64.39	5.61	132	68.48	1.52	92
75	68.06	6.94	72	60.00	15.00	60
80	74.81	5.19	135	69.01	10.99	142
85	76.92	8.08	26	63.64	21.36	22
90	72.86	17.14	199	76.80	13.20	181
95	67.21	27.79	61	81.25	13.75	64
100	80.10	19.90	407	88.85	11.15	556
	71.79	13.86	1399	78.82	10.51	1402

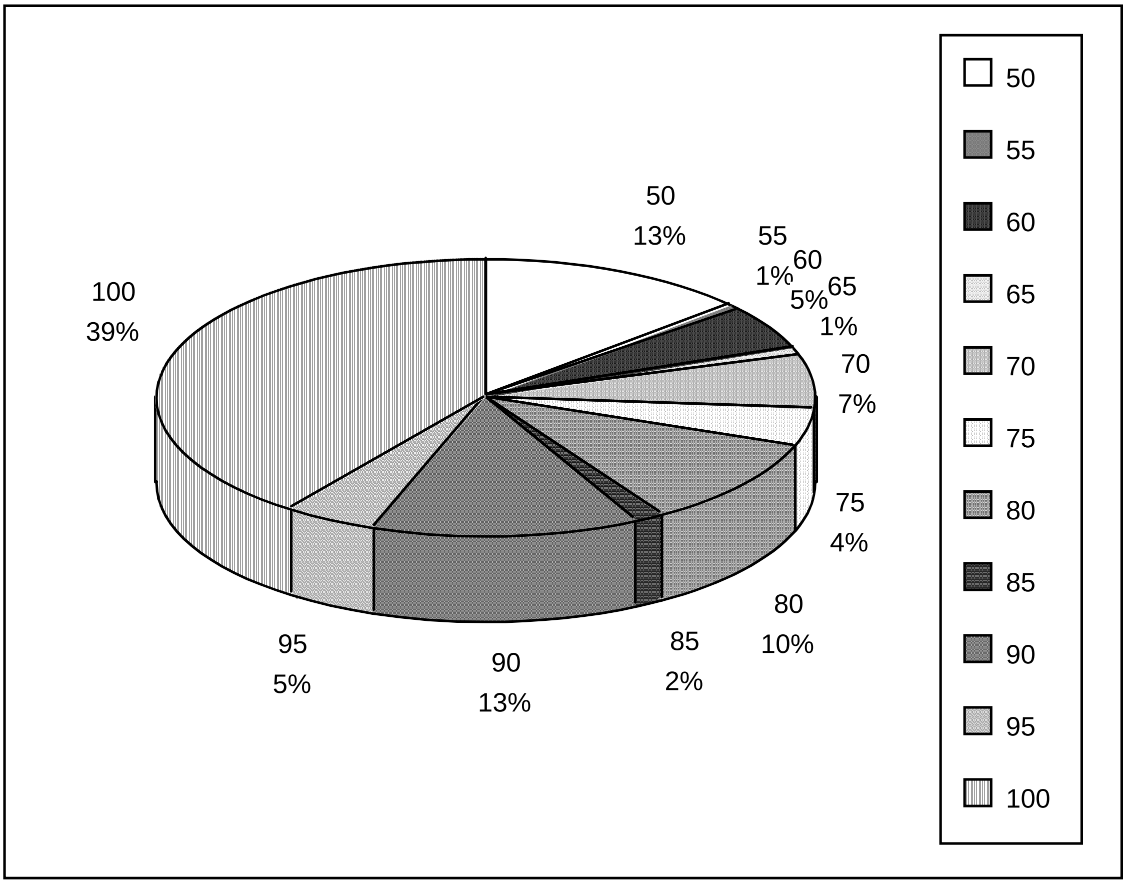

Figure 1.1. Proportions of judgements made with different levels of confidence.

Figure 1.1 shows how many of the total number of predictions were made with each confidence level. This chart shows that a large proportion of the predictions (39%) were made with 100 percent confidence, showing that subjects were happy to assign total confidence to a large number of predictions. Subjects were extremely confident in their predictions, expressing 90 percent or more confidence on 57 percent of all judgements. This chart shows that subjects obviously preferred to use ratings such as 70 and 90 in preference to 55, 65 and 85, as they may have found it easier to think on an 11 point scale than a 21 point one, which 5% intervals would entail.

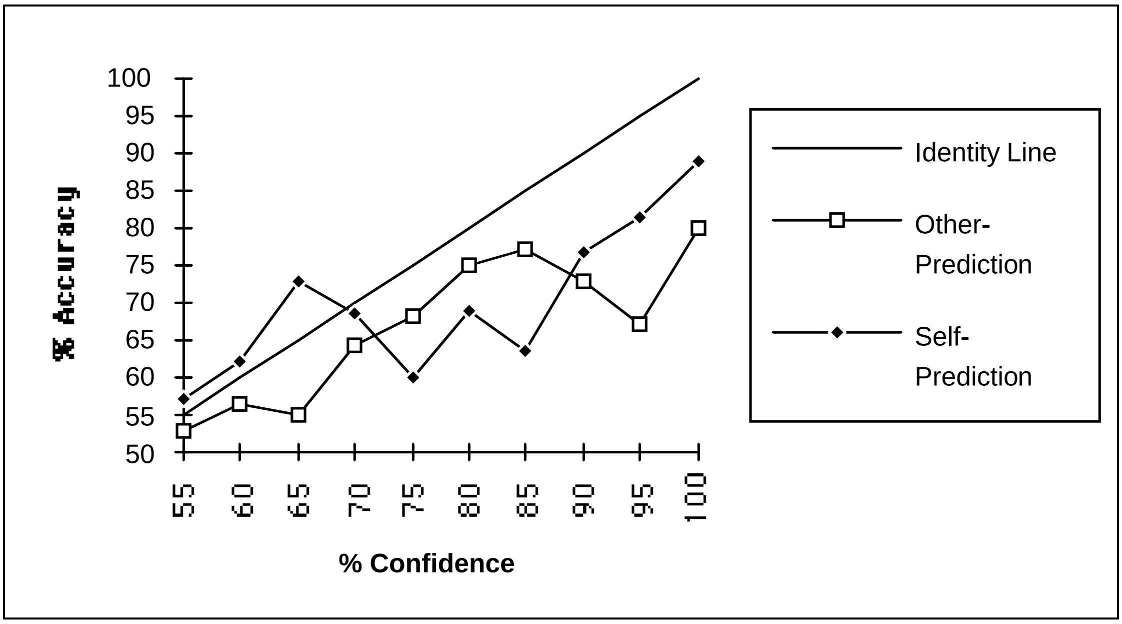

Figure 1.2. Calibration curves for predictions for oneself and one's partner by confidence levels.

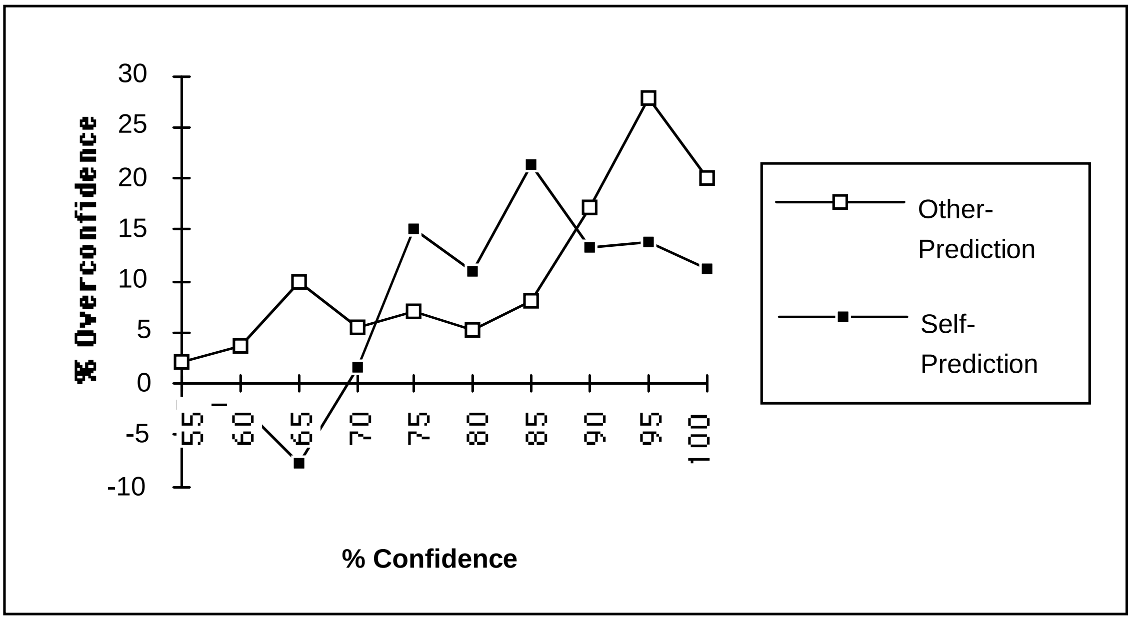

Figure 1.3. Overconfidence in predictions for oneself and one's partner for different confidence levels.

An inspection of the data in Table 1.5 and the calibration curves for predictions for oneself and for one's partner (Figure 1.2 and Figure 1.3), reveals that subjects were overconfident in their judgements for their partner at all levels of confidence. When predictions about one's own future behaviour were made with less than 70% confidence, good calibration was achieved and even slight underconfidence was found.

Overconfidence was higher for judgements about oneself than about one's partner when confidence estimates of 75% to 85% are used, but overconfidence is reversed and is very much higher for one's partner than for oneself when very high or low confidence is stated. Indeed the conclusion that seems apparent from Figure 1.3 is that when making judgements about one's partner most overconfidence occurs with 90% confidence or more, whereas for oneself the most overconfidence is obtained when judgements are made with above 75% confidence.

Analysis by Questions

Overconfidence is apparently highest when confidence is very high, but which questions provoke such high and unwarranted confidence? An analysis of overconfidence for each question is necessary to determine whether all predictions are made with equal overconfidence or whether some are more difficult to predict with subjects being unaware of that, and are thus more prone to overconfidence. Tables 1.6 and 1.7 show that the level of overconfidence did vary quite a lot between questions, ranging from underconfidence for Question 11 to a peak of 28.81% for Question 13 when making predictions for one's partner.

The amount of overconfidence varied between questions, 45% of the self-predictions were between 1 and 9% overconfident, 45% being between 10% and 20% overconfident, 5% exceeding 20 %, and 5% showing underconfidence. The predictions for one's partner were more overconfident with 20% from 1-9% overconfident, 45% being 10-20% overconfident, 10% being underconfident by up to 6% and 25% being over 20% overconfident. Eleven of the questions showed significant overconfidence for self-predictions and 14 showed overconfidence when the predictions were for one's partner. Only four questions showed a significant correlation between confidence and accuracy for predictions for one's partner, but eight were significant for oneself. Over all subjects, the relationship between confidence and accuracy was significant ($r = .59$ $p < .01$ for oneself and $r = .61$ $p < .01$ for one's partner), showing that those questions that were high in accuracy were also those with more confidence shown in them. Questions about which subjects were overconfident for themselves also tended to be the ones which they were overconfident about for their partner, $r = .76$, $p < .001$.

Table 1.6

Confidence, Accuracy and Overconfidence for Self-Prediction Questions

Qn. No.	Mean Self-Predictions Confidence	Accuracy	Over-confidence	*df*	*t*	*r*
1	92.00	83.08	8.92	64	2.03*	.36**
2	87.60	80.77	6.83	51	1.28	.28*
3	93.05	74.58	18.47	58	3.43***	.36**
4	90.98	83.61	7.38	60	1.66	.37**
5	89.82	85.45	4.36	54	0.89	.10
6	86.78	64.41	22.37	58	3.46***	.01
7	86.31	84.62	1.69	64	0.36	.03
8	86.85	80.65	6.21	61	1.17	.04
9	91.58	84.21	7.37	56	1.64	.39**
10	83.10	75.86	7.24	57	1.31	.24
11	93.28	95.52	-2.24	66	0.87	.22
12	90.08	78.13	11.95	63	2.42*	.32**
13	85.34	74.14	11.21	57	1.98*	.23
14	83.85	67.21	16.64	60	2.77**	.18
15	91.42	78.33	13.08	59	2.55**	.29*
16	91.29	77.27	14.02	65	2.64**	.07
17	93.06	85.48	7.58	61	1.60	-.01
18	88.00	76.67	11.33	59	2.12*	.25
19	85.69	68.97	16.72	57	3.05**	.47***
20	90.00	76.67	13.33	59	2.49*	.25
Total	89.00	78.78	10.22	19	7.76***	.59**

Note. All statistical comparisons are between confidence and accuracy for each question, the difference between which equals overconfidence.

* $p < .05$ ** $p < .01$ *** $p < .001$ (two-tailed)

Questions which subjects were more accurate in resulted in the least overconfidence, $r(19) = -.90$, $p < .01$ for self-predictions and for partner-predictions, $r(19) = -.95$, $p < .01$. Thus this shows a strong task difficulty effect. The difficulty of the task accounted for 81-90% of the variance in overconfidence.

Table 1.7

Confidence, Accuracy and Overconfidence for Other-Prediction Questions

Qn. No.	Mean Other-Predictions: Confidence	Accuracy	Over-confidence	*df*	*t*	*r*
1	86.53	74.58	11.95	58	2.16*	.26*
2	83.68	73.58	10.09	52	1.51	-.15
3	87.63	64.91	22.72	56	3.48***	.06
4	87.03	68.75	18.28	63	3.19**	.21
5	87.60	84.00	3.60	49	0.79	.50***
6	84.74	67.24	17.50	57	2.73**	.03
7	83.31	87.10	-3.79	61	0.94	.35**
8	83.28	60.66	22.62	60	3.49***	.04
9	84.72	73.58	11.13	52	1.76	.07
10	79.82	73.21	6.61	55	1.10	.14
11	91.09	96.88	-5.78	63	2.71**	.37**
12	83.08	63.33	19.75	59	3.16**	.17
13	77.97	49.15	28.81	58	4.15***	-.06
14	81.10	61.02	20.08	58	3.18**	.20
15	87.50	75.00	12.50	55	2.13*	.14
16	89.56	69.12	20.44	67	3.70***	.21
17	91.77	82.26	9.52	61	1.96*	.18
18	84.77	77.27	7.50	65	1.44	.14
19	85.93	68.52	17.41	53	2.73**	.14
20	86.98	69.84	17.14	62	2.83**	.04
Total	85.40	71.87	13.54	19	6.85***	.61**

Note. All statistical comparisons are between confidence and accuracy for each question, the difference between which equals overconfidence.

* $p < .05$ ** $p < .01$ *** $p < .001$ (two-tailed)

That different questions show different amounts of overconfidence is therefore apparent, but what factors distinguish these questions from one another? One factor which may be influential is the base rate of the question, and this was studied in this experiment. The mean level of overconfidence for each question for predictions for oneself and for one's partner are shown in Figure 1.4 according to the base rate of the question.

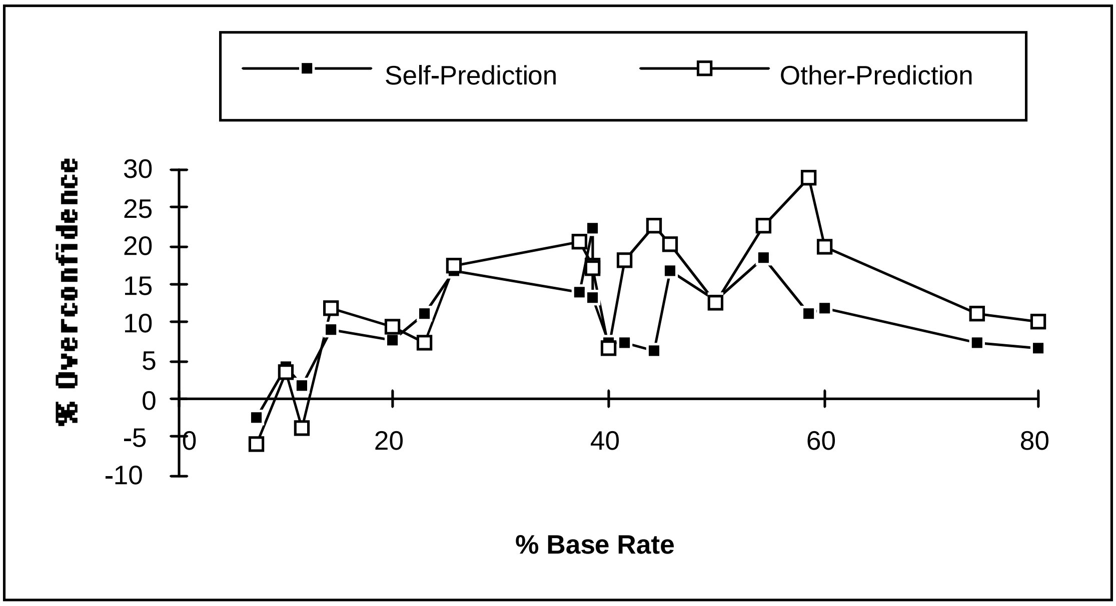

Figure 1.4. Overconfidence for oneself and one's partner according to the base rate of the question.

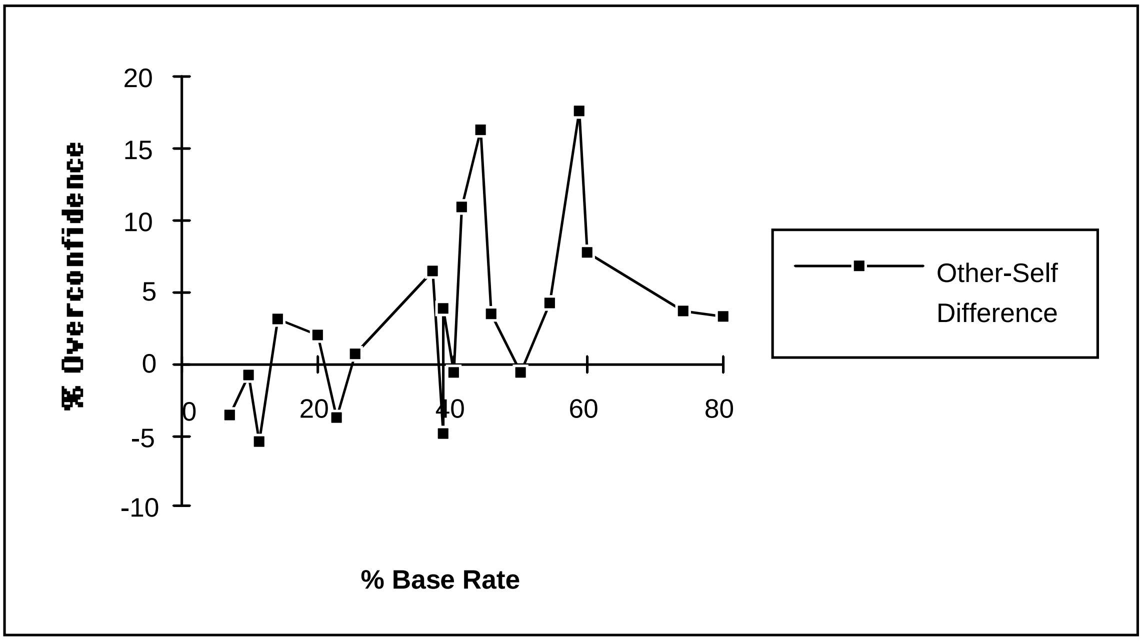

Figure 1.5. Difference between overconfidence for oneself and one's partner according to the base rate of the question.

To simplify the data graphed in Figure 1.4 and to aid understanding, Figure 1.5 shows the difference between the mean overconfidence found for oneself and for one's partner. It appears from Figure 1.4 that overconfidence is highest at the middle of the base rate range and this is also where the predictions for the other person become less well calibrated than for oneself. It is also apparent that predictions for one's partner are less overconfident than for oneself when the base rate is low and overconfidence is low. Overall, the best calibrated predictions were for Questions 5, 7 and 11 and these were the ones with the lowest base rates; they occurred to very few people, and were thus easily predicted and high in accuracy. Subjects were more accurate than confident, underconfident, when making predictions for Question 11, which asked if they would take part in an organized protest. This question had a very low base rate and subjects were very certain and confident that it would not occur, but they were not confident enough, as their predictions were highly accurate.

The largest discrepancies in overconfidence between self- and partner-predictions were for Questions 4, 8, 12, and 13 which had above average base rates, they occurred to a large number of people, and these predictions produced moderate overconfidence for self-predictions but very much higher overconfidence for predictions for one's partner. For example, Question 8 asked if subjects would feel homesick, and this was quite well predicted for oneself (81% accuracy) but poorly predicted for one's partner (61% accuracy). Similarly, the lowest accuracy was obtained for Question 13 where subjects had around chance levels of accurately predicting if their partner would eat out at a restaurant (49% accuracy), and out of all questions subjects were least confident about this prediction (78% confidence), but they were still too confident in comparison with this poor level of accuracy. For both Questions 8 and 13 much less overconfidence was found for predictions for oneself because they were more accurately predicted for oneself.

Discussion

A significant overall level of overconfidence was found in this experiment for both self- and partner-predictions. Over all questions self-predictions were less overconfident than those for one's partner, and differences were found when the base rate and the expressed confidence were considered. The subjects in this experiment were more overconfident for their partner's behaviour than for their own when they were extremely confident or not at all confident. However, when they were moderately confident in their predictions (75-85%) subjects were better calibrated, less overconfident, for predictions for their partner than they were for themselves.

There was a strong tendency for people to use very high confidence ratings, and this is where the highest overconfidence occurred -- corresponding levels of accuracy did not match up to the stated confidence. For example, when subjects were 100% confident in their predictions for themselves they were accurate in only 88.85% of those predictions, and 80.10% accurate in those predictions made with 100% confidence for their partner. These obtained accuracy rates are similar to the results of Dunning, Griffin, Milojkovik and Ross (1990), who also found about 20 percent of judgements made with complete certainty were in fact incorrect. It is evident from this experiment that expressing complete confidence in a prediction is often unwarranted -- as shown by high overconfidence for these judgements -- and this is especially the case when expressing complete confidence for predictions of other people's behaviour.

The largest discrepancies between self- and partner-predictions occurred for Questions 4, 8, 12, and 13 where predictions for one's partner showed more overconfidence than for oneself for all of these items. This was due to much lower levels of accuracy for other-person predictions than for oneself for these questions, predicting for another person was harder than subjects realized. Levels of confidence for these questions were not higher or lower than for the rest of the items, and neither was accuracy for self-predictions. This is backed up by the fact that for predictions for one's partner for these items, the levels of confidence did not significantly correlate with accuracy for each item. The results of this experiment are closely in line with those of Vallone et al. (1990), showing around one percent less overconfidence than they found for both self- and partner-predictions.

Figure 1.3 tends to point to the fact that subjects may have realised that their predictions for other people were less accurate, and their accuracy in those judgements made with 75-85 percent confidence for their partner turned out to be higher than when that level of confidence was expressed for themselves. Subjects also tended to use the 60-70 percent ratings more for partner-predictions than for self-predictions, but also used the 50 percent rating slightly less. Underconfidence was found when subjects were low in confidence in their self-predictions, indicating that when they were uncertain they were very uncertain. Subjects were less confident about making predictions for their partner, but still too confident.

It was hypothesised that much higher overconfidence would be shown for predictions for one's partner than for oneself when the base rate was mid range, because of the increased use of dispositional information rather than base rate information, which is less easy to determine at mid ranges. When dispositional information is used it is more accurate for oneself because the uncertainty about those dispositional factors is low, due to self-knowledge. The results of this experiment tend to support this hypothesis, as Figure 1.4 shows that much higher overconfidence was found for partner-predictions at mid base rate ranges than at low base rates.

It is also apparent that the reverse is true when the base rate is low; predictions for oneself are more overconfident than for one's partner. This may be because the base rate is being used much more for these predictions for the partner but dispositional information is used more for oneself, which leads to overconfidence because reliance upon dispositional information causes the impact of situational factors to be underestimated. Overconfidence may also be higher for oneself at low base rates because one can think of reasons why an event may occur to oneself, especially if it is desirable, but one may not be motivated to be optimistic for other people. It is possible that people use dispositional information for themselves at all base rates, which causes overconfidence due to failure to consider the situation. For predictions for other people though, when base rates are predictive (extreme) less reliance is placed on dispositional factors and so less overconfidence is found. The extreme base rates of over 80 percent were not covered in this experiment, and further research should include items with very high base rates.

The overall mean difference in overconfidence between predictions for oneself and one's partner was significant in this experiment. It may be the case that since the partner who was being predicted for was on average a good friend of, and well known to, the subject, that using dispositional information about them was useful and helped to reduced overconfidence. Although in this experiment no significant relationship between the friendship rating and overconfidence was found, it is possible that with a wider range of intimacy levels, such as close relationships with parents/children versus less close relationships with friends, relationships between intimacy level and the self-other difference in overconfidence may be found.

It is also interesting to note that being accurate in predictions of one's own behaviour was not significantly related to the ability to accurately predict one's partner's behaviour. This indicates that good self-insight is not automatically accompanied by good insight into other people's future behaviour. If subjects tended to be confident about their own behaviour they also tended to be confident about predictions of their partner's behaviour ($r = .75$), indicating a lack of flexibility in confidence and general tendency for predictions to be made with high or low confidence by people. In conclusion, when expressing confidence in predictions of other people's behaviour, one should temper high confidence, as it is likely to be proved somewhat unjustified, especially when the base rate is mid range and dispositional information has to be relied upon.

CHAPTER 2

Overconfidence, Base Rates, and Outcome Positivity/Negativity of Predicted Events

Overconfidence is said to occur when a person's confidence in a series of predictions exceeds the level of accuracy achieved. In this experiment, questionnaire items requiring predictions of personal life events were selected according to their objective base rates of occurrence and their outcome positivity/negativity for subjects in a pilot study. The 98 subjects in the main experiment predicted whether they would experience each event within the next week and rated their confidence in their predictions. Predictions were compared with responses to a follow-up questionnaire a week later to determine accuracy. Significant overconfidence occurred, and it was greater for positive-outcome than negative-outcome items. The results revealed a curvilinear relationship between base rates and overconfidence, with maximum overconfidence at intermediate base rate levels and underconfidence at both extremes of the base rate range. Subjects tended to overestimate base rates below 40 percent and to underestimate higher base rates.

Overconfidence, Base Rates, and Outcome Positivity/Negativity of Predicted Events

Evidence has shown that people tend to be overconfident in predicting both the behaviour of other people (Dunning, Griffin, Milojkovic, & Ross, 1990) and their own behaviour (Vallone, Griffin, Lin, & Ross, 1990). Across a wide range of social and self-predictions, subjective confidence has consistently exceeded the objective accuracy of predictions, especially when confidence was relatively high. This miscalibration of confidence and accuracy in judgements was first reported by Adams and Adams (1961) and has more recently been called the *overconfidence effect*. The findings associated with the effect suggest that the implicit psychological theories and other forms of intuitive psychology on which people base their social and self-judgements (Nisbett & Ross, 1980; Ross, 1977) may have less predictive validity than they generally assume, and that people's confidence estimates are especially likely to err when they are highly confident.

For any action that might be performed, or any other event that might be experienced by an individual, there is an objective base rate of occurrence defined by the relative frequency with which members of the relevant population experience the event in a specified time period. Dunning et al. (1990) found that when subjects made predictions contrary to the base rate, especially that a person would behave atypically, the predictions tended to be more inaccurate and also overconfident. By failing to give due weight to objective base rates and instead grounding their predictions on supposed dispositional factors internal to the actors, the subjects made errors in predicting future behaviour and displayed overconfidence in their predictions. If this interpretation is correct, then these errors in predicting *future* behaviour bear a family resemblance to the *overattribution bias*, first reported by Jones and Harris (1967) and subsequently confirmed by a number of researchers (e.g., Gilbert & Jones, 1986; Johnson, Jemmott, & Pettigrew, 1984; Ross, Amabile, & Steinmetz, 1977; Webster, 1993), according to which people tend to overestimate the causal role of dispositional relative to situational factors in explaining an actor's *past* behaviour.

In the case of self-predictions, the individual has maximum knowledge of the person and their dispositions, yet prediction errors still occur, arguably due to faulty inferences about the situation and its effects upon oneself. Vallone, Griffin, Lin, and Ross (1990) proposed that the overconfidence effect results from "insufficient allowance for the possibility of this erroneous inference or misconstrual" (p. 583). Predictions of self-behaviour that go against the norm or base rate still tend to lead to low accuracy and overconfidence, even though dispositional knowledge about oneself is more accurate than dispositional knowledge about other people.

By definition, overconfidence arises either when confidence is unrealistically high for a given level of accuracy, or what amounts to virtually the same thing, when accuracy is unrealistically low for a given level of confidence. Can the phenomenon be explained entirely by a tendency for subjects to make insufficient use of base rate information? If indiscriminate neglect of base rate information were the sole explanation, then what would be the relation between base rates and overconfidence? If subjects take full account of base rates, then at intermediate base rate levels, where prediction is relatively difficult because the prior probabilities of occurrence and non-occurrence of events are roughly equal, both accuracy and confidence should be relatively low. In regions of extremely low or extremely high base rates, where prediction is relatively easy because the prior probabilities of the event are close to either zero or unity, both accuracy and confidence should be relatively high if subjects make full use of the base rate information. Thus if subjects take full account of base rates, both their accuracy and their confidence levels should vary in parallel across the base rate range, and therefore no systematic relationship should exist between base rates and overconfidence. If, on the other hand, subjects neglect base rate information entirely, then neither accuracy nor confidence should be higher in the extreme than the intermediate base rate regions, and once again there should be no systematic covariation of base rates and overconfidence.

There is a third, more plausible, possibility, namely that base rates influence accuracy and confidence differentially, in which case base rates and overconfidence may turn out to be functionally related. The most natural assumptions would be (a) that people generally make at least some use of base rate information in formulating predictions, so that their predictions of events with extreme base rates tend to be relatively accurate compared to their predictions of events with intermediate base rates; and (b) that people generally take less account of base rates in modulating their feelings of confidence in their predictions. Although these assumptions have not been specifically tested, both are in line with closely related evidence in the field of decision making (e.g., Ajzen, 1977; Bar-Hillel, 1980; Tversky & Kahneman, 1980). If both assumptions are correct and applicable to situations in which the overconfidence effect occurs, then overconfidence should be greatest for events whose base rates of occurrence are in the intermediate range, and it should be relatively small for events with extreme base rates.

The overconfidence effect may thus be attributable at least partly to people's failure to make use of base rate information in modulating the confidence of their predictions. Another judgemental bias which has been reported is a tendency for predictions of future events to be unrealistically optimistic. People tend to over- or under-estimate the probability or the likelihood of the occurrence of events, and this varies according to whether the prediction is for positive- or negative-outcome events (Weinstein, 1980). The following experiment will examine whether such optimism is reflected in confidence judgements as well as in predictions of likelihood, and relate this to the base rate and the outcome positivity/negativity of the event. When testing for unrealistic optimism or overconfidence it is very difficult to see if an individual is correct about their long term predictions. This would involve a longitudinal study of one person over many years to see if their predictions come true (for example "I have a less than average chance of dying of cancer"). Instead the presence or absence of a bias can be measured over a group of people. For example, if the probability is one in a thousand (.001%) that a person in a population will be involved in a road traffic accident within a given time span, and within that population a sample state that their chances are on average one in ten thousand (.0001%) then unrealistic optimism is operating - as people are generally underestimating the likelihood of the event occurring. The analysis used in this chapter will be for the group as a whole, and since short-term judgements are used, each individual's accuracy will be measured and individual differences in unrealistic optimism will then be analysed in Chapter 4.

The effect of unrealistic optimism manifests itself in two ways: people underestimate the chances of negative (undesirable) events happening and overestimate the chances of positive (desirable) events. Motivational theories of optimism predict that people desire good things to happen and that this is wishful thinking, conversely they are defensive and thus want to avoid bad things. The opposing theory of optimism is a cognitive one which proposes that people may be unrealistically optimistic because of problems and biases in handling information about the event. For example, it may be easier to imagine the occurrence of an event rather than its non-occurrence. Thus people may not have enough information or may be biased in its usage which causes errors. Evidence from research on the fundamental attribution error (Jones & Nisbett, 1971; Ross et al,. 1977) shows that people find it hard to take another person's perspective, so in this context unrealistic optimism may be caused by not realising that other people share common factors with us and we are not uniquely different. Therefore our perception of the probability of an event for a population (the base rate) may be biased.

Unrealistic optimism has been found in relation to predictions about accidents (McKenna, 1993), automobile accidents (Robertson, 1977), disease (Harris & Guten, 1979; Kirscht, Haefner, Kegeles, & Rosenstock, 1966), health and safety (Weinstein, 1984), and crime (Weinstein, 1977). Weinstein (1980) compared subjects' levels of unrealistic optimism for positive- and negative-outcome events. He found unrealistic optimism for both types of events: people over-predicted the likelihood of positive events happening to them and underestimated the chances of negative events occurring. He did not however make any comparison with a control group of neutral events. Weinstein also proposed that the degree of desirability, perceived probability, personal experience, perceived controllability and stereotype salience would effect the level of unrealistic optimism expressed in a judgement. Having personal experiences of an event should make recall of factors leading to the event more easily recalled. This is the availability hypothesis, as demonstrated by Tversky and Kahneman (1973; 1974), and may result in a greater perceived probability (i.e. subjects may estimate the base rate to be higher that it really is).

Weinstein (1980) found that the amount of control people believe that they have over an event happening or not also influences whether people are optimistic. People can bring to mind actions to control events more easily in relation to themselves than for other people and so feel more certain that they can intervene in events and lower or raise their probability of events occurring. It is assumed that people imagine actions that will promote their chances of achieving or avoiding the event, rather than thinking of things they will not do. A cognitive explanation would say that it is easier to bring to mind events rather than non-events whereas a motivational explanation would say that selective recall of controlling actions is reassuring and ego-defensive. Thus the more control a person believes they have the more unrealistically optimistic they will be. Weinstein also looked at the stereotype salience for each event. If a stereotype readily comes to mind of the type of person who the event would happen to and that stereotype differs from us in its salient features then the representativeness heuristic would predict that we decide that we do not fit the stereotype and so the event is less likely to happen to us. Of course the stereotype may be wrong, thus causing inaccurate estimates of the base rate and our chances.

With regard to these factors Weinstein (1980) found that the positivity or negativity of the events had a powerful and independent effect, regardless of the other event characteristics previously mentioned. For positive events the main factor was perceived probability and for negative ones it was stereotype salience. In conclusion, for both positive and negative events unrealistic optimism occurs when the event is seen as controllable, and when there is an emotional investment in the outcome. Weinstein proposed that "under these conditions, optimism arises because people compare themselves with an inappropriate standard: a person who does little or nothing to improve his or her prospects" (p. 814). Weinstein predicted, but did not find, more optimism for events with strongly negative outcomes than milder outcomes.

An attempt to reduce unrealistic optimism in predictions was made by Weinstein (1980), by asking subjects to generate reasons which might influence whether the event will happen to them or not. This succeeded in reducing the level of unrealistic optimism for the positive events but not the negative ones. Then another group of subjects were asked to read the reasons listed by the first group, and the amount of optimism dropped for both positive and negative items compared to a control group who neither generated nor read reasons. The levels of unrealistic optimism were not totally eliminated however, and this may point to there being motivational factors at work. Weinstein thus concludes that the generation of mental lists of reasons is a crucial stage in evaluating if our chances are greater or less than average. Similar results were found by Koriat, Lichtenstein, and Fischhoff (1980) when they manipulated the generation of pro and con reasons to reduce overconfidence.

The principal aim of the experiment reported in this chapter is to examine the base rate-overconfidence relationship. A secondary aim is to clarify the underlying mechanism of the overconfidence effect by determining whether overconfident subjects are able to estimate base rates accurately but are unable to make adequate use of them, or whether they are simply unable to estimate the base rates in the first place.

Also, in the light of the well-established Pollyanna effect (Boucher & Osgood, 1969) according to which people tend to exaggerate the positive and minimize the negative aspects of their experience, which is akin to unrealistic optimism, a third aim of the experiment is to test the hypothesis that overconfidence is greater for events with positive, desirable outcomes than for negative-outcome events. The effect of base rate will be held constant for the comparisons of positive and negative item types. A wide range of naturally occurring events will be used so that generalisations to everyday judgements can be made.

It is hypothesised that overconfidence will be greatest in mid base rate ranges, where base rate information is not as useful for making predictions. No interaction between base rate and outcome-positivity/negativity should occur, as the availability of base rate information should not be dependent upon the outcome of the event.

Method

Subjects

The subjects who participated in the experiment were 98 undergraduate students (29 male and 69 female) with a mean age of 20.65 years (range 18 to 43 years). All of the subjects were first-year psychology undergraduates, who completed the experiment in return for course credits. The subjects were unaware of the aims of the experiment and had not been exposed to any information about the overconfidence effect or issues involving the calibration of confidence in judgements. Due to the week-long separation between the two testing sessions of the experiment, attrition of one female subject occurred.

Design

A 2 x 12 factorial repeated measures design (Positivity/Negativity x Base Rate) was used. The dependent variables, both measured as percentages, were the accuracy and rated confidence of the predictions, and from these data levels of overconfidence were calculated as explained below. Subjects' estimates of the base rate for each event were also measured on a percentage scale.

Materials

A pilot study was carried out to select the questionnaire items to be used in the experiment. An initial pool of 170 candidate items (see Appendix G), which included academic, social, and personal events, was assembled largely from the records of three volunteers who kept diaries of all noteworthy events that occurred in their lives within a week, and additional candidate items were added intuitively by the experimenters with the aim of including relatively infrequent events that had not turned up in the diaries.

Two categories of events were constructed: those with outcomes that were assumed to be positive and those with outcomes that were assumed to be negative. The items were all rated for outcome positivity/negativity by the subjects in the pilot study to establish empirically their true perceived positivity or negativity within the population being studied. Within these two categories (positive-outcome and negative-outcome), items were included with the aim of spanning the entire range of base rates, from low base rate events assumed to happen only very rarely within the target population of university students (e.g., “break a bone in your body”) to high base rate events that were assumed to be likely to happen to most students within a week (e.g., “wash your hair”). Other items included “fail to study sufficiently”, “smoke a cigarette”, “visit a doctor”, “have problems using a computer”, “have an argument”, etc.

The subjects who participated in the pilot study were 59 undergraduate students (7 males and 52 females) with a mean age of 19.74 years. Full instructions were given at the top of the pilot questionnaire (see Appendix G), and subjects were instructed to ask an experimenter for help if they had any difficulty understanding them. The subjects were given half an hour to complete the pilot questionnaire and were required to fill in their name, age and sex. The pilot study subjects (N = 59) were asked to state, by circling either “Yes” or “No”, whether each of the 170 events had occurred in their lives within the last week (e.g., “In the last seven days did you use a computer?”). They were also asked to rate the outcome of each event as either positive (+), negative (-), or neutral (Nu). Positive-outcome events were described as events that the subject would like to experience and negative-outcome events were ones that subjects would not like to experience; neutral events were ones about which subjects felt neither positive nor negative.

Subjects' responses to each item were scored 1 for a positive answer, 0 for a neutral answer and -1 for a negative answer. The total outcome positivity or negativity score was calculated for each item by summing the subjects' individual scores. The resulting positivity/negativity scores ranged from -42 to +43 compared with a theoretically possible range (if all subjects had responded identically to the most extreme items) of -59 to +59. Twelve of the items, classified as neutral (scoring from —10 to +10), were omitted from the main study because they were ambiguous inasmuch as there was insufficient consensus among students as to whether they were positive or negative. This left a total of 76 negative items and 82 positive items. The base rate for each event was defined as the percentage of students in the pilot study who experienced the event in the critical week. The resulting list contained 158 items, which were then rank-ordered according to their base rates.

The base rate percentage scale was divided into ten ranges representing 10% intervals. In addition the highest and lowest base rate ranges were sub-divided off at 3% and 97% to give two extra ranges of 3% at both extremes of the base rate range, because these were areas of special interest; where fine detail was required. The final set of base rate ranges are shown in Table 2.1.

Table 2.1

Ranges and their Corresponding Base Rates

Range	1	2	3	4	5	6	7	8	9	10	11	12
Base Rate %	0-2.9	3.0-9.9	10.0-19.9	20.0-29.9	30.0-39.9	40.0-49.9	50.0-59.9	60.0-69.9	70.0-79.9	80.0-89.9	90.0-96.9	97.0-100

Four positive-outcome and four negative-outcome items were selected out of the initial pool for each range, but only three negative-outcome items were available for range 10 and none for ranges 11 and 12. This was because, fortunately, there appeared to be very few negative-outcome events that were experienced by a high proportion of the population within a given week. Thus a smaller pool of 87 items was finally selected which covered most of the base rate range for both positive-outcome (48 items) and negative-outcome items (39), and these formed the main study questionnaire. The questionnaire was used twice in order to elicit both predictions and then to determine actual outcomes. The items were presented in the same order but instructions to subjects varied, with different responses required, according to which questionnaire was being completed (see Appendix G: 1st and 2nd Questionnaires).

Procedure

The first questionnaire was completed by 98 subjects who had not taken part in the pilot study. Subjects were asked to make self-predictions by circling either "Yes" or "No" on the questionnaire according to whether they thought that they would experience each of the events within the next week. They were also asked to write down a number, using a half-range percentage scale to indicate their degree of confidence that their judgement was correct. The anchors of the scale were 50% (*Zero confidence*) at one extreme and 100% (*Total confidence*) at the other.

50%-------------------100%

Zero confidence Total confidence

This scale and the following instructions were based upon those used by Dunning et al. (1990) and Vallone et al. (1990). The subjects were instructed as follows:

> "Your confidence estimate should correspond to the percentage of time you would expect to be right on judgements with that level of confidence. For example 90% means that you are very sure that you are right in your prediction and that you'd expect to be right 9 times out of 10, or the probability your answer is correct is 0.9. The 50% level shows complete uncertainty and you would expect to be right about half of the time, due to chance. (Confidence estimates below 50% are inappropriate as 40% confidence in predicting Yes is equivalent to saying 60% confidence in No)" (see Appendix G).

Subjects were re-tested one week later when they were given the second questionnaire, which included the same items as the first questionnaire, with instructions to state whether the event had occurred within the last week since completing the first questionnaire (see Appendix G). Subjects were also asked to rate each item as positive, negative or neutral in outcome and to estimate how many other first-year students they thought would have experienced the event within the previous week.

Subjects' self-predictions on the first questionnaire were then compared with their own answers on the second questionnaire to determine their accuracy for each item. A prediction that an event either would or would not occur within the specified time was defined as accurate if it was fulfilled in reality. The mean percentage confidence and accuracy were calculated for each subject across all of the items and for each individual item across all subjects, and then the percentage overconfidence was calculated for each subject and for each item using the following equation:

Overconfidence = Confidence - Accuracy

The base rate for each item was recalibrated using the follow up data for the main study alone, so that the base rate would refer precisely to the week in which the subjects were tested and about which they made their predictions. This resulted in many of the items shifting in base rate, and some moving across boundaries into different base rate ranges. This does not mean that the first estimate was inaccurate but that the week in question had different base rates for the items compared with the week used in the pilot study, because circumstances had changed slightly between the two studies. For example in the pilot study, which was conducted near the beginning of the first term, only 32.20% of students said that they would skip more than one lecture in the specified week. This base rate, however, increased to 71.13% several weeks later at the time of the main study, presumably because of changing attitudes towards lecture attendance and increased coursework commitments. The base rates were pertinent to the specific time span to which the judgements referred and to the specific target population under investigation. In general, base rates may apply to the future as a whole, but in this experiment they were limited to a specific week, after which the accuracy of the predictions for that week could be determined empirically.

Product-moment correlation coefficients and t-tests (within-subjects) were calculated between subjects' accuracy and confidence ratings across items and between subjects' estimates of item base rates and objective base rates. Data for statistical analyses are presented in Appendix H. Sex differences for all variables were analysed using between-subjects t-tests. Items were classified as either positive or negative according to the outcomes of the specified events and were grouped into base rate ranges according to their objective frequencies of occurrence during the week in question. Differences between positive-outcome and negative-outcome items and differences across base rate ranges in confidence, accuracy, overconfidence and estimated base rates were then tested for significance with 2 x 12 analysis of variance. The analysis was done with items as the unit of analysis, because the study aimed to investigate the effects of different kinds of items, rather than the differences between the subjects themselves. It must be noted when looking at tables and base rate graphs that not all means are equally weighted, some ranges are comprised of more items and judgements than others (see Table AG2 in Appendix G).

Results

Across all items combined, subjects were accurate 82.10% of the time, and their mean confidence was 84.42%. These figures yielded a significant level of overconfidence of 2.32%, $t(86) = 3.07$, $p < .01$ (all statistical results are shown in Appendix I). The levels of overconfidence varied with the base rate and the mean overconfidence scores alone do not reveal the curvilinear (inverted U shape) relationship between base rates and overconfidence that can be seen in Figure 2.1.

Base Rate

The effect of base rates on overconfidence (using a 2 x 12 ANOVA) was statistically significant, $F(11,65) = 8.58$, $p < .001$, and a posteriori Tukey-HSD test revealed that the extreme base rate ranges 1, 2, 10, and 11 were significantly lower in overconfidence than the central ranges 5, 6 and 7. Range 1 was also significantly lower than ranges 4 and 9 and range 3 was significantly lower than range 6. Range 12 did not differ significantly in overconfidence from the intermediate ranges, but this was probably due to the very small sample size of only two items within range 12. These results show that overconfidence at the extremes of the base rate range was significantly lower than at the middle ranges. In fact, at the extremes, *under*confidence occurred, with accuracy higher than confidence, whereas at intermediate levels of base rates *over*confidence occurred, peaking to a maximum of 10.16% in the 40—50 percent base rate range (range 6), or 13.98% in range 7 for positive items alone.

The reason for the inverted U shape between base rate and overconfidence can be clarified by examining levels of confidence and accuracy separately (as shown on Figure 2.1). The level of accuracy clearly followed a U-shaped curve across the base rate range, being significantly lower in ranges 5, 6, 7, 8 & 9 than at the extreme ranges 1, 2, 11 & 12 and ranges 5, 6 & 7 were lower in accuracy than ranges 3, 4 & 10, range 6 was lower than 9, and range 3 was lower than ranges 1 & 11, $F(11,65) = 16.50$, $p < .001$. Subjects' confidence in their judgements, however, followed a much shallower U-shape, dropping much less than accuracy did in the intermediate base rate ranges. Ranges 6 & 7 were significantly lower in confidence than ranges 2, 4, 11 & 12 and range 9 was lower than 11 & 12, $F(11,65) = 4.54$, $p < .001$. As a consequence, high levels of overconfidence occurred in the intermediate base rate ranges where confidence did not drop as low as accuracy. The data shown in Figure 2.1 may appear to suggest that subjects were reluctant to use the extremes of the scale when reporting their confidence. This does not appear to be the case, however, because of all the judgements made by the subjects, 25.76% were made with 90-99 percent confidence and 29.80 percent with 100 percent (i.e., total) confidence.

Accuracy and confidence were positively and significantly correlated across items, $r(87) = .69$, $p < .001$, whereas accuracy and overconfidence were found to be significantly negatively correlated, $r(87) = -.76$, $p < .001$, with high levels of overconfidence being associated with low levels of accuracy. The correlation between confidence and overconfidence over items was non-significant. It is apparent, therefore, that it is variation in accuracy rather than variation in confidence that accounts for variation in overconfidence across base rates.

Outcome Positivity/Negativity

The effect of the second factor, outcome positivity/negativity, was also significant, with mean overconfidence higher for positive-outcome items (M = 3.05) than for negative-outcome items (M = 1.42), $F(1,65) = 12.39$, $p < .001$. In fact, over all positive-outcome items there was significant overconfidence, $t(47) = -3.54$, $p < .001$, but over all negative-outcome items there was no significant mean overconfidence, $t(38) = -1.09$, $p > .05$, because overconfidence balanced out with underconfidence across the base rate ranges. There was no significant interaction between outcome positivity/negativity and range. Table 2.2 summarises the obtained confidence, accuracy and overconfidence and other measured variables for the two types of items. The difference in accuracy between positive-outcome (M = 83.61) and negative-outcome items (M = 80.24) was not significant. Confidence, however, was significantly higher for positive items (M = 86.67) than for negative items (M = 81.66), $F(1,65) = 7.25$, $p < .01$.

Figure 2.2 shows that the inverted U-shaped relationship between overconfidence and base rates was present for both positive-outcome and negative-outcome items. The figure also shows that, for both types of items, overconfidence occurs only in the intermediate base rate ranges, with no overconfidence, but rather *under*confidence, occurring at both low and high base rate ranges. Positive-outcome items elicited more overconfidence than negative-outcome items in all base rate ranges, except between 40-50% (Range 6), where overconfidence was about equal. There was no significant interaction between positivity/negativity and base rate range for either accuracy or overconfidence but there was one for confidence, $F(9,65) = 2.23$, $p < .05$.

Across items, the number of subjects predicting that each event would occur correlated highly with the numbers of subjects reporting that the events did in fact occur, $r(86) = .97$, $p < .001$. Over all items, subjects predicted that 51.59% of the specified events would occur. The actual number of events that did occur (the base rate) was 47.54 percent, so subjects were significantly over-predicting that the event would occur by 4.05 percent, $t(86) = 4.78$, $p < .001$.

When this over-prediction is broken down, no significant effects of outcome positivity/negativity can be seen (see Table 2.2). Subjects predicted that 56.06 percent of positive-outcome events would occur, an overestimate of 4.47 percent, and that 46.10 percent of negative-outcome events would occur, which again was an overestimate of 3.54 percent. Thus there was no unrealistic optimism apparent.

Table 2.2

Differences Between Positive-Outcome and Negative-Outcome Events in Predictions, Base Rates, Accuracy, Confidence, Overconfidence and Directional and Absolute Accuracy of the Estimated Base Rate

	Items			
Mean Percentage	All	Positive	Negative	*F*
Predicted to Occur	51.59	56.06	46.10	1.88
Objective Base Rate	47.54	51.59	42.56	2.34
Confidence	84.42	86.67	81.66	7.25 **
Accuracy	82.10	83.61	80.24	.87
Overconfidence	2.32	3.05	1.42	12.39 **
Estimated Base Rate	44.52	48.91	39.11	1.11
Estimated Base Rate: Directional Accuracy	-3.02	-2.68	-3.45	.53
Estimated Base Rate: Absolute Accuracy	13.05	12.63	13.56	.75

** $p < .01$

Estimates of the Base Rate

Outcome positivity/negativity had no significant effect on the subjects' estimates of the base rate or their resulting accuracy of their estimates. The objective base rates for positive-outcome and negative-outcome items were not significantly different. Subjects' estimates of item base rates correlated highly with objective base rates, $r(87) = .88$, $p < .001$, and the difference between the mean estimated base rate ($M = 44.52$) and the true mean base rate ($M = 47.54$) for all the items was not significant. Thus generally speaking subjects did make an accurate estimate of the objective base rate of each event. However, the relationship between estimates of base rate and real base rates is more complicated than this, as Figure 2.3 shows. Figure 2.3 shows that when the objective base rate of an item was less than 40% (i.e., below base rate range 6) subjects tended to overestimate the base rate by 10.79% on average, whereas for base rates above 40% they tended to underestimated the base rate by -12.78% on average ($Z = -5.28$, $p < .001$). The objective base rates for positive-outcome and negative-outcome items were not significantly different.

The accuracy of the base rate estimates was found by subtracting objective base rates from the estimated base rates - this was a directional measure. There was no significant correlation between directional accuracy of base rate estimates and overconfidence. This indicates that over- or underestimating the base rate did not correlate with overconfidence in an item. The directional accuracy of base rate estimates correlated significantly with the accuracy of predictions, $r(87) = .23$, $p < .05$, and with rated confidence of judgements, $r(87) = .31$, $p < .01$. Thus overestimates of the base rate were associated with higher accuracy and higher confidence in items, and the reverse for underestimates of the base rate.

Taking an absolute deviation measure of the estimated base rate from the actual base rate reveals that more accurate estimates of base rates (i.e., lower discrepancies, irrespective of the direction of the error) were associated with significantly higher confidence, $r(87) = -.23$, $p < .05$, but there was no significant correlation between the absolute estimate of the base rate and accuracy or overconfidence. The correlation between the objective base rates and the absolute accuracy of the base rate estimates was also non significant.

Predictions With and Against the Base Rate

Table 2.3 clearly shows that there is no overall overconfidence (-2.12%) for the 76.20 percent of the judgements that were made in line with the base rate. All of the overconfidence resulted from the 23.80 percent of judgements that were made going against the base rate, resulting in 23.34% overconfidence. Subjects did lower their confidence when making predictions against the base rate (M = 77.14 vs. M = 85.54) but evidently not sufficiently as their accuracy was only 53.80% for predictions against the base rate compared with 87.66% for predictions in line with the base rate.

The lowest level of accuracy (M = 46.37) resulted when subjects predicted against the base rate by saying that they would experience events when, objectively, most people would not in fact experience the event (i.e., when the base rate was less than 50%). Conversely, when subjects predicted that events would not happen to them when the base rate suggested that they probably would, (i.e. the base rate was more than 50%), accuracy was higher (M = 63.59). Subjects were approximately 77% confident when making predictions against the base rate, and this results in the most erroneous (highly overconfident) predictions being for those events where the majority of the population do not experience an event but the subject predicted that it would happen to them (30.36% overconfidence). Much less overconfidence (14.10%) occurred when subjects predicted that they would not experience events that most members of that population did, in fact, experience.

A further breakdown of the data reveals whether predictions that events would happen or would not happen resulted in most accuracy, confidence and overconfidence (Table 2.4). Obviously when only the correct judgements are looked at underconfidence is found, as by definition confidence could not exceed 100% accuracy. Overconfidence results from the judgements which turn out to be incorrect. Subjects who predicted that "Yes" an event would happen tended be more overconfident because their accuracy (M = 65.22%) was lower than when subjects predicted "No" (M = 77.20%), as confidence was similar for both groups.

Table 2.3

With Base Rate and Against Base Rate Predictions and Resulting Mean Percentage Confidence, Accuracy and Overconfidence

	Number of Predictions	Confidence	Accuracy	Over-confidence
With Base Rate	6409			
BR 0-50 (No)	3210	86.80	89.90	-3.10
BR 50-100 (Yes)	3199	84.27	85.42	-1.14
Mean		85.54	87.66	-2.12
Against Base Rate	2002			
BR 0-50 (Yes)	1138	76.72	46.37	30.36
BR 50-100 (No)	864	77.69	63.59	14.10
Mean		77.14	53.80	23.34

Subjects in this experiment were more accurate when they predicted that events would not occur, and less accurate when they said that they would. This may be because subjects can think of concrete reasons why events will not happen, which makes predictions more accurate, but not concrete reasons for why events will occur.

Table 2.4

Predictions and Outcomes with their Resulting Confidence Accuracy and Overconfidence

	Predicted Yes		Predicted No	
	Outcome		Outcome	
	Yes (Correct)	No (Incorrect)	Yes (Incorrect)	No (Correct)
Base Rate				
Less than 50%				
Confidence	82.71	72.87	82.51	87.28
Accuracy	100.00	0.00	0.00	100.00
Overconfidence	-17.29	72.87	82.51	-12.72
More than 50%				
Confidence	85.89	73.87	75.37	79.43
Accuracy	100.00	0.00	0.00	100.00
Overconfidence	-14.11	73.87	75.37	-20.57
Over All Items				
Confidence	84.34	73.35	79.07	83.49
Accuracy	100.00	0.00	0.00	100.00
Overconfidence	-15.66	73.35	79.07	-16.51

	Irrespective of the Outcome	
Confidence	80.37	82.40
Accuracy	65.22	77.20
Overconfidence	15.15	5.21

NOTE: N = 4337 for Yes and N = 4074 for No

From Table 2.4 it can be seen that the mean level of confidence in judgements which were correct was around 83.49 to 84.34%, less confidence was shown for judgements which subsequently turned out to be incorrect, 79.07% and 73.35% for Yes and No predictions. This indicates that subjects were aware that their judgements may be incorrect and did have lower confidence (certainty) in those predictions.

Calibration of Confidence

Irrespective of the base rate of the item the individual judgements and respective confidence and accuracy levels were analysed to see if high confidence leads to overconfidence, and which predictions were most inaccurate (see Table 2.5).

Table 2.5

Levels of Confidence in Predictions with their Resulting Accuracy and Overconfidence

Confidence in Predictions	Number of Predictions		Confidence	Accuracy	Over-confidence
	Total	%			
Low (50-69%)	1439	17.13	57.49	62.17	-4.68
Medium (70-79%)	1107	13.18	72.47	73.08	-.62
Medium High (80-89%)	1188	14.14	81.29	82.16	-.87
High (90-99%)	2164	25.76	92.26	87.94	4.32
Complete (100%)	2504	29.80	100.00	92.65	7.35
All	8402	100.00	84.45	82.16	2.30

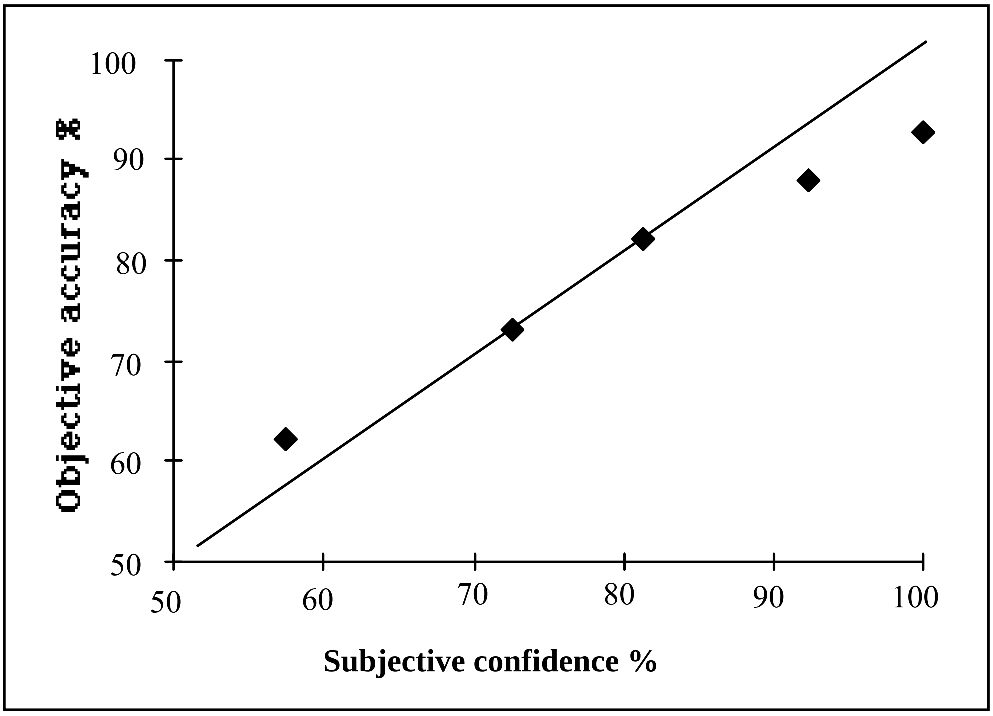

Figure 2.5. Calibration curve of confidence and accuracy in self-predictions.

The data in Table 2.5 shows that of the 8402 judgements made, over half of them (55.56%) were made with high (90-99%) or complete (100%) confidence. Thus there is no evidence to support the idea that subjects may be reluctant to use the extremes of the confidence rating scale. These were in fact the judgements which resulted in *over*confidence of between 4 and 8 percent, whereas judgements made with lower confidence levels (50-89%) resulted in a small overall level of *under*confidence of between -5 and 0 percent. Judgements which were low in confidence were also low in accuracy and as confidence increased so did accuracy. Predictions made with medium levels of confidence (70-89%) were the most appropriate and the least over- or underconfident, resulting in good calibration. In fact it would seem from the final calibration curve, shown in Figure 2.5, that in this experiment, which used a balanced test of self-predictions covering the entire base rate range, overconfidence only occurs when confidence is above 90 percent, and even at this level overconfidence is quite low.

Discussion

Across all items combined, subjects' confidence exceeded their accuracy, and this discrepancy produced a significant level of overconfidence in their judgements. These findings corroborate those of previous researchers (e.g., Dunning et al., 1990; Vallone et al., 1990) who have reported that confidence consistently exceeded the objective accuracy of predictions, especially when confidence was relatively high. These earlier results showed that predictions made with high levels of confidence tend to result in higher levels of overconfidence, but in this experiment items about which people were highly confident were not necessarily the ones that resulted in high overconfidence, because the correlation between confidence and overconfidence was non significant. This apparent inconsistency with earlier reported findings can be accounted for by the inclusion of items at the extremes of the base rate range, associated with very high accuracy and confidence, which produced underconfidence rather than overconfidence. Previous studies have found much greater levels of overconfidence for predictions made with high confidence, because they did not include very predictable, extreme base rate, items; for example, Vallone et al. (1990) found 20.2% and 22.6% overconfidence for judgements made with confidence over 90% and with 100% respectively. As regards item positivity/negativity, the results show that overconfidence was significantly greater for positive-outcome items than for negative-outcome items, as shown in Figure 2.2, and was non-significant for negative-outcome items taken alone, because among those items overconfidence was balanced out by underconfidence.

The outcome positivity/negativity of the events being judged had no significant effects on the subjects' predictions, their predictive accuracy, their base rate estimates or their accuracy of estimating base rates. Therefore subjects do not seem to have been over-predicting positive-outcome events and under-predicting negative ones, which the theory of unrealistic optimism (Weinstein, 1980) would suggest. What *is* affected by outcome positivity/negativity is the level of confidence and thus overconfidence in the judgements, which were both inflated for positive-outcome events. This experiment actually underestimates the difference in overconfidence between positive-outcome and negative-outcome events because there are no negative-outcome events at the top of the base rate range that would have lowered the mean overconfidence for negative items even further below the level for the positive items.

The difference between this study and the previous ones (e.g. Weinstein, 1980) that have reported unrealistic optimism, may be due to the time scale sampled. In this experiment the predicted events were to occur within one week, whereas in other studies the time scale was in terms of subjects' whole future lives. In this experiment, subjects over-predicted events for both positive-outcome and negative-outcome items, leading to unrealistic optimism of predictions for positive-outcome items only, and unrealistic *pessimism* for negative-outcome items. The level of over-prediction for both types of items was relatively low, probably because there was a high level of reliable information, about current situational factors, available to the subjects for making their predictions, thus reducing possible informational retrieval biases. Unrealistic optimism may be more likely to occur for predictions involving a long time scale, because of cognitive or motivational biases, and less likely for short term, everyday judgements.

The higher levels of confidence and overconfidence in judgements involving positive-outcome events can probably be explained by motivational factors. Perhaps unrealistic optimism in *predictions* taps the informational/cognitive aspect of judgements, and unrealistic optimism shown in *confidence* reflects the motivational/ emotional aspects. Weinstein attempted to reduce unrealistic optimism in predictions by asking subjects to generate reasons which might influence whether the event will happen to them or not. This succeeded in reducing the level of unrealistic optimism for the positive events but not the negative ones. Then another group of subjects were asked to read the reasons listed by the first group, and the amount of optimism dropped for both positive and negative items compared to a control group who neither generated nor read reasons. The levels of unrealistic optimism were not totally eliminated however, and this may point to there being motivational factors at work. Weinstein concluded that the generation of mental lists of reasons is a crucial stage in evaluating if our chances are greater or less than average.

Weinstein (1980) did not find any effect of the level (or degree) of outcome desirability on unrealistic optimism of predictions. This may be because in his research the events were all seen as either extremely desirable or extremely undesirable, so there may have been a ceiling effect. The events used in this experiment were much less emotive than those used by Weinstein and produced lower levels of unrealistic optimism in predictions for positive items and unrealistic pessimism for negative items. This suggests that there may be a motivational component operating in unrealistic optimism and that motivation was lower in this experiment because of the less emotive items used. Another consideration is that whereas in previous research subjects compared their likelihood of experiencing each event with other people's likelihood of experiencing those same events, in the experiment reported in this chapter, subjects' self-reported predictions were compared with *objective* self-outcomes, which may provide information about the *absolute*, rather than the *relative*, level of unrealistic optimism.

A possible likely explanation for the significantly higher confidence for positive-outcome than negative-outcome items may lie in the Pollyanna effect (Boucher & Osgood, 1969), named after the unrealistically optimistic eponymous heroine of Eleanor Porter's 1913 novel and the films based on it. According to the Pollyanna effect, people tend to exaggerate the positive and minimize the negative aspects of their lives. It is hardly surprising in this light that people tend to feel more confident about the likelihood of positive than negative experiences, as clearly shown in Table 2.2, and one consequence of this is that overconfidence was greater for positive-outcome than negative-outcome items, as clearly shown in Figure 2.2. The Pollyanna effect seems to operate on confidence more so than on predictions.

Alternatively, Ginossar and Trope (1980) proposed that when more information has to be processed confidence drops, and the differences in overconfidence between positive-outcome and negative-outcome events could be due to differences in information loads for the two types of task. In this experiment positive-outcome and negative-outcome events elicited similar levels of overconfidence in the 40-50% base rate range, and this may be due to similar levels of information being processed in this range where uncertainty is near to 50:50. Further empirical evidence will be needed to determine if there are different amounts of information processed in predictions for positive-outcome and negative-outcome events across the base rate range. The subjects may have more confidence in their positive-outcome judgements because they feel they have more control over their occurrence in comparison to negative-outcome events, and the relation of confidence to perception of control also needs addressing in future research.

The against base rate predictions reveal that denying that an event would occur to oneself, when it happens to most people, results in overconfidence, and is indicative of the item being negative in outcome. Much higher overconfidence occurs if subjects wanted events to happen to them (positive events) when they would not occur for most people (going against a low base rate).

Perhaps the most important finding of this experiment was that overconfidence varied across the base rate range, which resulted in a curvilinear, inverted U-shaped relationship between base rates and overconfidence shown clearly in Figures 2.1 and 2.2. Overconfidence was greatest for intermediate base rate ranges 5 to 7, which spanned base rates between 30 and 60 percent where there was maximal uncertainty as to whether or not the events would occur. This shows the hard-easy effect, that the greatest levels of overconfidence occur when accuracy is low due to uncertainty being high. Subjects evidently expressed excessively high levels of confidence and failed to consider or allow for the true level of uncertainty of the situations in the middle base rate ranges. In extremely low base rate ranges 1 and 2 (corresponding to base rates between 0 and 10 percent) and extremely high base rate ranges 10 to 12 (corresponding to base rates between 80 and 100 percent), overconfidence was not observed, and in fact *under*confidence occurred. This finding is in line with an earlier finding of Lichtenstein, Fischhoff, and Phillips (1982), and is accounted for by the fact that very easy items generally fall at the extremes of the base rate range. It seems reasonable to conclude that the overconfidence effect is confined to events with intermediate base rates of occurrence, which are judgements under conditions of uncertainty. Extreme base rates, where there is relative certainty about outcomes, do not appear to have been adequately investigated in previous research into the overconfidence effect.

Is the relationship between overconfidence and base rate range an artifact? The answer to this question is probably not. One could argue that overconfidence is unobtainable at the extremes of the base rate range as accuracy will always be so high as to exclude overconfidence. However, on examining Table AH1 in Appendix H it can be seen that accuracy reaches 97.40% in range 12 and 95.88% in range 1 (being the most extreme ranges) thus overconfidence could exist in these ranges up to a maximum of 2.60% and 4.12% respectively. Although this is a small percentage it still exists and there is a possibility of 6-9% overconfidence in ranges 2 and 11 which fails to occur, underconfidence of approximately -2% being the norm.

These findings throw some light on the psychological mechanism underlying the overconfidence effect. They suggest that the effect may be explained by a tendency for subjects to make insufficient use of base rate information in making self-ratings of confidence. The findings do not suggest that subjects were oblivious of base rates, or that they were unable to estimate base rates or make any use of them. Furthermore, the overconfidence effect does not appear to be due to inaccuracies in subjects' estimates of base rates, because there was no significant correlation between the accuracy of base rate estimates and overconfidence. Figure 2.3 shows a strong relationship between objective and estimated base rates ($r = .88$), and the mean difference between estimated and objective base rates was small and non-significant. Subjects were thus able to estimate the objective base rate of events reasonably accurately, although Figure 2.3 shows clearly that they tended to overestimate low base rates (below 40 percent) and to underestimate high base rates (above 40 percent). Figure 2.1 shows that the accuracy of predictions was highest in regions of extremely low and extremely high base rates, where base rates provide the most useful information as to whether or not an event will or will not occur, and lowest in intermediate base rate ranges, which shows that subjects were probably using base rate information either directly or indirectly to inform their predictions. It is possible that people generally use base rate information more when the event is externally controlled than when an outcome is within their own control.

Figure 2.1 also shows, however, that the subjects took less account of base rates in modulating their ratings of confidence in their predictions. However, items for which subjects could accurately estimate base rates also generated higher confidence ratings. In intermediate base rate ranges where accuracy was lowest, confidence ratings were hardly lower than in extreme base rate regions where accuracy was significantly higher. At the extremes of the base rate range, where certainty was high subjects were underconfident, they were either unaware of the extremity of the base rate or they did not allow for it.

There is, however, a possible methodological problem in this experiment, the subjects gave their estimates of the base rates for the rest of the student population at the end of the week, i.e. in the second testing session. This should have been done in the first session when subjects made their initial judgements and confidence ratings. Asking them a week later may be problematical as their own experiences over the week may have biased their estimates of the base rates. Subjects were making retrospective rather than prospective judgements, and this should be reversed in future experiments.

The overconfidence effect is evidently explained by the gap between accuracy and confidence resulting from unrealistically high confidence in relation to the accuracy of difficult predictions for events with intermediate base rates of occurrence. This shows then the hard-easy effect, where overconfidence is greatest when difficulty of predictions is greatest. Overconfidence results when subjects go against the base rate, this occurs by definition most frequently in the middle of the base rate range, and explains why the highest level of overconfidence is found there. This also explains why the effect apparently disappears, and is even reversed, for events with extremely low or high base rates of occurrence, as shown in Figures 2.1 and 2.2.

The predictions against base rate reveal that denying that an event would occur to oneself, when it happens to most people, results in overconfidence, and is most common in the case of items that are negative in outcome. Much higher overconfidence occurs when people want a (positive) event to happen to them when the base rate is low, so it is unlikely to happen. This may indicate that self-presentation biases are operating, for example to protect self-esteem, or perhaps that there is a higher level of selective recall for information when events are positive in outcome.

Accuracy dropped in the middle base rate ranges where the statistical uncertainty of the judgements was highest. The lack of equal change of confidence estimates in the middle base rate ranges may be due to the effect of a general level of task confidence that is appropriate to a task but not flexible within that task. Having high confidence in a judgement or belief may make it more resistant to change, and this has obvious implications for health-related and risky forms of behaviour. Strategies for improving people's calibration of confidence may result in more effective attitude change and a less unrealistic perception of risky behaviours.

The principal aim of the experiment reported in this paper was to investigate the relationship between base rates and overconfidence. The results clearly show that the relationship is strongly curvilinear, with overconfidence peaking at intermediate base rates around 40 to 60 percent and decreasing for both lower and higher base rates. The experiment also sought to clarify the underlying mechanism of the overconfidence effect by determining whether people are aware of base rates but cannot make full use of them, or whether they are simply unaware of them. The results suggest that subjects were probably aware of base rates, and made use of them directly or indirectly in formulating predictions, but that they did not take adequate account of them in adjusting their confidence ratings. It would be interesting to know whether these conclusions apply not only to self-predictions but also to social predictions.

What emerges from the results reported here is that the so-called overconfidence effect does not appear to be a unidirectional bias in human judgment across the full range of item base rates, as hitherto assumed, but a more subtle and complex phenomenon. It is evidently a bi-directional bias involving overconfidence for items with intermediate base rates of occurrence and underconfidence for items at both extremes of the base rate range, although in experiments using only items with intermediate base rates it may appear to be a unidirectional overconfidence effect. Overconfidence as a bias in human reasoning may not be a bias if all the underconfidence is never tested and included in the means. Researchers investigating this phenomenon should in future take care to include items from across the entire spectrum of base rates.

CHAPTER 3

Cognitive Styles and Probabilistic Reasoning Ability in Relation to Overconfidence in General Knowledge and the Fundamental Attribution Error

Individual differences in overall levels of overconfidence or underconfidence were hypothesised to be related to other factors such as probabilistic reasoning ability, cognitive styles such as field dependence, and the ability to perceive the influence of situational factors on oneself and for another person (the Fundamental Attribution Error). The experiment was conducted with 60 subjects who completed tests to measure field dependence, probabilistic reasoning, overconfidence, and attributions for oneself and other people. Within-subjects correlations were used to test for any relationships between variables. No evidence was found to support the proposed relationships between these factors, and the hypothesis that field-dependent and field-independent people would show different levels of overconfidence was not supported. Methodological problems and the newly designed Embedded Words Test are discussed.

Cognitive Styles and Probabilistic Reasoning Ability in Relation to Overconfidence in General Knowledge and the Fundamental Attribution Error

As stated in earlier chapters, overconfidence occurs when a person's confidence exceeds their accuracy when making judgements. Overconfidence is usually regarded as a bias or error, which implies that it is inherently undesirable. The question arises: does it matter if a person has a general bias towards over- or underconfidence? There are associated costs and benefits of being cognitively biased. For example, more time, effort and resources may be put into following through with judgements and decisions that people have high confidence in than those with lower confidence levels but perhaps more accuracy. Another problem with overconfident judgements is that disconfirmatory evidence, which should cause a judgement to be reassessed, is often ignored if the person is highly confident (Ross & Lepper, 1980).

Overconfidence can have negative consequences, as it may also discourage fall-back measures from being taken. Consider, for example, a student who is highly confident about obtaining good "A" level grades and who applies only to universities that require high entry grades. A second student, who is less confident and better calibrated, may also have a backup measure (such as having applied for a job) in case of exam failure. Both students may end up with the same grades, high or low, but the consequences for the less confident student in the case of low grades are more favourable than for the highly confident student. Thus overconfidence can have negative consequences, but a lack of confidence can also prove negative, as in the case of a third student who achieves high exam grades but did not apply to a university at all, because of *under*confidence in their ability to get high grades. A general bias towards over or underconfidence may have advantages and disadvantages depending upon the situation. Flexibility in when to be over- or underconfident may be the optimal strategy.

Individual Differences

Cognitive factors alone probably do not account for the huge individual differences observed between subjects in the calibration research literature; personality traits should also be studied for their influence on confidence. Wright and Phillips (1976) took seven measures of calibration and found two correlations with authoritarianism (.41 and .34) but not with any measures of conservatism, dogmatism, intolerance of ambiguity, or verbal expressions of uncertainty used in answering questions. The amount of evidence relating to individual differences and how they affect *overconfidence* is currently quite limited, but related research may lead to some predictions. An early experiment, by Adams and Adams (1961), found that students who had a higher discrepancy score between their confidence and accuracy on exam questions also had a higher fear of failure prior to the exam ($r = .36$). The actual level of the students' knowledge was unrelated to their fear of failure. The correlation was with the absolute discrepancy scores of overconfidence, the algebraic discrepancy score hardly correlated with the variables measured, thus it was the magnitude of the overconfidence or underconfidence (poor calibration) rather than the direction of the bias which was related to a fear of failure. They concluded, therefore, that it is the inappropriate assessment of one's knowledge which resulted in greater fear. Those people who cannot accurately assess how much they know tend to suffer from higher apprehension and fear of failure. This implies a negative consequence of being poorly calibrated. The relationships between overconfidence and stable factors within the person should also be considered.

Cognitive Styles

Two different approaches to understanding the causes of human behaviour arose in the history of psychology. One perspective maintains that a given stimulus will cause a given response from an organism and only these two factors need be studied, whereas the other perspective includes the organism itself as a mediating factor between the external stimulus and a response by an organism. The latter perspective proposes that only by understanding the processes going on within the organism can behaviour be understood and/or predicted. Such processes include motivation, ability, temperament, etc. Another important idea is that the organism's cognitive construction of reality may be more important than the objective reality itself. With respect to this idea, the way that the organism views the environment, that is, its cognitive styles of thinking and categorizing, have been studied.

A cognitive style is a way that a person conceptually organises the environment, and this refers to the structure rather than the content of what is stored. Bieri (1971) proposed that people have *cognitive structures* which are strategies or programmes which extract and translate meaning from the incoming stimuli. Cognitive structures are also used to structure the outgoing behaviour of the person as well as the incoming information (Zajonc, 1968). A person's cognitive style is a consistent and stable preference for organizing and categorizing the external environment, within their cognitive structures.

Cognitive styles have been studied from many approaches such as scanning, levelling-sharpening, constricted-flexible control, tolerance for incongruous or unrealistic experience, field dependence, cognitive complexity, reflection-impulsivity, styles of conceptualization and styles of categorization (see Goldstein & Blackman, 1978, for a review). A cognitive style which has received a lot of attention in the past is that of field dependence. The area of field dependence has received much research attention and is the area which will be looked at in this chapter and Chapter 4. Witkin, Dyk, Faterson, Goodenough, and Karp (1962) proposed that people either have an articulative or a global cognitive style. The articulative style suggests differentiation and a piecemeal tendency to separate figure from ground, whereas the global style is more holistic and has a more undifferentiated understanding (Cacioppo & Petty, 1982).

Research on the field dependence cognitive style has mainly used two types of test: the Rod-and-Frame Test (RFT) and the Embedded Figures Test (EFT). Subjects in the RFT have to orient a luminous rod to the vertical without reference to a surrounding frame. Subjects who can orient the rod accurately, or who can recognise simple geometric shapes within a complex background in the EFT, are termed *field-independent*. Those subjects who cannot do these tasks because their perceptions are dependent upon the surrounding cues, are termed *field-dependent*. Distinctions have been drawn between field-dependent and field-independent people, on the grounds that field-dependent people rely upon surrounding cues more than field-independent people who can ignore extraneous information more easily. Witkin, Lewis, Hertzman, Machover, Messiner, and Wapner (1954) studied perceptual tasks and found that large individual differences existed, which after further research turned out to be related to other factors such as personality and cognitive style.

Ability to do the EFT is not only a perceptual ability, it relates to non-perceptual intellectual tasks, and so Witkin, Oltman, Raskin, and Karp (1971) widened the construct to mean how differentiated people are. The more differentiated the person is, the more field-independent they are, that is, they see the field as more discrete and structured. Differences have been shown to exist between field-dependent people, who are more socially sensitive, and field-independent people, who have some improved cognitive abilities.

Many versions of the EFT have been used and shown to have satisfactory reliability, however the use of naive subjects is advisable because practice and training can improve EFT scores (Goodenough & Witkin, 1977). Early research pointed towards a gender difference in field dependence, with females being more field-dependent than males. Research on this issue has been contradictory and it has been suggested that sex differences only arise because of factors within the test situation, which can be eliminated if controlled for (Naditch, 1976). A more consistent finding is that field-independent subjects score more highly on intelligence tests, (Dubois & Cohen, 1970; Wachtel, 1971) although the independence of the two measures has been questioned.

Because field dependence relates to how the person conceptualises and perceives the environment, people may differ in how accurate they are in their judgements. Cognitive styles may influence how people assess and categorise probabilities and confidence and so the relationship between field dependence and overconfidence will be examined in this chapter. There are several possible ways that field dependence may relate to overconfidence in judgements. Field-dependent people cannot pick out relevant information from the field of information as easily as field-independent people, are more easily overwhelmed by information, and can see more than one perspective, and this may result in them being less confident. Field-independent people may also be more overconfident than field-dependent people because they need less information to make decisions, so their confidence may be higher because they base their judgements on less information; as more conflicting information lowers confidence.

Rationale for Experiments 3 and 4

Field-dependent and field-independent people may have different levels of accuracy and confidence in their judgements. Different types of judgements may also result in differences between field-dependent and field-independent people in overconfidence. Wolitzky (1973) pointed out that field-independent subjects are more accurate in social perceptions, even though field-dependent subjects may be more responsive to them. Therefore, if field-independent subjects are more accurate in social judgements then they should be less overconfident than field-dependent subjects if they have similar confidence levels. Field-dependent people are, however, more socially aware than field-independent people and thus may be more confident in social judgements, because they are more comfortable making social judgements. In contrast, field-independent people may be more overconfident at general knowledge type judgements. Chapter 4 will examine the effect of field dependence on overconfidence in relation to social judgements, and this chapter will examine it in relation to a general knowledge task.

The ability to see another person's perspective and realise the influence of situational factors on their behaviour has been widely studied. The tendency to underestimate the role of the situation on other people's behaviour but perceive it as influencing oneself has been termed the Fundamental Attribution Error (FAE). This phenomenon is often called the Fundamental Attribution Bias or overattribution bias (Jones and Harris, 1967; Gilbert & Jones, 1986; Johnson, Jemmott, & Pettigrew, 1984; Ross, Amabile, & Steinmetz, 1977; Webster, 1993) according to which people tend to overestimate the causal role of dispositional relative to situational factors in explaining an actor's past behaviour. A measure of people's degree of committing the FAE will be measured in this, and the following, chapter to see if the ability to perceive the influence of situational factors on behaviour is related to the cognitive style of the person, or to their overconfidence in judgements.

Overconfidence is usually measured as a mean value across many judgements for one person, or across a population of people for one judgement. The analysis in this chapter, and Chapter 4, which both look at individual differences in overconfidence for different types of tasks, will use the person as the unit of analysis to determine if people generally have a bias in one direction or another, to what extent it exists, and if it relates to other individual differences such as probabilistic reasoning ability and the ability to see another person's perspective (as measured by the FAE score).

Method

Subjects

The subjects were 60 students from the University of Leicester. The subjects ranged in age from 18 to 41 years old (mean age 23.55 years) and 27 males and 32 females, and one subject of unknown sex (missing data), participated. Subjects were either first-year psychology students, who participated as part of their course requirements, or students who volunteered in return for a raffle ticket for a prize of twenty pounds.

Design

A within-subjects correlational design was employed. The order of presentation of the two Fundamental Attribution Error tests was counterbalanced and the order of the questions on this test was different for the self- and other-questionnaires.

Materials

Five tests were used in this experiment: the Embedded Figures Test; a probability test; an Embedded Words Test; a general-knowledge overconfidence test; and a test of the Fundamental Attribution Error (all test materials, their answers, and instructions to subjects are presented in Appendix J). The Embedded Figures Test (EFT) was copied from the examples given in the paper by Witkin (1950). The eight simple figures were presented on one side of the paper and the twenty four complex figures on the reverse side. An example was given along with detailed instructions on how to complete the test. Subjects were required to find and mark the shape of the simple figure within the complex one.

The Embedded Words Test (EWT) was specially designed for the purpose of this experiment as a possible alternative measure of field dependence. It consisted of ten words, each of ten letters in length. The words were selected to contain many smaller words embedded within them. Subjects were required to write down as many words as they could see embedded within the longer word.
The ten words were: *typewriter*; *brainpower*; *coincident*; *importune*; *everywhere*; *weathering*; *researcher*; *malefactor*; *protestant*; and *photograph* (see Appendix J). An example of the task was given to the subjects:

ATTEMPTING contains the words: at; tempt; tin; in; attempt; tempting

Each individual's overconfidence, displayed in a test of general knowledge, was measured by asking for the answers, and associated confidence in those answers, for twelve questions with numerically objective answers. For example: how many grammes are there in an ounce? Subjects were asked to rate how confident they were that their answers were correct - to within ten percent of the true answer. They did this by ringing a number on an 11-point rating scale from 0% (*No confidence*) to 100% (*Total confidence*), using 10 percent intervals (see Appendix J). The probability test, which was based on the questions asked by Blackmore and Troscianko (1985), consisted of five probability questions; four with multiple choice answers, and a task involving the generation of a list of 20 random numbers using the numbers 1 to 5.

The Fundamental Attribution Error test consisted of two questionnaires asking subjects to rate the causes of their own behaviour and that of an unknown person in twelve short scenarios. Half of the scenarios had positive outcomes and the other half had negative outcomes. The behaviours were rated on a seven-point scale, ranging from 0 (*Entirely due to the person involved*) to 6 (*Entirely due to the situation*), indicating the internal or external nature of the attribution of causality.

Procedure

Subjects were tested in groups of between one and nine and they were instructed that they could answer the questionnaire anonymously if they wanted to, but were required to state their age and sex. Subjects were given ten minutes to complete the Embedded Figures Test, eight minutes for the Embedded Words Test, and unlimited time to complete the remaining tests. Subjects were instructed that they should ask for help at any time if they did not understand anything. After all subjects had completed the questionnaire they were debriefed and any questions that they had were answered.

Test Scoring

The number of correct answers on the probability test was calculated, out of a maximum possible score of 4 (answers are shown in Appendix J). The number of runs produced in the random number generation task on the probability questionnaire was counted for each subject (the word runs shall be used to indicate doubles or triples, for example, 33 or 444).

The Embedded Figures Test had a maximum score of 24 correctly perceived and outlined figures. The subjects' answer sheets were scored by placing a transparency with the correct figures on it over the answer sheet. The answer was marked as correct if the subject had clearly seen and marked the simple figure within the complex one. A higher score on the EFT would indicate that the subject could disembedd more figures within the limited timespan and is thus more field-independent.

There were 106 real words which could possibly be seen in the Embedded Words Test, that were checked in the *Oxford Encyclopedic English Dictionary* (1991) and considered to be correct answers (presented in Appendix J). When scoring the subjects' responses, non-words, names, abbreviations, prefixes, suffixes, combining forms and languages other than English were not allowed. The number of errors that each subject made, along with the total number of correct answers (out of 104), was recorded. A higher score on the EWT would indicate that the subject could disembedd more words.

The answers on the general knowledge test were marked as correct if they fell within 10 percent on either side of the true answer (answers and 10% boundaries are shown in Appendix J). The mean number of correct answers was recorded as the subject's accuracy score. The mean confidence rating across all of the answers was also calculated. Each subject's overconfidence score was calculated by subtracting the percentage of correct answers (accuracy) from the mean percentage level of confidence.

The Fundamental Attribution Error score was equal to the subject's mean rating score for self-attributions minus the mean score for other-attributions across all twelve scenarios. Thus a positive FAE score would indicate that subjects saw themselves as more influenced by situational factors, and a negative score indicates that the other person was seen as more under the control of situational factors than oneself. The subject's mean scores for positive-outcome and negative-outcome scenarios were also recorded for self- and other-attributions. The possible range of attributions was 0 to 6, with 3 indicating that both internal and external factors were equally responsible for the situation described in the scenario.

Two questionnaires were discarded totally as the subjects had not understood the instructions on two of the tests and had put vague and unclear answers on a third. Relationships between scores on these tests were tested for, using Pearson correlation coefficients. Order-of-presentation effects and the effects of gender on the scores were made using t-tests. All data are presented in Appendix K and statistical analyses in Appendix L.

Results

Field Dependence

The mean score on the Embedded Figures Test was 17.37 ($SD = 5.53$) out of a possible maximum score of 24. The mean score of correct answers on the Embedded Words Test was 53.02 ($SD = 8.49$) out of a possible 104. A small but significant positive correlation was found between the score on the Embedded Figures Test and the score on the Embedded Words Test, $r(58) = .28$, $p < .05$. A high score on the Embedded Words Test was not significantly associated with the error score on the test, indicating that subjects were not just rapidly guessing answers. Subjects who scored high on the Embedded Words Test tended to score lower on the general knowledge test, $r(59) = -.31$, $p < .05$, but this did not occur for the Embedded Figures Test, where there was no correlation with the accuracy score on the general knowledge test. No other measured variable significantly correlated with the scores on these two measures of field dependence. No evidence was found to support the predicted relationship between field dependence and overconfidence.

Overconfidence

The mean level of confidence expressed in the answers given on the general knowledge test was 40.16% which was significantly higher than the mean level of accuracy (19.98%) achieved, $t(59) = 8.23$, $p < .001$. This resulted in an overall significant level of overconfidence of 20.18%. Mean confidence and accuracy scores for each subject were correlated, $r(59) = .34$, $p < .01$, indicating that subjects who were more confident were also more accurate, although the relationship is quite small. Subjects who were more accurate were less overconfident than those less accurate subjects, $r(59) = -.37$, $p < .01$, and those who were more confident were more overconfident than those subjects who were less confident on average, $r(59) = .74$, $p < .01$.

Male subjects had significantly higher confidence in their knowledge than female subjects (M = 46.53% vs. M = 33.77%), $t(57) = 2.90$, $p < .01$, and this appeared to be because the male subjects had a significantly higher accuracy score on the general knowledge test than female subjects (M = 24.19% vs. M = 16.27%), $t(57) = 2.31$, $p < .05$. There was no significant difference between male and female subjects in mean overconfidence, although it was slightly higher for males. Thus, in this experiment, there was no significant sex difference in overconfidence in general knowledge.

Fundamental Attribution Error

The order of presentation of the two FAE tests was counterbalanced, and a manipulation check, using between subjects t-tests, showed that there was no significant influence of the order of completing the two questionnaires on overall self-attributions, other attributions, or the difference between the two (FAE score). In fact this experiment did not discover any Fundamental Attribution Error at all; as mean attribution ratings for oneself ($M = 2.39$) were on average indicating internal attributions, but they did not differ significantly from attribution ratings for another person ($M = 2.30$), and the mean ratings were highly correlated with each other, $r(59) = .72$, $p < .01$. Neither did the desirability of the outcome have many significant effects on attributions for oneself or another person. There was neither a significant difference between positive- and negative-outcome scenarios for oneself on attribution ratings, nor for attributions for another person. The only difference between predictions for oneself and another person lay in positive-outcome scenarios, which tended to be attributed as more due to internal factors for the other person and more due to both internal and external factors for oneself ($M = 2.21$ vs. $M = 2.41$), $t(59) = 2.04$, $p < .05$.

The data in Appendix K show that some subjects attributed more internally for themselves, and others attributed more internally for the other person, resulting in no overall mean FAE bias being discovered. However the direction of the bias for each subject may relate to other individual differences. Indeed, a significant negative correlation between the subject's FAE score and their confidence on the general knowledge test was found, $r(59) = -.29$, $p < .05$. Since the individual's FAE score did not correlate significantly with their accuracy on the general knowledge test, but it did with confidence, this resulted in a significant correlation between the person's FAE score and overconfidence, $r(59) = -.31$, $p < .05$. There were significant differences between males and females when attributing causality to events on the FAE test. Male subjects rated the scenarios more internally than female subjects for both themselves, ($M = 2.20$ vs. $M = 2.57$), $t(57) = 2.19$, $p < .05$, and for another person ($M = 2.07$ vs. $M = 2.48$), $t(57) = 2.60$, $p < .01$. It may be the case that female subjects opted for the neutral position on the scale more, and perceived the scenarios as more due to both internal and external factors.

Probabilistic Reasoning

The mean number of runs produced was 1.54 ($SD = 1.74$) and the mean number of correct answers on the test of probabilistic reasoning was 2.40 ($SD = .69$) out of 4. No significant relationships between probabilistic reasoning ability and either overconfidence, field dependence, or the FAE score was found.

Discussion

The hypothesis that field-dependent and field-independent subjects would have different levels of accuracy and confidence in their judgements was not supported. Neither was there any relationship between field dependence and overconfidence. Subjects who were the most accurate on the general knowledge test tended to have lower overconfidence, and the subjects with the highest confidence were the most overconfident, but no relationship between overconfidence and cognitive style (field dependence) or probabilistic reasoning ability was found. As no significant relationship between scores on the Embedded Figures Test and accuracy, confidence, and overconfidence was found, it can be concluded that there is no evidence in the results of this experiment that the cognitive style of field dependence has any influence on overconfidence in one's general knowledge judgements.

Subjects who scored more highly on the Embedded Words Test tended to have lower accuracy scores on the general knowledge test, that is, they knew significantly less ($r = -.31$). One explanation is that the EWT measured field dependence and that those subjects who were more field-independent were less accurate on the general knowledge test. There was, however, no significant correlation between accuracy and scores on the EFT, which indicates that perhaps some other factor, such as the verbal component, of the EWT was responsible for the negative correlation with accuracy, rather than an effect of field dependence. The general knowledge test was numerically based, which tended to favour males over females in terms of accuracy, and thus the correlation may be explained if those subjects who were more numerically able scored well on the general knowledge test, and those people who were more verbally able scored more highly on the EWT test.

The positive correlation between the Embedded Figures Test and the EWT indicates that if a person is good at one they also tend to be good at the other. The EFT test has already been established as a test of field dependence and it may be the case that the EWT also taps the same underlying field independence skills. A problem with the EWT test, as administered in this chapter, is that as well as testing field dependence it also tests people's knowledge of the English language and vocabulary. Thus, different levels of education and familiarity with words may confound the results; hiding the true level of field dependence/ independence. However, if the words used are simple/common enough for everyone to recognise then this problem could be reduced. For example, the word EDUCATION has the simple words A, I , ON, CAT, AT which are familiar to everyone. It may be useful to extend the instructions on the embedded word test, to include all words which subjects can see/disembedd but which may not be real English words, as long as they follow the rules of English language construction. For example, the word VOCABULARY contains the words: voc, bul, lary, cabul, which are not English, but could be. Subjects whose first language is not English would still be at a disadvantage though, and the test should really only be used with the subject's native language.

In the EWT the word "I" stood out as an exception -- some people saw it as a word and reported it, and others never reported it. The EWT used uppercase letters when presenting the words and thus a translation of i to I was not the cause of non-reporting. It could be argued that only the most field-independent people would report a single letter as a word, as field-dependent people would not be able to distinguish it from the field of other letters. Perhaps a simple classification of subjects into those who can and cannot see the letter I or A as a word would indicate how field-dependent they were. It could be argued that the subjects who reported I as a word were very field-independent as they could break down the large word into its tiniest constituent parts, which field-dependent people could not do. If field-independent people can see the parts of words more easily than field-dependent people this may make learning to spell easier for them, as breaking the word into smaller parts may make spelling it easier. Further research could investigate this possibility and its applications to education.

There were small but significant negative correlations (-.31) between the FAE score and confidence and overconfidence on the general knowledge test. This indicates that those subjects who showed more overconfidence (unjustified confidence) in their general knowledge also showed a lower FAE score. The mean FAE score was not significant, indicating that subjects did not show a bias for underestimating the impact of situational factors on other people relative to themselves. Subjects rated the scenarios, on average, to be under the control of the person. Overconfidence in subjects was not significantly correlated with the mean attribution rating score for either oneself or another person. Thus it seems that people who see themselves and other people as equally under the control of situational factors, or who see the other person's behaviour as more due to situational factors than their own behaviour (negative FAE score) were more confident and overconfident. Those who saw their behaviour as more internally controlled relative to other people were less confident and overconfident. Thus the ability to perceive the influence of situational factors on behaviour is slightly related to the person's overconfidence in judgements. The cause of this relationship is unclear and further research is required.

The experiment found that female subjects tended to rate the FAE scenarios as more due to both internal and external factors than the male subjects. This may be because female subjects see the situation as more responsible in causing events to happen, and they may have less perceived control over the situation. Previous research has not found, or has not reported, consistent sex differences in the Fundamental Attribution Error.

The scores on the test of probabilistic reasoning did not significantly correlate with any other variable measured, indicating that either there is no relationship between probabilistic reasoning ability and overconfidence and field dependence, or that the test which was used was not sensitive enough to discriminate between different levels of probabilistic reasoning ability. A longer and more sensitive test should be used in further experiments. It may, however, be the case that people who do have a good understanding of the rules of probability only understand the theory of probability and that this does not influence the way that they actually perceive and estimate subjective probabilities. A comparison of statisticians' and lay people's overconfidence would be valuable to test this hypothesis.

CHAPTER 4

Individual Differences in Overconfidence for Self-Predictions of Future Events

Self-predictions of the occurrence of future events were made and associated levels of confidence for those judgements compared with actual outcomes to measure the degree of realism of confidence (overconfidence). Confident subjects were no more or less accurate than less confident subjects, but were more overconfident. Those subjects who were more accurate, and found the task easier, were less overconfident. Subjects who could accurately estimate the true base rate were better calibrated. Science students were more overconfident than arts students and male subjects showed significantly better resolution of confidence and accuracy, as measured by the correlation between the two, than female subjects. Subjects who were overconfident for positive events tended also to be overconfident for negative events. Large differences in subjects' overall overconfidence were found, but no significant relationships between self-esteem, field dependence, probabilistic reasoning ability, and overconfidence in self-predictions were found.

Individual Differences in Overconfidence for Self-Predictions of Future Events

According to the literature, overconfidence occurs when a person's confidence exceeds their accuracy when making judgements. If, however, the reverse is true and accuracy exceeds confidence then the person is being underconfident in their judgements. The history of overconfidence research is embedded in how people understand and use probability, and early research in this field generally used general knowledge type questions. Recent research has, however, spread overconfidence research into the realms of social judgement, including predicting how other people, or we ourselves, may behave in future situations (Dunning, Griffin, Milojkovic, & Ross, 1990; Vallone, Griffin, Lin, & Ross, 1990). This latter area is the topic of this chapter.

Self-Esteem

Engaging in social behaviour can have significant effects on people's self-esteem, as the comments other people make and the way they behave can influence how positively we perceive ourselves, but we do have coping mechanisms to protect self-esteem in the social world. The measurement of self-esteem by self-ideal discrepancy scores can be traced back to William James (1890), and early research using this measure found that as psychiatric patients' self-esteem improved through psychotherapy their self-ideal discrepancy score also decreased (Rogers & Dymond, 1954). Research by Baumgardner (1990) and Campbell (1990) indicated that people with low self-esteem show less self-knowledge. This was backed up by Gould (1993) who used a measure of self-ideal discrepancy, and showed that people with large self-ideal discrepancies engaged in less self-monitoring than those with lower scores. Predictions about one's future behaviour require self-knowledge and if this is lacking in subjects with low self-esteem, then poorer accuracy will surely result. McFarlin (1985) showed that those people with high self-esteem were able to use information about a task, to determine the level of effort required, more accurately than people with low self-esteem. It is predicted that subjects who are high in self-esteem will benefit from better self-knowledge and will have higher accuracy scores. High self-esteem subjects may also show higher confidence because of their exalted opinions about their own judgement accuracy, and thus be more overconfident than subjects with lower self-esteem.

Some degree of overconfidence may have benefits, as Taylor and Brown (1988) suggested that positive cognitive illusions can protect one's self-image, and illusions are associated with higher self-esteem, better mental health and balanced adjustment. Illusions may be beneficial for mental health, but may also lead to errors of judgement because they may bias people's ability to accurately assess and predict their abilities or future performance. Distortions of perception and judgement which protect self-esteem will probably also be associated with high confidence and thus overconfidence. Thus the hypothesis of this chapter is that there will be a positive correlation between self-esteem and overconfidence. The positivity or negativity (valence) of the outcome of the event being predicted by the subjects will also be considered when relating self-esteem to overconfidence. This is because subjects with low self-esteem may be more overconfident about negative events and less so about positive events than those subjects with high self-esteem. If this occurs it may be due to higher pessimism about bad events occurring, in those subjects with lower self-esteem, which raises their confidence that the event will occur.

Rationale for the experiment

Protection of self-esteem is one possible cause of overconfidence but other factors, such as cognitive styles, as described in the introduction to Chapter 3, may also influence subjects' overconfidence levels. The aim of the experiment described below is to see whether all people are overconfident in social judgements and, if differences in overconfidence exist, what factors may be linked with them. Another aim is to see whether differing levels of over/underconfidence are related to types of people, for example arts/science students, male/female subjects.

This experiment will also look at each subject's resolution of their judgements. Resolution "refers to the judge's ability to discriminate correct from incorrect judgments by differentially assigning confidence judgments to accurate and inaccurate judgments" (Sharp, Cutler, & Penrod, 1988, p. 272). The accuracy-confidence relationship can be examined to see how well resolved a subject is across a range of judgements. Previous research has tended to find low, but positive, relationships between confidence and accuracy within subjects using correlation tests. A high correlation value for a person would show that high confidence is associated with higher accuracy and vice versa, which is good, however this does not mean that the person is necessarily well calibrated as they may still be over- or underconfident. For example, a person may always be 20% more confident than their accuracy warrants and be well resolved; high confidence resulting in high accuracy and low confidence in low accuracy but overconfidence still being present.

Data from Chapter 2 will be used and analysed in this chapter using a within-subjects design which is relevant to individual differences, this angle was not examined in Chapter 2 because it more logically belongs in this chapter. The probabilistic reasoning test and tests of field dependence used in Chapter 3 will be re-used to examine their relationship with subjects' overconfidence in self-predictions.

Method

Subjects

The experiment used 98 students as subjects, (29 males and 69 females) with a mean age of 20.65 years (range 18 to 43 years). All of the subjects in this study were first-year psychology undergraduates at the University of Leicester who completed the experiment as part of their course requirement. Of these subjects, who had all taken part in the experiment reported in Chapter 2, 35 out of the initial 98 subjects volunteered to complete further tests. The mean age of this sub-group was 20.63 years and there were 7 male and 28 female subjects.

Design

A within-subjects design was employed. In this chapter the data collected in Chapter 2 shall be analysed further, by subjects instead of by questions, to look at individual differences in self-prediction of future behaviours.

Materials

The materials and procedure for the first part of the experiment were described in Chapter 2. In addition, the Embedded Figures Test, an Embedded Words Test, a probabilistic reasoning test (all previously described in Chapter 3, and see Appendix J) and a test of self-esteem were completed by 35 of the original 98 subjects. The self-esteem test was a questionnaire comparing ratings of one's self and one's ideal self on 32 self-descriptive statements (see Appendix M) including for example: "persistent", "avoids arguments", "articulate" etc. This test was developed as a test of self-esteem by Dr A. M. Colman in the Psychology Department of the University of Leicester. The basis of the list was formed from items most frequently used by students when describing themselves and others in an experimental group discussion, and was thus specific to the age group and student population.

Procedure

The procedure for the first three tests was identical to that used in Chapter 3. For the self-esteem test subjects were asked to read through a list of 32 self-descriptions and to rate each one on a scale of 0 (*Doesn't describe me at all*) to 100 (*Perfectly describes me now*). After completing the ratings for themselves subjects re-rated all of the items for, and whilst imagining, their ideal-self as they would like to be. The absolute discrepancy between the score for self and ideal-self for each item was calculated and the final score for the self-esteem questionnaire was the mean of these discrepancy scores. The higher the self-esteem test score the larger the discrepancy between self and ideal self and thus arguably lower self-esteem.

$$Self\ \text{-}\ esteem = \sum \sqrt{(Self - Ideal)^2}$$

The scores on these four tests were analysed by correlating them with the levels of overconfidence and other variables measured in the first part of the experiment (see Chapter 2 for methodology) to determine if any relationships existed between them. Analysis of variance tests were also carried out to see if differences existed between subjects who had high, medium and low levels of field dependence, probabilistic reasoning, and self-esteem, with respect to confidence, accuracy, overconfidence, predictions, base rate, accuracy in estimating the base rate, and demographic factors. Subjects were asked to record if they were male of female, and whether they were registered for arts or science degree courses.

The data from Chapter 2 were further examined to discover relationships between how confident, accurate, and overconfident a subject was and how this related to their ability to estimate the base rate and how well resolved their judgements were (as measured for each subject over the set of predictions by the correlation between each confidence estimate and subsequent accuracy).

Results

Predictions

Subjects who predicted that they would experience a lot of positive events also predicted that a lot of negative events would also occur, $r(97) = .33$, $p < .01$. Subjects who predicted that a lot of events would occur did in fact experience more events later on, $r(97) = .67$, $p < .01$, (see Table 4.1), ($r(97) = .68$ for negative items and $r(97) = .62$ for positive items), thus subjects who predicted the occurrence of fewer events went on to experience fewer. This resulted in a high accuracy rate of 82.10% of predictions coming true.

Subjects who predicted that more events would occur to them tended to have significantly lower overall confidence, $r(97) = -.38$, $p < .01$, this being true for both positive and negative events, and correspondingly those subjects who predicted that fewer events would occur had higher confidence. There was, however, no significant relationship between how many events a subject predicted would occur overall and their accuracy in those predictions.

Accuracy and Confidence

There was no significant relationship between subjects' mean level of confidence and their mean accuracy over all 87 predictions, $r(97) = .02$, thus more confident subjects were neither more or less accurate than less confident subjects (see Figure 4.1), and those who were more accurate were no more or less confident. No relationship between confidence and accuracy could be seen, even when the subject's own base rate was statistically controlled for. Subjects who differed in their base rates and their estimates of the base rate did not significantly differ in mean confidence, and the overall level of a subject's accuracy did not correlate with the subject's own base rate or the mean number of events predicted to occur.

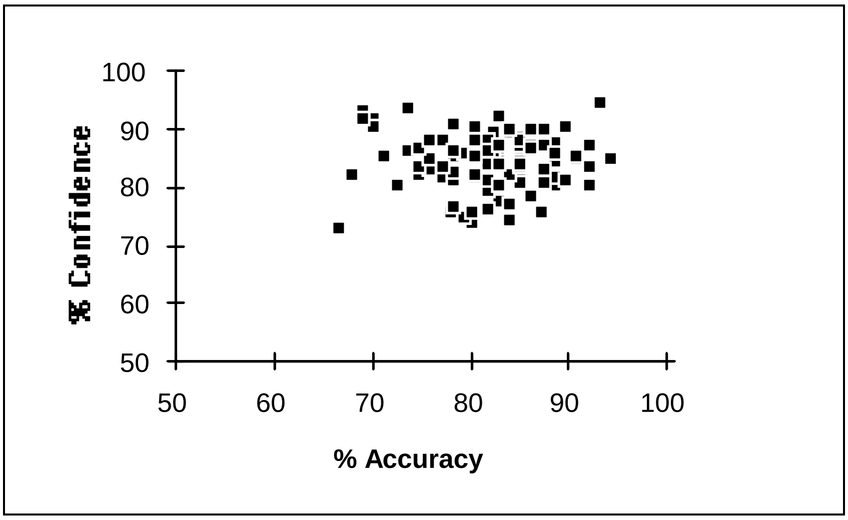

Figure 4.1. Overall confidence and accuracy for each subject.

Overconfidence

Significant correlations were found between a subject's accuracy and their mean overconfidence in self-predictions, $r(97) = -.79$, $p < .001$, and between mean confidence and overconfidence, $r(97) = .62$, $p < .001$, which was expected because, by definition, overconfidence is equal to confidence minus accuracy. More confident subjects were also more overconfident and, correspondingly, less confident subjects were less overconfident. These correlations show that accuracy was the main factor in determining a person's overconfidence, as it accounted for 62% of its variance, and confidence accounted for 38%. A negative correlation was found between the individual subject's mean base rate and their level of overconfidence, $r(97) = -.21$, $p < .05$, such that students who experienced fewer of the events were more overconfident and vice versa.

Subjects were significantly more overconfident for positive than negative items (as reported in Chapter 2) and there was a positive correlation between these two overall means, revealing that subjects who were more overconfident in positive-outcome predictions were significantly more overconfident for negative items as well, $r(97) = .49$, $p < .001$. Subjects who were confident about positive-outcome predictions tended also to be confident about the negative ones too, $r(97) = .68$, $p < .001$, and the directional estimate of the base rate was similar for both positive and negative items, $r(97)= .79$, $p < .01$. This indicates that people are responding in very similar ways for both positive- and negative-outcome predictions but the resulting accuracy for both types is less highly related, $r(97) = .26$, $p < .01$, but still significant. Thus subjects who were more accurate at positive-outcome predictions were also significantly more accurate at negative-outcome predictions.

Individual Differences

Table 4.2 shows the percentage of subjects in this experiment who exhibited overall under- or overconfidence or who were accurate, divided into quartile groups. There is a wide variety of overconfidence levels between people. Analysis of variance was used to compare the four quartile groups, divided up according to their final level of over- or underconfidence into four ranges:

1 = < -4%, 2 = -3% to 1%, 3 = 2% to 6%, 4 = > 7%. Subjects who were underconfident in self-predictions (ranges 1 and 2) were more accurate in those predictions than subjects in ranges 3 and 4, as Figure 4.2 shows, and subjects in range 3 were more accurate than those in range 4, $F(3,96) = 41.42$, $p < .001$, Tukey-HSD $p < .05$.

Table 4.2

Numbers of Overconfident, Underconfident, and Accurate Subjects and their Associated Levels of Confidence, Accuracy and Estimates of the Base Rate

	Underconfident		Accurate	Overconfident	***F***
Range	1	2	3	4	
Level of Over- or Underconfidence	< -4	-3 to 1	2 to 6	> 7	
Number of Subjects	24	25	24	24	
Predicted	54.60	50.31	49.88	51.67	1.36
Base Rate	50.65	47.68	45.32	46.55	1.46
Accuracy	86.90	85.18	81.11	75.08	41.42*
Confidence	80.17	84.42	85.69	87.56	15.73*
Abs Error of Estimated Base Rate	19.34	17.79	19.56	19.26	1.36
Dir Error of Estimated Base Rate	-0.61	-2.06	-4.79	-5.35	1.41

*$p < .001$

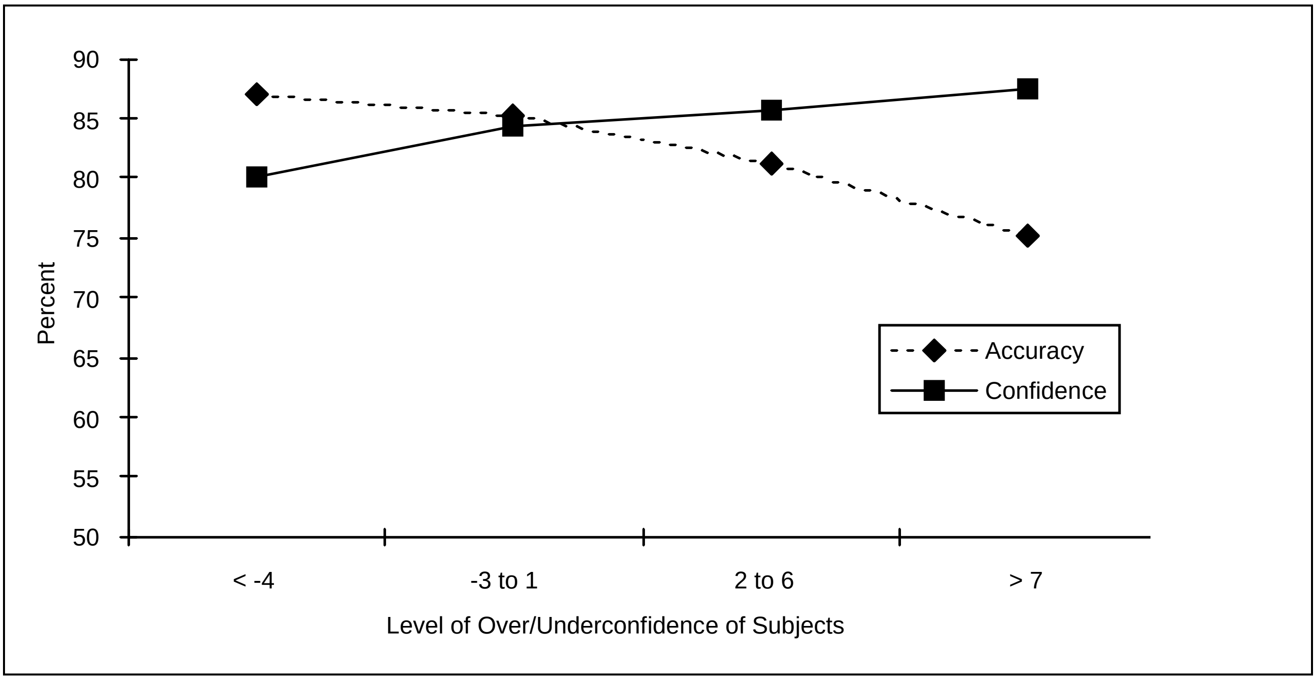

Figure 4.2. Subjects' levels of overconfidence plotted against accuracy, confidence, and estimated base rate.

Table 4.1

Means and Pearson r Correlation Coefficients of Variables in Self-Prediction Task

			Correlation With							
	Mean	*SD*	1	2	3	4	5	6	7	8
1 Accuracy	82.10_a	6.03	1.00	-.02	-.79**	.17	-.07	-.05	.19	-.21*
2 Confidence	84.46_a	4.69		1.00	.62**	-.14	-.38**	-.04	-.14	-.11
3 Overconfidence	2.36_a	7.71			1.00	-.21*	-.18	.01	-.23*	.10
4 Base Rate	47.55_a	9.30				1.00	.67**	-.20	.40**	-.15
5 Prediction	51.60_a	9.03					1.00	.01	.32**	-.04
6 Acc/Con Correlation	$.31_a$	.12						1.00	-.09	-.06
7 Estimate of Base Rate Directional Error	-3.19_a	9.36							1.00	-.17
8 Estimate of Base Rate Absolute Error	18.98_a	3.46								1.00

a - Averaged across 97 subjects

* $p < .05$ ** $p < .01$

Table 4.3

Means and Pearson r Correlation Coefficients between Measures of Individual Differences and Overconfidence

			Correlation With						
			Arts/	Embedded		Probability		Self-	
	Mean	*SD*	Science	Figures	Words	Score	Runs	Esteem	Sex
Accuracy	82.10_a	6.03	-.15	.05	-.07	-.22	-.33	.09	.07
Confidence	84.46_a	4.69	.20*	.08	.06	-.10	.03	.22	-.10
Overconfidence	2.36_a	7.71	.24*	.02	.09	.10	.28	.08	-.12
Acc-Con Correlation	$.31_a$	.12	-.01	-.33	-.20	-.07	-.30	-.22	-.20*

a - Averaged across 97 subjects

* $p < .05$

Subjects in the four ranges also differed in confidence, range 1 (the underconfident subjects) showed significantly lower confidence than the other three ranges and range 2 had lower confidence than range 4 $F(3,96) = 15.73$, $p < .001$, Tukey-HSD $p < .05$. Thus the difference in overconfidence between people does not seem to be wholly attributable to only one factor, such as an increased level of confidence for some people, but is an interaction of both confidence and accuracy.

The subjects who showed slight overconfidence, that is, those in range 3, were found to have significantly higher accuracy/confidence-correlations in their predictions (Mean $r = .36$) than subjects in range 1 who were underconfident (Mean $r = .27$), that is, they showed better resolution, $F(3,96) = 2.69$, $p < .05$, Tukey-HSD $p < .05$, with the other two groups having intermediate mean scores of $r = .31$ and $r = .30$. There was no significant difference between the different ranges in how accurate the subjects were at estimating the base rate.

The overall level of overconfidence in this experiment, across all subjects, was 2.4% (confidence = 84.46% accuracy = 82.09%, $t(96) = 3.01$, $p < .01$). Most previous studies found *over*confidence because the mean of all the subjects is taken, but there are large differences between people in over/underconfidence as Figure 4.3 shows. Figure 4.3 shows that the scores are negatively skewed which results in the mean score falling on the overconfidence side of the range. The mean overconfidence score is usually reported because it shows that there is a general bias operating in the population, but this hides the fact that many people have highly appropriate confidence in their judgements, and some are underconfident. In this experiment 37% of students were found to have virtually no overconfidence (to within $\pm$3%) in their judgements. Just under a quarter of subjects were found to be underconfident.

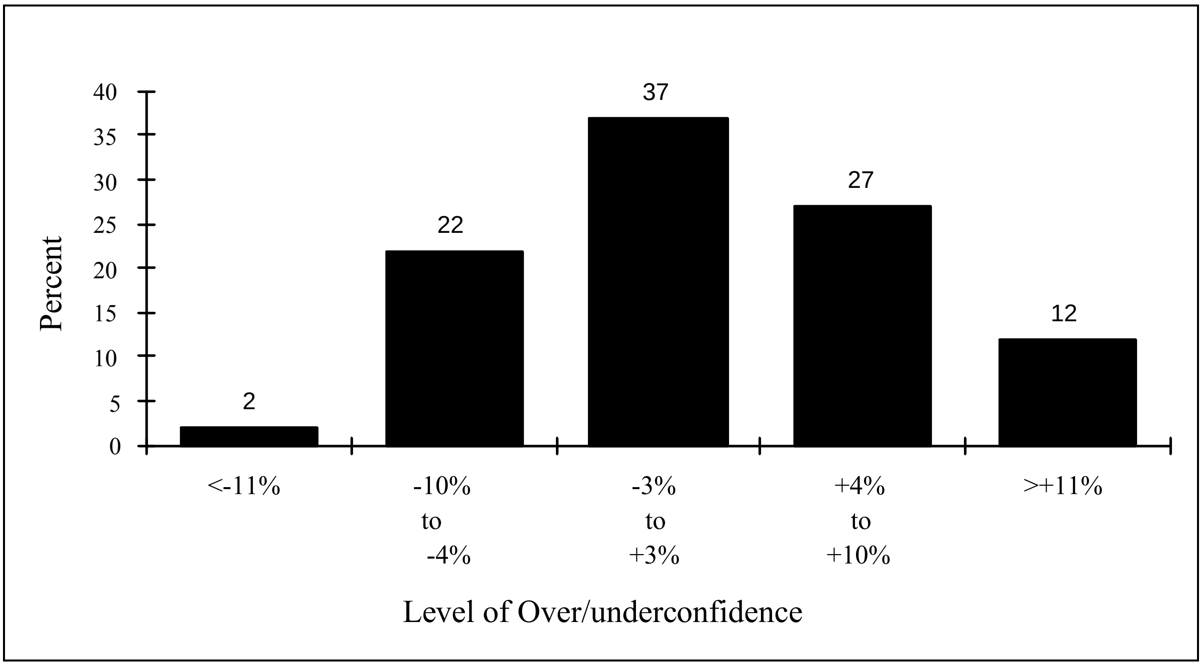

Figure 4.3. Percentages of subjects who were over/underconfident.

Base Rate

Subjects who experienced more events (higher base rates) were significantly less overconfident, $r(97) = -.21$, $p < .05$. There was however no relationship between a subject's own base rate and the *absolute* level of overconfidence that they had; showing that experiencing more or fewer events did not have any effect on the magnitude of the over- or underconfidence, but only on the direction of the error.

The base rate for positive events correlated $r(97) = .32$, $p < .01$ with the base rate for negative events, so subjects who experienced more positive events also experienced more negative events. The subject's own base rate did not correlate significantly with subject's mean level of confidence, even when outcome desirability was considered. Neither did the base rate correlate with overall accuracy, but accuracy for negative events significantly correlated with the base rate for negative events, $r(97) = .32$, $p < .01$.

There was a positive correlation between the base rate and score on the Embedded Words Test, $r(35) = .36$, $p < .05$, probably because older subjects had a lower base rate and also scored less well on the Embedded Words Test. The base rate had no relationship with self-esteem, sex, art/science status of the subjects, the Embedded Figures Test score, accuracy/confidence correlation, or probabilistic reasoning ability.

Estimates of the Base Rate

The absolute error of estimating the base rate is defined irrespective of the direction of the error and shows the magnitude of the error, whereas the directional error indicates if the subject's mean error was an over- or underestimate. The absolute error of estimating the base rate can theoretically equal the directional error if a subject always either over- or underestimates the base rate so that no balancing out of over- and underestimating occurs. The higher the value of an absolute error estimate the larger the discrepancy and the worse the subject is at estimating the base rate.

There was a significant positive correlation between the number of events a subject predicted would occur and how accurate their estimate of the base rate was, $r(97) = .32$, $p < .01$. This was a directional measure of accuracy which reveals that people who predicted (and subsequently tended to experience) that more events would occur also tended to estimate the base rate to be higher than those subjects who predicted and experienced fewer events; who also estimated the base rate to be lower. There was no relationship with absolute error of estimation so the number of events experienced did not influence the absolute magnitude of the error only its direction.

A significant negative correlation between subjects' directional error at estimating base rates and overconfidence, $r(97) = -.23$, $p < .05$, was found. Thus the lower they estimated the base rate to be, the higher they were in overconfidence and vice versa. This corresponds with the observation that a subject's estimate of the base rate for their peers significantly positively correlated with their own individual mean base rate, $r(97) = .40$, $p < .01$, indicating that subjects presumed other students to have similar base rates to their own. Thus subjects with low base rates estimate the base rate to be low and thus become overconfident. Subjects with a high base rate estimated the base rate to be high too, and became less overconfident.

There was no relationship between a subject's accuracy, confidence, accuracy/ confidence-correlation and the directional error of estimating the base rate. The *absolute error* of base rate estimates was found to be significantly related to the subject's overall accuracy level, $r(97) = -.21$, $p < .05$. That is, the most accurate subjects were also the ones who were more accurate in their estimates of the base rate (the ones who scored lowest in terms of discrepancy).

The absolute and directional measures of estimating the base rate were correlated significantly, $r(97) = -.23$, $p < .05$, which indicates that subjects who showed the least discrepancy in estimating the base rate tended to be the ones who scored high on the directional measure; they were overestimating the base rate. This shows that people who overestimated the base rate were less in error than people who underestimated the base rate.

The subject's absolute error in estimating the base rate significantly correlated with their measure of absolute overconfidence, $r(97) = .28$, $p < .05$. Thus people who were better at estimating the base rate were less over/underconfident and more well calibrated. Those subjects who were accurate estimators of the base rate were generally also more accurate in their predictions, $r(97) = -.21$, $p < .05$. The directional error of estimating the base rate correlated negatively with overconfidence, $r(97) = -.23$, $p < .05$, showing that people who overestimate the base rate for the group tend to be the ones who are underconfident and those who underestimate the base rate are overconfident.

A subject's overconfidence was not significantly correlated with their absolute error of estimating the base rate. Neither was the directional error of estimating the base rate significantly correlated with the absolute level of over-confidence, accuracy or confidence. The base rate positively correlated with the directional error of estimating the base rate, $r(97) = .40$, $p < .01$, but not with the absolute error of estimating the base rate. Thus people who overestimated the base rate were ones who had high base rate themselves, but high base rate people were not more likely to be any higher than low base rate people in the absolute degree of error in their base rate estimates.

With respect to the positivity/negativity of the questions subjects were better at estimating the base rate for positive items (absolute error of estimate $M = 17.94$) than for negative items ($M = 20.47$), $t(96) = 6.50$, $p < .001$.

Accuracy/Confidence Correlation

The accuracy-confidence correlation coefficient (resolution) for each subject was calculated by correlating the subject's accuracy in each prediction with their confidence in it. The mean overconfidence score is a measure of something slightly different to this measure as, for example, one subject could have 100 percent confidence in their first ten judgements which turn out to be wrong and 0 percent confidence in the next ten, which turn out to be correct. Another subject could have 0 percent confidence in ten incorrect answers and 100 percent confidence for each of ten correct answers. Both subjects would have 50 percent confidence and 50 percent accuracy mean scores and look perfectly well calibrated on average. However it is obvious that there is no correspondence between confidence and accuracy for the first subject and this is revealed by a correlation of $r = -1.00$. The second subject however would score $r = 1.00$ as confidence scores always predict the outcome and the person shows perfect resolution. The accuracy/confidence-correlation reveals whether a person's confidence generally reflects their accuracy in a judgement. The mean value of all the subjects was $r = .31$ ($SD = .12$) with a range of -.06 to .55.

A comparison was made between the subjects who were most highly resolved (mean $r = .45$, $N = 23$) compared with those who were least resolved (mean $r = .16$, $N = 24$) (top and bottom quartiles). These two groups did not differ significantly from each other with respect to how accurate they were at estimating the base rate, age, gender, arts/science, self-esteem, probabilistic reasoning, overconfidence, confidence, or predictions.

There was also no difference between the two groups on the Embedded Words Test, but the least well resolved subjects scored significantly higher on the Embedded Figures Test ($M = 16.3$ vs. $M = 9.67$) and were thus more field independent, $t(17) = 2.69$, $p < .05$.

The most highly resolved group were actually less accurate in their predictions for negative questions ($M = 76.28$ vs. $M = 81.30$), than the poorly resolved group, $t(45) = 2.02$, $p < .05$. This was related to the base rate for negative questions which was higher for poorly resolved than well resolved subjects ($M = 45.51$ vs. $M = 36.70$), $t(45) = 2.13$, $p < .05$. No difference between the two groups existed for positive questions, or over all the questions combined.

The accuracy/confidence correlation was negatively related to the absolute overconfidence of the subjects, $r(97) = -.26$, $p < .01$ (see Table 4.3). Thus the more over/underconfident the subjects were, the lower their accuracy/confidence correlation was. Subjects who are poorly resolved are thus more likely to be over- or underconfident rather than well calibrated. In addition to this, there was no relation between the accuracy/confidence correlation and confidence, thus indicating that subjects did not seem to realise that they were poorly resolved and thus did not reduce their confidence and thus only the magnitude or their overconfidence and not the direction was influenced.

Embedded Figures Test

There was no significant relationship between a subject's score on the Embedded Figures Test and any other variable. A correlation between scores on the Embedded Figures Test and Embedded Words Test was predicted because both should measure field dependence, but no relationship was found, and no curvilinear relationship was apparent. Subjects were divided into three tertiles, to compare those who did well on the Embedded Figures Test ($M = 19.67$) with those who did averagely ($M = 13.33$) and those who did poorly ($M = 7.43$). These three groups did not differ significantly from one another for any variable other than self-esteem, where the low scorers (field-dependent subjects) had significantly lower self-esteem ($M = 17.50$) than the middle scorers ($M = 26.40$), $F(2,34) = 3.23$, $p = .05$, HSD $p < .05$, but the high scorers did not differ from these two groups in self-esteem ($M = 21.75$).

Embedded Words Test

The only significant correlation between subjects' scores on the Embedded Words Test and any other variable was with the base rate, $r(35) = .36$, $p < .05$ (see Table 4.3) and more specifically with the negative base rate, $r(35) = .44$, $p < .05$) not the positive base rate ($r(35) = .17$). A one-way analysis of variance did not, however, find any difference between the subjects who scored at the top, middle or bottom of the scale on the Embedded Words Test, thus these correlations are probably spurious. Thus it seems that either there is no relationship between field dependence and overconfidence or that the Embedded Words Test does not measure field dependence or is not sensitive enough to measure it. A discussion about the Embedded Words Test can be found in Chapter 3.

Self-Esteem

The range of self-esteem scores was 7.97 to 42.97 with a mean score of 21.25 (*SD* 8.75). Higher scores indicate a larger self-ideal discrepancy and lower self-esteem. It was predicted that subjects with high self-esteem would have higher levels of overall confidence, but no evidence was found to support this hypothesis. Subjects with high self-esteem did not have significantly higher or lower overconfidence than subjects with low self-esteem. There was a relationship between a subject's level of self-esteem and their mean level of confidence for negative predictions, $r(35) = .35$, $p < .05$ (see Table 4.3), but no significant relationship for positive predictions. Thus subjects with higher self-esteem had lower confidence than those subjects with lower self-esteem when making predictions about negative events. This did not make these subjects have lower overconfidence, however, as self-esteem was not significantly correlated with any other measured variable (confidence, accuracy, overconfidence, or accuracy of estimating the base rate).

Probability Tests

Subjects who scored more runs on the probability test also had a higher absolute level of overconfidence (less well calibrated), $r(33) = .38$, $p < .05$, but an analysis of variance did not find a significant difference between groups with different scores on the runs test. The subject's level of probabilistic reasoning did not significantly relate to the subject's arts or science status, age, gender, or confidence. The only relationships observed were with accuracy and thus over-confidence, as the more runs a subject generated on the probability test the less accurate they were on predictions for positive events, $r(35) = .36$, $p < .05$, but they were not more confident on positive events so there was a resulting higher level of overconfidence on positive predictions, $r(35) = .35$, $p < .05$. There was no relationship between the number of runs generated and any measure of confidence or accuracy for negative predictions. The more runs scored, the better the understanding of probability.

The other measure of understanding of probability was the number of questions correct on the probability test, and this showed that subjects who had a high score (better understanding) had a significantly lower level of accuracy when predicting negative events, $r(35) = .42$, $p < .05$, but there was no significant relationship with any variable measured for positive events. There was no significant correlation between the number of questions answered correctly on the probability test and the number of runs generated during the production of random numbers ($r = -.11$). This perhaps raises doubts as to how good a measure of probabilistic reasoning these tests were .

The subjects who got one out of four questions correct on the probability test were significantly higher in self-esteem ($M = 29.79$) than subjects who scored two ($M = 18.33$), $F(2,34) = 4.69$, $p < .05$, HSD $p < .05$. Subjects who scored three correct did not differ from the other two groups in their confidence ($M = 21.37$).

Sex Differences

The sex of the subject had few discernible effects upon any of the variables measured, there being no significant differences between male and female subjects in terms of accuracy, confidence, overconfidence or accuracy in estimating the base rate and no interactions with the outcome positivity/ negativity. Neither did male and female-subjects differ significantly in the number of events that they predicted would occur ($M = 51.41$ and $M = 51.69$ respectively) or that did in fact occur (base rate $M = 47.05$ and $M = 47.77$ respectively), and again there was no interaction with the outcome positivity/ negativity of the event. There was no significant difference between the male and female subject groups in terms of age or the numbers of arts/science students.

No significant gender differences occurred in terms of field dependence, for either the Embedded Figures Test or Embedded Words Tests, or for self-esteem or probabilistic reasoning. The results of this experiment did, however, reveal that there was a gender difference in how well resolved the subjects were. Male subjects were more highly resolved, they had a higher accuracy-confidence correlation ($M = .34$) than female subjects ($M = .29$) on self-predictions of the type used in this experiment, $t(95) = 2.03$, $p < .05$. That is, male subjects confidence paralleled their resulting accuracy more so than for females.

Age Differences

The age of the subjects was found to be negatively correlated with the real base rate of events, $r(97) = -.31$, $p < .01$, which indicates that the older subjects were less likely to have experienced items on the questionnaire than younger subjects. This relationship between age and base rate was significant for positive events, $r(97) = -.37$, $p < .01$, but not for negative events, $r(97) = -.15$. This effect of age was predictable as the questionnaire was designed to be specific to first year students likely to be living in halls of residence, and thus mature students would be less likely to experience some events associated with halls, leaving home etc. Correspondingly, older subjects also predicted that fewer positive, $r(97) = .34$, $p < .01$, and negative, $r(97) = .28$, $p < .01$, events would occur, and thus there were no significant relationships between age and accuracy, confidence and overconfidence and neither were older subjects better or worse than younger ones at estimating the base rate.

Arts and Science Student Differences

Science students were found to be significantly more overconfident ($M = 4.05\%$) than arts students ($M = 0.32\%$) by on average 3.72%, $t(95) = 2.43$, $p < .05$ (Table 4.1). This difference between arts and science students in overconfidence occurred for both positive, ($M = 1.25\%$ vs. $M = 4.56\%$), $t(95) = 2.02$, $p < .05$, and negative items, ($M = -0.82\%$ vs. $M = 3.42\%$), $t(95) = 2.14$, $p < .05$ (see Table 4.4). In terms of the absolute level of overconfidence, irrespective of direction, arts students were significantly less in error than science students, (M = 4.87% vs. M = 7.40%), $t(95) = 2.52$, $p < .05$.

Table 4.4

Differences Between Positive-Outcome and Negative-Outcome Events in Predictions, Base Rates, Accuracy, Confidence, Overconfidence and Directional and Absolute Error of the Estimated Base Rate for Arts and Science Students

		Subjects		
Mean Percentage		Arts	Science	*t*
Predicted to Occur				
	All	53.80	49.78	2.23*
	Positive	57.61	54.81	1.44
	Negative	49.13	43.58	2.14*
Objective Base Rate				
	All	48.73	46.58	1.14
	Positive	51.70	51.44	0.13
	Negative	45.00	40.54	1.61
Confidence				
	All	83.43	85.31	2.00*
	Positive	85.84	87.36	1.59
	Negative	80.47	82.78	2.07*
Accuracy				
	All	83.11	81.26	1.51
	Positive	84.59	82.80	1.29
	Negative	81.28	79.36	1.10
Overconfidence				
	All	0.32	4.05	2.43*
	Positive	1.25	4.56	2.02*
	Negative	-0.82	3.42	2.14*
Estimated Base Rate: Directional error				
	All	-1.38	-4.69	1.75
	Positive	-0.89	-3.25	1.38
	Negative	-1.07	-5.43	1.82
Estimated Base Rate: Absolute error				
	All	18.50	19.37	1.23
	Positive	17.77	18.08	0.43
	Negative	19.66	21.16	1.66

df = 95 throughout

Between-subjects t-tests

* $p < .05$

The groups of arts and science students did not differ from each other with respect to age of participants or the gender proportions. For both groups subjects were more overconfident when making predictions about positive events, which they desired to occur.

Arts and science students were not significantly different in the accuracy of their predictions (M = 83.11 vs. M = 81.26), but science students were significantly more confident than arts students (M = 85.31 vs. M = 83.43 respectively), $t(95) = 2.00$, $p < 0.05$. There is an interaction between these two factors, science students are more confident and slightly less accurate than arts students, and this causes the science students to be significantly more overconfident than arts students.

No significant difference between arts and science students was found in terms of field dependence; both groups had similar scores on both the Embedded Figures Test (M = 12.78 vs. M = 13.53) and the Embedded Words Test (M = 45.50 vs. M = 42.82). Neither was there any difference between these two groups in terms of probabilistic reasoning: scores on the probability test were 1.94 for arts students and 2.35 for science students, although this difference was approaching a significant level, $t(33) = 1.80$, $p = .08$. The two groups did not differ in their levels of self-esteem either.

No significant arts/science difference was found for the objective base rate of the occurrence of events, accuracy of the predictions or estimates of the base rate and no interactions with outcome positivity/negativity (desirability) occurred with these factors. The outcome desirability did however make a difference with respect to confidence and overconfidence. Science students were more confident about negative event predictions (M = 82.78 vs. M = 80.47), $t(95) = 2.07$, $p = .05$, than arts students, but the groups did not differ for positive event predictions (M = 87.36 vs. M = 85.84 respectively).

Arts students predicted that significantly more events would occur to them than science students did (M = 53.80 vs. M = 49.78), $t(95) = 2.23$, $p = .05$. This was especially true for negative events (M = 49.13 vs. M = 43.58), $t(95) = 2.14$, $p = .05$, and did not significantly occur for positive events (M = 57.61 vs. M = 54.81). No significant difference occurred between the two groups in terms of the true base rate. There was no significant difference between arts and science students in terms of resolution of judgements, (accuracy-confidence correlations were M = .31 vs. M = .31).

Discussion

This experiment has attempted to discover what differences between people exist in overconfidence and what factors may be related to these differences. It was found that subjects who predict they will experience more events do in fact experience more events, and they also tend to believe that other people will also experience more events. These high predictors also have lower confidence but this does not make them any more accurate or better calibrated in their predictions. Subjects who predicted a lot of events would occur, and those who experienced a lot of events (who tended to be the same people), expected that other people would also experience more events, but this did not influence how accurate their estimates were, only the direction. Theories of unrealistic optimism would predict that subjects would overpredict the occurrence of positive-outcome events and underpredict the occurrence of negative-outcome events, but this experiment does not tend to support this idea as a positive correlation was found between the overall number of positive and negative events predicted to occur, indicating that some subjects perceive more events will happen to them irrespective of the type of outcome.

Overall it was found that subjects who were more accurate in their predictions were also less overconfident, and those who were more confident were more overconfident, because no significant relationship between confidence and accuracy was found. The results show that subjects who were less accurate in their estimates of the base rate for the group were less accurate in their self-predictions and less well calibrated (more over- or underconfident irrespective of direction) than subjects who were more accurate in their estimates of the base rate. This indicates that either having an ability to be more accurate in your estimates of the base rate for the group causes lower overconfidence, or some other ability causes both to be improved.

The relationship between resolution and overconfidence seems quite straightforward, subjects who were least over- or underconfident showed the best resolution. This may seem like an obvious statement, but it could have been the case that the least over/underconfident subjects (irrespective of direction) appeared less over/underconfident due to a balancing out of over- and underconfidence over the questions rather than subjects actually showing better resolution.

The finding that subjects with higher self-esteem tended to have lower confidence about negative judgements could be due to a tendency of those subjects to avoid thinking about negative events as a strategy to protect their self-esteem. Thus when they were forced to consider negative events in this questionnaire they were less confident. The relationship between self-esteem and overconfidence needs further investigation as theoretically there should be a link, which this experiment has not discovered. Another area also deserves further study, as those subjects who were better at probabilistic reasoning in this experiment were significantly less well calibrated than those who were worse at probabilistic reasoning. This contradicts the hypothesis that an understanding of probability or statistics would reduce over/underconfidence (improve calibration).

Science students in the current experiment were more overconfident about their predictions than arts students. This may reveal something about the nature of different types of people who study different disciplines. However, these differences may possibly be an artefact of the measurement system used. It may be the case that science students were more confident than arts students due to their familiarity with using numbers to express their confidence, or else may be more at home using the extremes of the measurement scale. However, looking at the standard deviations for both accuracy and confidence reveals that they are virtually identical for both arts and science students, so they do not appear to be using the scale differently. Alternatively, arts students may be less used to filling in questionnaires or being in experimental situations. This may make them more self-conscious and hence introspective, possible making their judgements with a more realistic level of confidence. Whatever the cause of the difference, the arts students ended up being the most accurate and least overconfident in their predictions. It may be possible to replicate this experiment using a non-numerical measure of overconfidence, such as marking a point on a scale, to see if this difference would still persist.

The amount of overconfidence found depends upon the questions asked and the population tested. Most studies in this area use undergraduates, who may be much more confident than the population as a whole, due to their higher level of education and achievement. There may also be cultural differences, and these possibilities deserve further research.

To conclude, people are all different and to talk only of overconfidence is to dismiss many underconfident judgements. It seems intuitively obvious that confidence varies over time and is situationally dependent. The boundaries of this varying level of confidence may be fixed at different levels for different people, however, and thus over a wide range of tests relationships with other individual differences may be found. This experiment did not find significant relationships with such factors as self-esteem, perhaps because many judgemental domains need to be measured to get an accurate measure of an individual's overconfidence. People may be overconfident in one task but not in another or high in both. Because self-esteem is built up of confidence in many aspects of oneself correlations between self-esteem and overconfidence may only be found when many judgemental tasks are considered together. Comparisons of people's overconfidence in different tasks would be interesting. A correlational study across types of judgements could determine whether some individuals are generally better judges than others. Better judgements may be related to introspectiveness, impulsivity/reflectivity and self-deceptiveness.

The higher levels of overconfidence found for positive items indicates that the type of prediction being made affects people's confidence. Factors such as optimism and hope may raise confidence and others may lower it. Are people who are generally overconfident life's optimists, conversely are underconfident people life's pessimists? How does confidence vary with mood, feedback about progress, physiological factors and so on? This is an area of psychology ready to be researched and applied. Clearly one of the most significant areas of application could be in the evaluation of decision making abilities of individuals in organisations. By studying the factors that lead to appropriate levels of confidence we could learn how to incorporate them into training. We also need to know how individual differences (for example self-esteem, personality and cognitive styles) affect overconfidence, if we are ever to be able to predict when it will occur and whether or not it is justified.

In conclusion, people who are most confident are not necessarily also high in accuracy, however, people high in accuracy tend to be lower in overconfidence and those high in confidence tend to be higher in overconfidence. Because mean accuracy and confidence show no significant relationship for subjects it can be concluded that those people who express the most confidence over a range of self-predictions need not be the most accurate in reality. It also seems that people who can more accurately estimate the base rate have higher accuracy in their own self-predictions. The causality of this relationship is unclear, however, either accuracy in self-predictions leads to a better judgement about the base rate or a good estimation of the base rate lends itself to improving the accuracy of judgements about oneself.

CHAPTER 5

Feedback and Item Difficulty Effects on Overconfidence

Immediate feedback of results was used in an attempt to de-bias overconfident subjects in a general knowledge test. In a 2 x 3 x 4 mixed factorial design (Feedback x Question Difficulty x Trial Blocks), the accuracy, confidence, and overconfidence of judgements of 150 subjects (48 males and 102 females) were measured. Hard questions produced significantly higher levels of overconfidence than medium-difficulty and easy questions, which in turn resulted in under-confidence. Combining all levels of difficulty, females were significantly less overconfident than males. No significant effect of external feedback was found, although better calibration in latter trial blocks for hard-level questions suggests that intrinsic feedback through self-monitoring occurred but was only effective in reducing the bias for hard questions.

Feedback and Item Difficulty Effects on Overconfidence

From the preceding chapters it has become apparent that some people show poor calibration of confidence in relation to their corresponding levels of accuracy in everyday judgements. For such people, confidence is markedly lower or higher than is warranted by the accuracy of their judgements, and such a mismatch between confidence and accuracy is a common and well-established phenomenon (Lichtenstein, Fischhoff, & Phillips, 1982). Several questions arise from this; for example, under what circumstances does poor calibration occur, how does this lack of calibration develop in the first place, and can people become better calibrated? If any attempt is to be made to improve calibration, it is important to understand how people become poorly calibrated, that is, how poor calibration develops. Thus we need to know what factors make some people better calibrated than others.

Confidence judgements can be examined in novel situations where few existing skills will transfer, or in new situations based on familiar themes where existing skills are relevant to the current task. The former situation may occur mostly in childhood, whereas in adulthood few judgement tasks are completely novel. I propose that when a new judgement task is encountered people's initial confidence in their ability to perform it is low and unstable, more open to revision than when more experience has been gained. The initial confidence level will depend upon whether people believe that they have skills from a related area that will transfer to the new task and provide some confidence in the new task. Confidence will change as experience reveals how capable the person is at performing the task: if feedback is positive and higher accuracy is being achieved, then confidence should increase, whereas if the feedback is negative, then confidence should decrease (or remain stable if the feedback is disregarded, for example to protect self-esteem). Poor calibration may be due to a lack of feedback during the relevant time span when the level of confidence in a task is forming, or alternatively feedback may be available but may be misinterpreted, or not used to alter confidence sufficiently. When it results in overconfidence, poor calibration shares many properties in common with unrealistic optimism (e.g., Harris & Middleton, 1994).

To test the robustness of judgemental biases, studies have been carried out that try to de-bias subjects by different means and specify the conditions under which the biases operate. Fischhoff (1982) pointed out that the root of a bias may lie in either the task performed, the people making the judgement, or a mismatch between the two. Biases may arise because the task may be inappropriate as subjects may not understand the instructions, may perceive the task in a different way from how the experimenter sees it, may not be able to explain their responses clearly, may not be motivated highly, or may use stereotypic answers to avoid thinking about each question.

Fischhoff (1982) therefore proposed several ways in which people may be de-biased: warning of the bias, describing the bias and its magnitude, providing subjects with feedback, and training subjects. As Fischhoff pointed out, however, even if a de-biasing technique is found to work, the underlying mechanisms still need to be understood, and if de-biasing can prove effective then why does it not occur in everyday life? There have been many experimental attempts to reduce overconfidence using these de-biasing techniques.

Attempts to De-Bias Via Feedback and Training

Comprehensive training in calibration and probability assessment was used by Lichtenstein and Fischhoff (1980), after testing sessions involving general knowledge type questions, and the training did result in improved calibration and reduced overconfidence, mainly after the first feedback session. This study showed that training can produce improvements in calibration for some people but not for others -- not, for example, for those who were initially well calibrated. The improved calibration was shown to generalise to some other tasks but not others, indicating that learning had occurred but with limited generalization. Some criticisms of this research seem justified, on the grounds, for example, that the range of subjects used was limited -- only 12 acquaintances of the experimenters. Also, the feedback employed at the end of each session of 200 general knowledge questions was very artificial and statistical in nature, which means that the findings are of debatable relevance to understanding how people learn to be better calibrated in everyday life where this type of training does not occur.

A general knowledge type test was used by Sharp, Cutler, and Penrod (1988) who, over a period of four weeks, gave subjects either a detailed statistical breakdown of their individual performance on a similar test a week before, or no feedback at all. No significant improvement in calibration or reduction in overconfidence resulted from this type of statistical feedback. Another unsuccessful attempt to reduce overconfidence was made, this time by Fischhoff and Slovic (1980), by warning subjects that the task that they were attempting -- to discriminate between the drawings of Asian and European children -- may be impossible. In this case considerable overconfidence was reported and only reduced by about 5% when the warning was given.

In an earlier investigation, however, Adams and Adams (1961) compared the calibration of two groups of subjects on a word judgement task. A feedback group was given feedback about their calibration and the control group was not given this information. The control group became slightly but not significantly less well calibrated (mean absolute discrepancy scores increased from 11.16 to 15.22), but the feedback group became significantly better calibrated (mean absolute discrepancy scores fell from 13.20 to 6.92).

To extend this result, Adams and Adams reported another experiment to see if improved calibration could be transferred from one task to another. Over five days, eight experimental subjects, who were given feedback about the calibration of their performance and eight control subjects carried out tasks such as general knowledge tests, weight discrimination tests and perception of dots and lines, and finally the word pair task from the first experiment. The feedback group were again significantly better calibrated than the control group (9.28 versus 13.61). The improvement in subjects' calibration, when information about calibration was given as feedback, was found to be transferable to some related tasks.

In the studies just mentioned, the experimenters tried to train subjects to be better calibrated by giving them information about their overall calibration of confidence or warning them of the bias. Alternative de-biasing techniques have also been used, such as where the correct answer is given to the subjects after every trial or group of trials or where subjects have been required to supply justification for answers given.

Arkes, Dawes, and Christensen (1986) suggested that high overconfidence in a task may result in corrective feedback being ignored when people believe that their accuracy is already high. Their later research, however, showed that corrective feedback, in the form of the correct answers, given to a highly overconfident group, did significantly improve the subjects' calibration (Arkes, Christensen, Lai, & Blumer, 1987). Arkes et al. manipulated the apparent difficulty of the practice session questions and found that the greatest reduction in confidence results from feedback that contradicted subjects' feelings the most -- subjects overcompensated and became underconfident when they believed they were doing well but were told that they were performing poorly. The subjects who believed that they were doing well, on the apparently easy questions, were initially highly confident, but feedback which revealed how poorly they had in fact done reduced their confidence, resulting in final underconfidence of -4.67% on the main set of questions compared to the no-feedback group's score of 9.13%. Subjects in the apparently hard condition already believed that their performance was poor and had lower overconfidence, and when given feedback reduced their overconfidence slightly to 4.36%, which was not significantly lower than the no-feedback easy (9.13%) or no-feedback hard (6.29%) groups. Telling subjects that they are doing badly when they were already aware of it did not have such a marked effect on confidence.

Humble, Keim, and Hershauer (1992) also achieved a reduction in overconfidence with feedback. The subjects' task was to repeatedly make a decision about which of six software packages to recommend, based on certain selection criteria with varying weighted importance. The experimental group received feedback as to whether their last decision was the optimal one. Subjects in the feedback group significantly raised their accuracy and lowered their confidence, thereby lowering their overconfidence in comparison with the control group. Overconfidence was reduced by feedback but still remained at a very high level of 40% compared to 56% without feedback. Humble, Keim, and Hershauer concluded that overconfidence is reduced by specific immediate feedback, but that the bias still remains high as a result of the experiences of one's lifetime.

Other methods of de-biasing subjects have shown some success. Koriat, Lichtenstein, and Fischhoff (1980) proposed that overconfidence may result from biases in information processing, such that recall from memory, either during or after the decision making process, tends to be biased towards evidence supporting a tentative answer rather than contradicting it. To test this hypothesis, attempts were made to de-bias subjects (that is, to reduce their overconfidence) by asking them to write down all the reasons they could think of why each of the multiple choice answers was right or wrong. This method resulted in subjects becoming very well calibrated for all levels of confidence, except the very highest, and a significant reduction in overconfidence resulted. The generation of pro/con reasons had no significant influence on subjects' accuracy levels, but it did lower confidence, resulting in reduced overconfidence. However, when the subjects were required to generate only one piece of supporting evidence or one supporting and one contradicting, this had no effect on calibration. Their results show that the important factor was the production of reasons that contradicted the subjects' chosen answers. This finding suggests that it is the act of generating reasons why the answer may be wrong that reduces overconfidence and is a partial remedy.

According to Koriat, Lichtenstein, and Fischhoff (1980), people search their memories to come up with an answer and then review the evidence and assess their confidence in that answer. These researchers concluded from their findings that debiasing techniques are likely to be most effective if they encourage the generation of contradicting evidence and suppress the generation of supporting evidence.

Some experimenters have found a natural improvement in calibration without the need for feedback from an external source. In tasks where subjects perform the same judgement task over and over again, or where the accuracy of their judgements is obvious to the subjects, self-feedback occurs almost inevitably. Understanding how self-feedback operates is obviously important because it occurs so frequently in everyday situations where feedback from other sources is unavailable.

In one of the earliest studies in this field, Adams and Adams (1961) found that subjects who were intentional learners showed less overconfidence than incidental learners (who were exposed to the same material but did not deliberately attempt to learn it), and improved their calibration with practice, over 16 trials, which incidental learners did not. This improved calibration occurred in the absence of specific external feedback about performance or calibration; it was a natural improvement. In this experiment the same task was repeated up to 16 times, so subjects presumably saw the test materials again and again and thus indirectly had feedback about their performance in the previous session. It seems that intentional learners were motivated to monitor their performance with the goal of improving it, and thus they gave themselves feedback which improved their calibration. Although the authors did not comment on this, a possible reason why the incidental learners did not show corresponding improvement may have been that they were not motivated enough to either monitor their performance or to modify their confidence. Motivation may play a crucial role in improving calibration of confidence ratings by providing internal feedback through the increased monitoring of performance.

Paese and Sniezek (1991) found that subjects' confidence in a task (of judging the future performance of professional baseball players) increased with practice without any feedback being supplied. Although no improvement in performance had in fact occurred subjects seemed to believe that they must have improved with practice and raised their confidence levels. As a result overconfidence increased with practice because the subjects internal feedback was incorrect. If subjects who do not receive feedback are unable to improve their performance or to lower their confidence through accurate self-feedback, then differences between the feedback and no-feedback groups are generally found in experiments.

Apart from accurate feedback some other factors may account for lack of calibration in confidence. According to Lichtenstein, Fischhoff, and Phillips (1982), there may be times when people are not motivated to be honest in their assessments of confidence, because situations may reward or punish honesty differentially. Thus subtle pressures to conform, impress, or deny may be strong reasons to be mis-calibrated in one's judgements. Arkes et al. (1987) found that subjects who anticipated a discussion of their responses with peers lengthened the amount of time that they spent making decisions and significantly reduced their overconfidence which suggests that motivational factors regarding self-presentation were in operation.

Social pressures may influence actual and portrayed confidence. Such pressures may have different effects on different people. For example, the level of confidence expressed by women compared with men appears to be situationally dependent, and the type of task being undertaken is also relevant (Lenney, 1977). According to Lenney, women are less self-confident and devalue their performance as compared with men when feedback about their performance is either ambiguous or absent.

Lenney (1977) asserts that when people are informed of norms or a pass grade prior to a task, then females' expectations tend to be lowered. She cites evidence for the pressure of social cues on women's self-confidence and concludes that situational factors, including the individuals or groups with whom self-comparisons are made, influence self-confidence in women much more than men. Another important factor that affects expressed confidence in predicting how well a task will be performed is whether the task is perceived as being congruent with one's sex role. Tasks that are presented as sex-role inappropriate result in lower expectations of success and lower confidence (Stein, Pohly, & Mueller, 1971).

Rationale for the experiment

The experiment described below was designed to investigate how the type of feedback available in everyday life can alter confidence and whether it can make the level of confidence more appropriate to the level of accuracy achieved. In previous experiments the effect of feedback may have been confounded with the level of difficulty of the questions, thus this experiment manipulates the level of difficulty to see whether feedback differentially affects overconfidence when the task difficulty varies. When the level of difficulty is constantly either very hard or very easy, subjects should learn quite quickly to adjust their confidence to an appropriate level. However, if the set is a mixture of hard, middling, and easy questions then feedback will be of little use in determining confidence for the next question of unknown difficulty. Many judgements in everyday life are of course of the latter type, and this is probably why learning to be accurately calibrated in confidence is difficult.

Accurate calibration is said to exist when people have a realistic understanding of how right they are or how well they are performing. Feedback should therefore reduce overconfidence only if it changes people's perception or beliefs about their performance. The greatest improvements in calibration should occur when feedback about performance is consistent, and this in turn should occur when the questions are consistently hard or easy, but when there is a mixture of difficulty levels the feedback about the difficulty of the task is necessarily ambiguous and calibration should not improve as much or at all.

Detailed statistical information on calibration, which has sometimes been given as feedback to train subjects to be better calibrated, seems very artificial and is not the type of information that people receive in everyday life. Thus the following experiment aims to be more true to life by using immediate right/wrong feedback about the accuracy of the subject's response to each question. With this form of feedback, subjects should either realise that they are answering many questions correctly, which should raise their confidence, or that they are getting many answers wrong, which should lower their confidence.

Two treatment conditions (feedback and no-feedback) are used. In the feedback condition the correct answer is revealed to the subjects after each of their answers. In this condition the accuracy level should remain stable, and the subjects' confidence levels should either increase or decrease, depending on how well or badly they are performing, thereby producing better calibration in the feedback group.

The effect of the presentation position of the questions (trial block position) is also examined to see whether subjects learn to reduce their confidence over time. It is expected that subjects will lower or raise their confidence estimates for the later questions after having received feedback on the accuracy of their answers in earlier trials. In the first trial block no difference in overconfidence should be found between the feedback and no-feedback conditions, because the subjects will not yet have received enough feedback. If feedback has an effect then the differences between the two conditions should increase over trial blocks. The level of over/underconfidence should remain stable across trial blocks for the no-feedback condition. Thus an interaction between feedback/no-feedback and trial block is predicted for overconfidence.

Method

Subjects

For the purpose of selecting items for the main study, a pilot study was conducted with 57 subjects (13 males and 44 females) drawn from the same pool of subjects as for the main study. The subjects in the pilot study were under-graduate and postgraduate students studying a variety of subjects. In the main study, the subjects were 150 undergraduate students (48 males and 102 females) with an average age of 20.69 years (range 18 to 48 years).

Design

A 2 x 3 x 4 mixed factorial design (Feedback x Question Difficulty x Trial Blocks) was used. The first factor, Feedback, was manipulated by giving some subjects feedback in the form of the correct answer to each of the questions immediately after they had answered, and others no feedback. The second factor, Question Difficulty, was varied across three levels: hard, medium, and easy. The last (within subjects) factor was trial block position over four levels: Trial Block 1 (the first 5 questions), Trial Block 2 (Questions 6 to 10), Trial Block 3 (Questions 11 to 15), and Trial Block 4 (Questions 16 to 20). The dependent variables were the levels of confidence, accuracy and overconfidence (where overconfidence = confidence - accuracy) of the subjects' judgements. Subjects were randomly assigned to each of the six treatment conditions (Feedback x Question Difficulty) with 15 females and 8 males in each condition.

Materials

For the pilot study, 60 questions were selected from the board game Trivial Pursuit. The criteria used to select the questions were that they should not be multiple choice, should have short one-word or two-word answers, and should not be subject to rapidly becoming out of date. A typical example of a question used is number B5: "What flavour is *Grand Marnier*?" (answer: orange).

Subjects completed the pilot study questionnaire at their own pace within a half-hour session. They were requested to give an answer for each of the 60 questions in turn, and to state their level of confidence in their answer, using any number between 0% ("No confidence") and 100% ("Total confidence"), as shown in the following scale:

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

No confidence -- Total confidence

Subjects were told that if they made a wild guess they should indicate that they had little confidence that their answer was right, and thus choose a number close to zero percent; if, however, they were confident that the answer was correct, then to choose a number nearer 100 percent. Subjects were asked to treat the experiment seriously even though it was light-hearted in format, and asked not to discuss the experiment or questions with other students who may do the experiment in future. The subjects' responses were anonymous and confidential, and they were informed of this before they began the test.

The pilot study was used to determine the level of difficulty of each of the questions (accuracy score), i.e. the base rate proportion of subjects who answered each question correctly (see Appendix Q for data). The questions were ranked in order of difficulty and then divided into three categories: hard (0-33% accuracy), medium (34-66%), and easy (67-100%) to make up questionnaires A, B, and C respectively. Thus twenty questions were used for each of these three categories and the final questionnaires can be found in Appendix P.

Procedure

In the main study, subjects were randomly assigned to one of the six treatment conditions and tested in small groups. Each of the 20 questions was read out and subjects were given 30 seconds to write down their answer and their level of confidence that the answer they had given was correct, using the rating scale and instructions described above, used in the pilot study (see Appendix P). In the feedback condition the correct answer was read out after the 30 seconds had elapsed, and the subjects marked their answer as either correct or incorrect before proceeding to the next question. In the no-feedback condition the correct answers were not given until after the subjects had responded to all of the questions, at which point all the correct answers were read out and subjects marked their own answer papers. At the end of the testing session subjects were debriefed minimally to avoid knowledge of the purpose of the experiment leaking into the student population. The subjects were told that a full description of the experiment and its findings would be displayed on a notice board after all the subjects had been tested, and any questions would be answered at that time.

Results

Manipulation Check

The subjects' mean accuracy levels (see Table 5.1) were 21.00% for hard questions, 57.53% for medium questions and 79.00% for easy questions, and these accuracy levels were all significantly different from each other, $F(2, 138) = 207.07$, $p = .001$, Tukey-HSD = 10.50, $p < .05$, which provides a manipulation check confirming the categorization of the questions into three levels of difficulty respectively.

Table 5.1

Mean Confidence, Accuracy, and Overconfidence for the Three Levels of Question Difficulty

	A Hard	B Medium	C Easy	*F*
Confidence	28.50	48.03	67.58	88.30*
Accuracy	21.00	57.53	79.00	207.07*
Overconfidence	7.50	-9.50	-11.42	38.04*

* $p < .001$

Question Difficulty

The obtained level of confidence for the three levels of difficulty were: hard (M = 28.50%), medium (M = 48.03%), and easy (M = 67.58%), which all differed significantly from one another, $F(2, 138) = 88.30$, $p = .001$, Tukey-HSD = 11.41, $p < .05$.

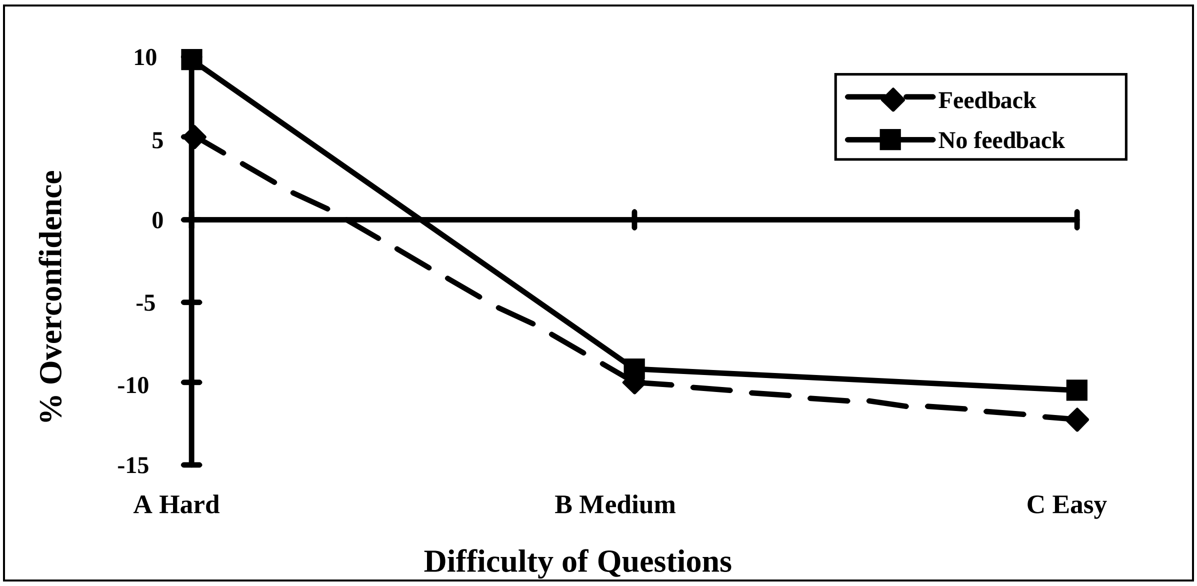

Figure 5.1. Levels of overconfidence by question difficulty and feedback/no-feedback.

The level of difficulty of the questions also had a predictable effect on overconfidence (see Table 5.1), F (2, 138) = 38.04, $p < .001$. A Tukey-HSD test showed that hard questions (M = 7.50%) resulted in significantly higher overconfidence than medium (M = -9.50%) and easy questions (M = -11.42%), Tukey-HSD = 8.49, $p < .05$ (see Figure 5.1). This is in line with previous research and is explained in terms of the *hard-easy effect*. Medium and easy questions did not differ significantly from each other in overconfidence. Overall, hard questions produced the most appropriate levels of confidence, being close to perfect calibration, although on the side of overconfidence, than medium-difficulty and easy questions which produced underconfidence.

Effects of Feedback

The mean overconfidence was -5.62% for the feedback group and -3.33% for the no-feedback group. This difference, though in the predicted direction, was non-significant (see Appendix R for all statistical calculations).

Irrespective of the level of difficulty, the experiment resulted in a significant overall level of underconfidence (*M* = -4.47%), (accuracy = 52.51% vs. confidence = 48.04%), due to the large number of easy questions, $t(149) = 3.74$, $p < .001$.

No significant differences were found between the feedback and the no-feedback groups for either accuracy (*M* = 53.22% vs. *M* = 51.80% respectively) or confidence (*M* = 47.60% vs. *M* = 48.47%). Thus, feedback had no significant effect on confidence, accuracy or overconfidence across all sixty questions combined. There were no significant interactions involving the factors feedback and question difficulty for accuracy, confidence or overconfidence.

Trial Blocks

Considering both the feedback and no-feedback groups together (bearing in mind that no significant differences were found between them) a small though significant effect of trial block on overconfidence was found, $F(3, 414) = 7.27$, $p < .001$ (see Table 5.2). A posteriori comparison, using the Tukey-HSD test, revealed that Trial Block 3 (*M* = -8.78%) yielded significantly lower overconfidence than Trial Blocks 1 and 2 (*M* = -1.58% and *M* = -1.60% respectively) (Tukey-HSD = 5.20, $p < .01$). Trial Block 4 (*M* = -6.00%) was not significantly different from the other three trial blocks.

The predicted difference between the feedback and no-feedback groups in amounts of confidence, accuracy and overconfidence, did not occur and there was no significant interaction with trial block. The two groups show very similar patterns of overconfidence across trial blocks, with the feedback group having only slightly (though non-significantly) lower overconfidence. This may be because subjects intuitively knew whether they were getting the questions correct or not, and thus feedback from the experimenter was not necessary.

Table 5.2

Overconfidence for Different Levels of Question Difficulty Over Trial Blocks

	Trial Block				
	1	2	3	4	*F*
A Hard	13.30	18.00	-0.18	-1.14	16.57*
B Medium	-9.27	-12.77	-12.25	-3.82	2.37
C Easy	-8.80	-10.02	-13.90	-13.03	1.11
Mean Over-confidence	-1.59	-1.60	-8.78	-6.00	7.27*

*p < .001

Paradoxically, the subjects were better calibrated in the first two trial blocks, that is Questions 1 to 10, and became more **under**confident on later trials. This may be less anomalous than it appears, however, because the mean scores conceal different patterns of overconfidence which are revealed when the differing levels of question difficulty are taken into account. This becomes clear when overconfidence is broken down across trial blocks for each level of question difficulty.

Interaction of Question Difficulty and Trial Blocks

A significant interaction was found between question difficulty and trial blocks with respect to overconfidence, $F(6, 414) = 7.12$, $p = .001$ (see Table 5.2 and Figure 5.2). This appears to be due to the high levels of overconfidence for hard questions which declined sharply in later trial blocks. Almost perfect calibration was achieved for hard questions in Trial Blocks 3 and 4, which elicited significantly lower levels of overconfidence than Trial Blocks 1 and 2, $F(3, 196) = 16.57$, $p = .001$, Tukey-HSD = 11.81, $p = .01$ (see Table 5.2 for all means). There was no improvement in overconfidence over the course of the trial blocks for either medium-difficulty or easy questions, contrary to prediction for easy questions.

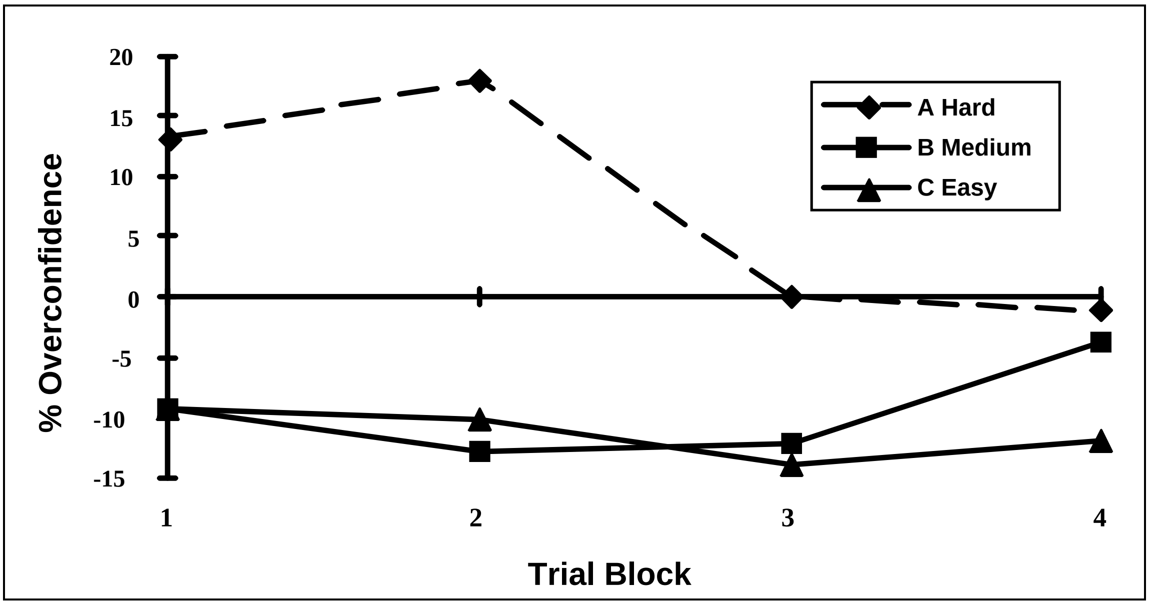

Figure 5.2. Overconfidence by levels of question difficulty over trial blocks.

Sex Differences

Combining all levels of question difficulty male subjects were found to be more overconfident than female subjects (M = -0.74% vs. M = -6.23%), $F(1, 138) = 6.90$, $p < .01$, although overall both groups were underconfident. Male subjects also found the task much easier than female subjects , at all levels of question difficulty, and achieved significantly higher accuracy scores (M = 59.06% vs. M = 49.43%), $F(1, 138) = 14.59$, $p < .001$. Male subjects correspondingly expressed significantly higher confidence in their judgements (M = 58.32% vs. M = 43.20%), $F(1, 138) = 34.53$, $p < .001$. The level of question difficulty did not differentially influence male and female subjects: both groups responded with similar patterns of overconfidence (see Figure 5.3), but the female subjects were significantly less confident and accurate than male subjects for all levels of question difficulty.

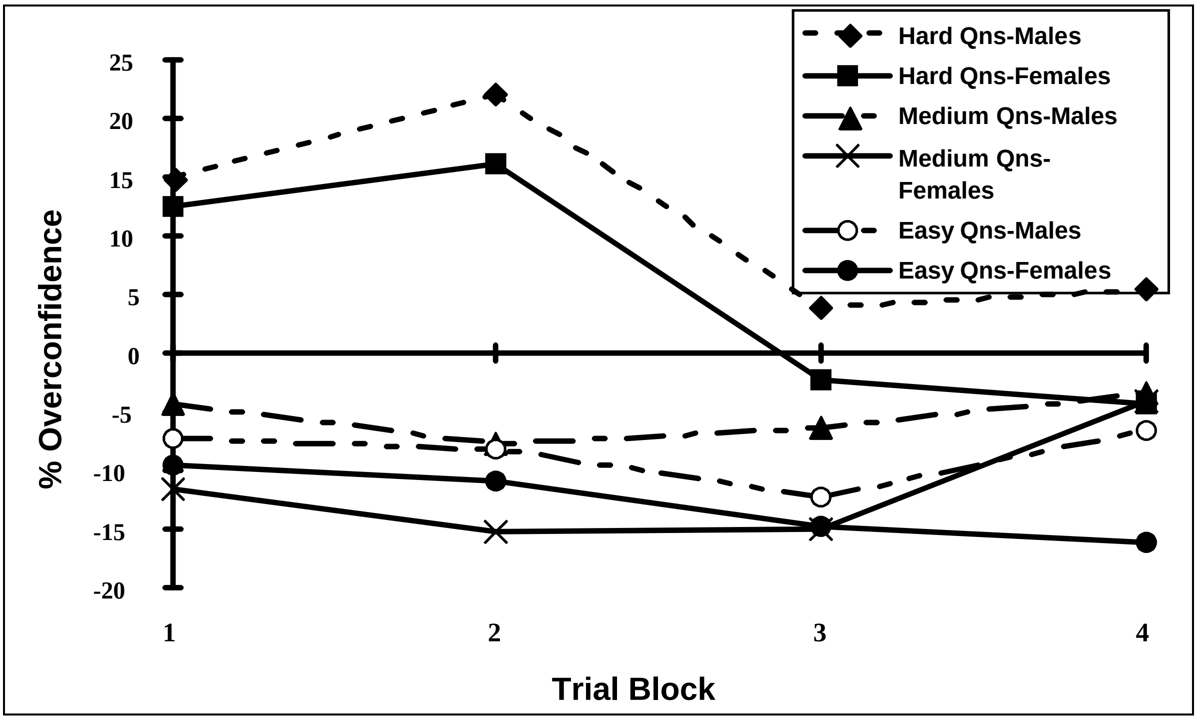

Figure 5.3. Levels of overconfidence for male and female subjects over trial blocks for different levels of question difficulty.

Discussion

No evidence was found for any significant effect of feedback on any of the dependent variables. This is probably because it is relatively easy, at least in this type of general knowledge task, for subjects to self-monitor their own accuracy and thus to reduce their confidence appropriately. Subjects can do this even without direct feedback from the experimenter, because they know roughly whether they are answering questions correctly or not. Support for this self-monitoring interpretation comes from the observation that subjects in the no-feedback, hard-level group showed much better calibration in the latter half of the trial blocks than earlier on. Thus subjects reduced their overconfidence for later questions, which suggest that they monitored their poor performance and in effect provided their own intrinsic feedback.

If internal monitoring accounted for the improved calibration of subjects in the no-feedback, hard-level group over trial blocks, then this process was apparently slightly less effective than external feedback at reducing over-confidence. In the absence of external feedback, subjects had to guess their accuracy rates and modify their confidence accordingly. This may have caused subjects to feel less certain about the validity of their own feedback in comparison to external feedback.

Subjects given medium-difficulty or easy questions showed no improvement in calibration of confidence ratings: their level of overconfidence remained stable across trial blocks and they were underconfident rather than overconfident. Several interpretations of this lack of improvement are possible. First, perhaps the subjects did not monitor their performance. However, the fact that subjects given hard questions and no feedback reduced their overconfidence levels over trial blocks suggests that self-monitoring occurred in that treatment condition, and it is reasonable to assume that subjects in other treatment conditions similarly engaged in self-monitoring. A second interpretation is that subjects may have attempted to self-monitor their performance but were unable to improve their calibration of confidence because the difficulty of the task was less clear-cut than in the hard-level treatment conditions. This is the most likely explanation for the lack of reduction in overconfidence over trial blocks in the medium-difficulty group. Third, perhaps the feedback, whether internal or external in origin, provided unambiguous information about the level of task difficulty, but subjects for some reason did not use it to modify their confidence ratings. This is probably the explanation for lack of improvement of calibration among subjects in the easy-level group.

Why did subjects in the easy-level group apparently not use the feedback to increase their confidence and improve their calibration over trial blocks? The feedback groups did not differ from the no-feedback groups in confidence or overconfidence for questions of any level of difficulty. Thus feedback was obviously not the important factor, because it could be provided internally. Perhaps even though the subjects realised that the questions were very easy, through feedback or self-monitoring, there were social or psychological barriers preventing them from increasing their confidence. Subjects may have failed to increase their confidence for easy questions because they appeared too easy, and perhaps subjects suspected that there may be a trick to catch them out.

Perhaps people only change their confidence levels when past performance seems to be a reliable predictor of future performance, and this may be easier when questions are consistently hard rather than consistently easy. Subjects in the easy-level group did not seem to change their impression of the level of difficulty of the questions. They may have believed that luck rather than knowledge accounted for the apparent easiness of the questions, and luck has an associated level of uncertainty with it as it may "run out", so confidence is lower than it could be. If people attribute higher than expected accuracy, when it occurs, to luck (external factors) rather than to internal factors such as skill, then perhaps they tend not raise their confidence in judgements, because they feel that they have less control over those external factors.

The evidence from this experiment indicates that feedback, either internal or external, seems to improve calibration of confidence only when the questions are consistently hard. This could be because there is more social pressure to reduce confidence during hard tasks to save face in case of failure. This social pressure does not seem to raise confidence for easy questions where people are underconfident, because it may be more socially desirable to be underconfident and may even be a way of boosting self-esteem in cases of success when low confidence was previously expressed. Maintaining low confidence is also a protective strategy, because if confidence is raised during an easy task, but the achieved accuracy is not as high as expected, then the task cannot easily be blamed for being too hard and thus failure must be internally attributed, which may threaten self-esteem.

Alternatively, a more cognitive explanation may account for the improvement in calibration for hard but not easy questions, in that feedback for hard questions may encourage subjects to generate evidence contradicting the chosen answers, which is known to reduce overconfidence (Koriat, Lichtenstein, & Fischhoff, 1980). Feedback to easy questions, however, does not encourage any generation of confirming evidence, which in any case is less effective in improving calibration.

In some situations feedback is disregarded when assessing future confidence levels. People may, for example, fail to alter their confidence levels because of motivational factors. However, it also seems logical to suppose that in the absence of external feedback subjects will not alter their confidence ratings if they do not realise that their confidence ratings are out of kilter with their true accuracy levels. If people cannot accurately assess their true accuracy levels then calibration will not improve. In the absence of external feedback improved calibration will only occur if subjects realise that their beliefs about their accuracy rate are out of line with their true accuracy levels. Thus to improve calibration, subjects must either be realistic in their assessments of their accuracy or have external feedback to give them an accurate measure of their true performance.

The most inappropriate levels of overconfidence in this experiment occurred at the extremes of the difficulty continuum, with judgements that were either very hard or very easy, where subjects did not realise just how difficult or easy the questions were. Some judgement tasks in this experiment are misleading in their level of difficulty inasmuch as they seem much harder or easier than they really are. This in turn leads to poor calibration, because the objective accuracy rate does not turn out to equal the subjective accuracy, which is what subjects are basing their confidence on. For example, question A9 asked “what is considered to be the luckiest number worldwide?” (the answer being 9) and this produced the most overconfidence of 51%. This question did not have very high confidence associated with it (only 55%) but the accuracy rate was only 4%. The question is deceptively hard because in Western countries the number seven is considered lucky, so this answer seems most likely to be true to anyone from the West. Questions which do not have an appealing but inaccurate alternative answer tend to result in very little overconfidence, for example question B12 asked “what is the name of the number system using only the symbols 1 and 0?” (answer: Binary) resulted in zero overconfidence as both accuracy and confidence were 58%. No other answer readily springs to mind to mislead subjects here. At the other extreme, questions can be easier than anticipated and result in higher accuracy than expected. Sometimes the subject does not know the answer so expects to have low accuracy, but their guess turns out to be correct because of the paucity of other alternatives. For example question B15 produced underconfidence of -31% as 64% of subjects accurately guessed that World War 2 was “the war waged by 57 countries” but were only 33% confident in their answer. Subjects match their confidence to the expected level of accuracy, it is when the accuracy obtained is radically different for the expected that overconfidence occurs. Thus it is when accuracy is unexpectedly low or high that over or underconfidence is found.

Gender also appears to have an effect on overconfidence: male subjects were significantly more overconfident than females. The female subjects found the task significantly harder than the males -- they exhibited lower accuracy overall. Harder tasks usually produce more overconfidence, but the female subjects in this experiment were more *under*confident than the males. The female subjects believed either that the general knowledge task was harder than it actually was or that their general knowledge was poorer than it actually was. Because they believed that they were doing less well than male subjects, their confidence was correspondingly lower. As much as 20 percent of the variance in confidence ratings was explained by gender.

It is possible that there was more social pressure on the female than the male subjects in this experiment to show low confidence for this type of task. In most previous experiments, for example in social and self-predictions reported earlier in this thesis, gender differences in overconfidence have not been found. This raises the question: are female subjects actually less confident or are they under-reporting their confidence to conform to social pressure to be modest in an intelligence type task? Lenney (1977) argued that women are not universally low in confidence in all tasks, and thus the desire to appear modest cannot invariably account for lower self-confidence, but that “social factors do influence women’s actual levels of self-confidence” (p. 9). This could have consequences for women in terms of success and achievement, and it merits further research.

To conclude, feedback seems to be effective in improving calibration of confidence and accuracy only when the questions are consistently difficult; it does not improve calibration very much for medium-difficulty questions or easy questions. Feedback, be it intrinsic or extrinsic, is important for regulating the level of overconfidence but is not always effective in improving calibration of confidence unless the task is consistently difficult. In everyday life, tasks are often a mixture of hard and easy components and thus overconfidence may not be reduced to such a large extent. Overconfidence or underconfidence in everyday life appears therefore to be due to a combination of motivational factors associated with self-presentation and also uncertainty about the level of task difficulty leading to mis-calibration of confidence.

CHAPTER 6

Discussion

Discussion of Findings

This thesis has shown that there are situations where the perceived probability of a judgement being correct (confidence) does not equal the actual probability of that judgement being correct; people are over- or underconfident. The types of judgements which lead to this miscalibration have been investigated, and some influencing variables have been identified. In Chapter 1 it was found that judgements for oneself show significantly less overconfidence than predictions for another person, and that this difference was most noticeable at mid-range base rates, where base rate information is less predictive. It was concluded that people over-rely upon dispositional information, and that dispositional information is more likely to be incorrect for other people than for oneself leading to more overconfidence in predictions about other people. Subjects did, however, show less confidence in their predictions about their partner than in predictions about themselves, which indicates that to some extent they were aware of the difficulty of making predictions about other people. This reduction in confidence was, however, insufficient, as subjects were more overconfident for predictions about their partner, indicating that the subjects underestimated the difficulty of the task.

The study also intended to discover whether the findings of Vallone, Griffin, Lin, and Ross (1990) who used a half-range scale of measurement could be replicated with a slightly different response format. Thus in Chapter 1 a single response format was used to measure both accuracy and confidence in each prediction. The two studies did indeed show similar levels of overconfidence, with a difference in overconfidence of only one percent on average. The full-range scale of measurement did, however, have a significant drawback in that the 50% prediction indicated that the subject was unsure about whether the event would or would not occur. In this circumstance the objective accuracy could not be measured because no definite prediction in either direction had been made. This response format therefore excluded an analysis of how accurate subjects were when they said they were 50% confident. This problem was avoided by using a half-range scale in Chapter 2, with separate responses for predictions and confidence judgements. This problem does not arise in later chapters because only confidence in knowledge was being assessed and no predictions are made.

The relationship between base rates and outcome positivity/negativity was examined in Chapter 2, which focused on self-predictions of future behaviour, and high overconfidence for items with mid-range base rates was found. Overall it appears that when a wide selection of items are used, covering the entire base rate range, people are fairly well calibrated in self-predictions of their future behaviour. Being more than 90% confident resulted in the most overconfidence; indicating that people should be more careful when they are extremely confident about their predictions. The exception to this statement occurred when the base rate of the event being predicted was extreme, where underconfidence was found; when confidence should be very high it was not high enough. The desirability of the outcome of the event being predicted also affected overconfidence, as more overconfidence was found for predictions with positive outcomes than for those with negative outcomes. There was no significant difference in accuracy for predictions with positive and negative outcomes, but confidence was significantly higher for positive-outcome events. This may occur because people are more able or more motivated to bring to mind supporting evidence for positive events which they desire to happen, and the generation of supporting evidence raises overconfidence (Koriat, Lichtenstein, & Fischhoff, 1980). Chapter 2 also discovered that when subjects were asked to estimate population base rates they overestimated low base rates and underestimated high ones; they did not seem to appreciate that some events occur very frequently or very infrequently and thus the extremity of their estimates was reduced. The neglect of base rate information in assessing confidence was implicated as a cause of the increased overconfidence at mid base rates, where base rate information is less predictive but still useful.

In both Chapters 1 and 2 subjects were required to predict their future behaviour and the accuracy of their predictions was measured by comparing their predictions with self-reported outcomes after the specified time period. The follow-up questionnaires asked subjects if they had experienced the event in the previous month/week. In Chapter 1 there was a difficulty in establishing exactly when the follow-up questionnaires were completed, as subjects could have filled in the follow-up questionnaires during the month in question, before they were due to fill them in and hand them back to the researchers at the end of that month. This problem was eliminated from Chapter 2 by ensuring that the follow-up questionnaires were completed in the presence of the researchers at the end of the time period. This latter method is obviously much more reliable and should be used in future experiments.

It could also be argued that subjects were motivated to appear consistent in their answers to the questionnaire and follow-up questionnaire. The questions on both questionnaires were presented in the same order and this could lead to better recall of answers from the first questionnaire. It would be advisable to randomise the order of presentation of questions in future experiments of this type. If a subject were deliberately repeating their answers to the first questionnaire on the follow-up questionnaire, or reversing their answers, then accuracy levels would appear either very high or very low for that subject. It would be expected that if subjects were motivated to appear consistent in their answers then accuracy would be high. However, from looking at the accuracy data from Chapters 1 and 2, presented in Appendices B and H, it can be seen that accuracy varies considerably for different questions, and the mean accuracy for subjects is 78.9% in Chapter 1 and 82.1% in Chapter 2. If mere reporting of outcomes, to be consistent with predictions, was occurring then such variation should not be so apparent, and a higher level of accuracy would be expected.

The possibility that the accuracy scores have been affected by misleading reporting of events by the subjects cannot, however, be totally ruled out. It is to be expected that if subjects were not telling the truth then it would be in order to increase consistence and accuracy. In this case the overconfidence effect would be reduced because accuracy would be artificially raised. If subjects were mis-reporting the true occurrence of events then it would appear that calculations of overconfidence in these chapters are below the true level due to subjects covering up their mistaken judgements. The self-reporting method used here does rely on the honesty of the subjects. The subjects were asked to be truthful and accurate in their answers to the follow-up questionnaire, and they were told in the instructions given to subjects in Chapter 2 that truth, not consistency with the previous questionnaire, was important.

Another question which arises from Chapter 2 is whether the base rate alone is influencing confidence and accuracy, or whether another factor could also be involved. This experiment controlled for outcome desirability, but another factor, such as the controllability of the outcome of the predicted event, could also be related to the base rate and may influence overconfidence: for example, at the extremes of the base rate, where events happen to a large or small proportion of the population, controllability may differ from mid-base rate ranges. If people perceive that they have more control over an event occurring or not, then they may be more confident in their prediction of its occurrence (see Weinstein, 1980, for a discussion of how valence of outcome interacts with perceived controllability in affecting unrealistic optimism). The subjects' perception of how much control they have over the event may, however, be erroneous and thus their accuracy may be low and overconfidence high. The experiment reported in Chapter 2 could be repeated to assess the role of controllability and its influence on overconfidence, by including a section where subjects are asked to rate their perceived degree of control over the outcome of each event. It would also be informative to ask subjects, in an open ended question format, about their confidence and their responses to the predictions, in order to determine what other factors may also be responsible for influencing overconfidence.

The type of prediction and the person being predicted for were examined in Chapters 1 and 2. Chapters 3 and 4 then went on to examine whether some people are more prone to showing overconfidence than other people, and what factors may influence this. Previous research had examined the relationship between overconfidence and intelligence (Lichtenstein & Fischhoff, 1977), and authoritarianism, conservatism, dogmatism, and intolerance of ambiguity (Wright & Phillips, 1976). No previous research had investigated whether differences in cognitive styles was related to people's levels of overconfidence, so Chapters 3 and 4 set out to discover if such a relationship existed. The Embedded Figures Test of field dependence was chosen because it was readily available and it has been shown to give a reliable measurement of field dependence. A test of probabilistic reasoning ability was also included in these two chapters to determine whether people's ability to understand numerical data was influencing their subjective probabilities and their overconfidence. The hypothesis was that people who understand the laws of probability, or those who can more accurately express their degree of certainty in numerical form, may be less overconfident in their predictions. Another factor which was investigated was the extent to which the subjects could perceive the influence of the situation upon their behaviour. A test of the degree to which subjects committed the Fundamental Attribution Error was used to see if they perceived situational factors as affecting behaviour more for themselves than for other people.

It was concluded that a person's overall overconfidence in both self-predictions (Chapter 3) and general knowledge (Chapter 4) was not significantly associated with how field dependent they were, their self-esteem, or how good they were at probabilistic reasoning. The probabilistic reasoning test used in these experiments was limited to five questions due to time constraints. The small range of possible scores on this test may not be large enough to accurately distinguish between probabilistic reasoning skills in the subjects. It would be advisable to use a more reliable, standardised, test of probabilistic reasoning in future research. The limitations associated with the Embedded Words Test (EWT) have already been mentioned in Chapter 3. The EWT is a new test and its construct validity has not been established. This experiment was useful in determining problems with the EWT and after modifications it may be useful and should be investigated more deeply. The Embedded Figures Test's (EFT) validity as a measure of field dependence is long established in the literature. In these experiments there was no evidence of a relationship between field dependence and overconfidence using the EFT.

A small (r = .31) and unexplained relationship between the subject's degree of committing the Fundamental Attribution Error and overconfidence in general knowledge was found in Chapter 3. Problems arose in interpreting the results of the Fundamental Attribution Error test because no significant error was found. The test was designed specifically for this experiment and had not been previously validated. There may have been problems with the rating scale used in the FAE test, in that the mid-point was labelled as "neutral", with the end-points labelled "entirely due to the person involved" and "entirely due to the situation". The subjects' interpretation of the mid-point may have differed from that intended by the experimenter, as they may have seen it as an abdication of an attribution in either direction rather than a 50:50 weighting of both situational and dispositional factors. Thus it seems probable that the findings relating to the FAE score are unreliable. This test was deemed not to be adequately measuring the effect and, therefore, not used in Chapter 4.

As previously mentioned, very little research has been done to find out what individual differences affect overconfidence. There are several factors which may deserve attention in the future. This thesis has not been able to examine all these factors, but some ideas about what could be occurring can be mentioned. Firstly, what is the role of private self-consciousness in overconfidence? It may be the case that those people who are more privately self-conscious, or high self-monitors, are more aware of the causes of their behaviour, and are thus better able to accurately predict their future behaviour. This may mean that private self-consciousness is related to a better match between confidence and accuracy, because as accuracy increases overconfidence usually decreases. In retrospect it would also have been valuable to include personality tests in Chapters 3 and 4 to see if there was a relationship between introversion/extroversion and overconfidence. Further research relating overconfidence to intelligence would also be useful because Lichtenstein and Fischhoff's (1977) study only compared graduate students to undergraduate students, and a wider range should be employed. Lack of previous research in these fields precludes any firm hypotheses.

The fact that more overconfidence was found for positive-outcome than negative-outcome events in Chapter 2 indicates that the valence of the outcome is important. Thus it would seem logical to suppose that if people differ in how positively or negatively they see events, then their overconfidence may also be affected. The relationship between optimism/pessimism and overconfidence is worthy of investigation, especially in relation to the valence of the outcome of the event being predicted for. For example, pessimists may process more information about negative events but not about positive events, resulting in lower overconfidence for such events than shown by optimists. If people who are defensive pessimists set lower expectations and think through possible outcomes more than optimists (Norem & Illingworth, 1993) then they may also be better calibrated. Shepperd, Fernandez and Ouellette (1996) suggested that pessimism prompts a person to avoid the perceived negative outcome by increasing their efforts. Therefore the resulting accuracy of predictions by pessimists and optimists may differ and overconfidence may be affected. Optimists may also have a higher perceived level of control over events and this may raise their level of overconfidence.

Overconfidence may also be affected by the perceived locus of control, because if people see control of events as external to themselves then they may have lower confidence in their ability to influence the future (see Furnham & Steele, 1993, for a review of research into locus of control). Those people who tend to perceive the locus of control as external to themselves may feel less confident and more helpless and be more underconfident.

The causes of individual differences in overconfidence thus need much more investigation to determine what factors make some people well calibrated and others more or less overconfident. What can be said is that more confident subjects tend to be more overconfident and those who are more accurate are less overconfident, due to the poor correlation between the mean level of subjects' confidence and accuracy for the group as a whole (Chapter 4). It was also found that those subjects who were more accurate at estimating the base rate of the event in their population showed less over- or underconfidence, probably because some people have a better ability to accurately perceive the factors which affect both the probability of the population experiencing an event and themselves experiencing it. An interesting difference between arts and science students in overconfidence for self-predictions was also found and discussed in Chapter 4, though no significant sex differences were found. The influence of these factors should be investigated further to discover if they apply to other judgemental domains.

The effect of feedback on overconfidence was the main focus of study in Chapter 5. Earlier work in Chapters 3 and 4 indicated that individual differences in overconfidence exist. Therefore some mechanism must operate to improve some people's calibration. Chapter 5 was, therefore, carried out to determine whether feedback is influential in reducing overconfidence and whether the difficulty of the task interacts with feedback in determining overconfidence. Chapter 5 revealed that overconfidence in a general knowledge task was reduced over time when the task was consistently difficult. The fact that confidence could be altered to fit the circumstances of the experiment does indicate that subjects are aware of task difficulty and can modify their confidence if they are sufficiently convinced of high task difficulty by the consistent nature of the task. Sex differences were found in Chapter 5, as male subjects were significantly more overconfident than female subjects in a general knowledge task, possibly due to self-presentation biases, which did not occur for behavioural predictions in Chapter 2.

Although the results of Chapter 5 do interestingly indicate what occurs to overconfidence in the short term, the methodology of this experiment could have been improved by the expansion of the task to obtain a more reliable measurement of overconfidence by including more questions. A serious criticism of this experiment can also be made due to the failure to randomise or counterbalance the order of presentation of the questions. If by chance some questions were placed in the latter trial blocks which differed in difficulty from those in earlier blocks then an artificial change in overconfidence may have resulted. This should be considered in future experiments and controlled for by randomising the question presentation order for subjects in the same condition. Manipulation checks should also be carried out to determine if the difficulty level of the questions remains constant across trial blocks. The feedback interval time could also be considered as a variable in future research, as it is currently difficult to know how confidence alters over time. The feedback condition may have failed to show significantly lower overconfidence than the no-feedback condition, contrary to the hypothesis, because the time given in between questions was not long enough for the subjects to reflect upon, and be influenced by, the feedback information.

The results of Chapter 5 tally with those of Arkes, Christensen, Lai, and Blumer (1987), who also found that feedback in the form of correct answers given to a highly overconfident group did significantly improve their calibration (in this case a reduction in overconfidence). It has been suggested that the generation of reasons why an answer may be wrong is the critical factor in reducing overconfidence (Koriat, Lichtenstein, & Fischhoff, 1980). I propose that when subjects in the hard level condition were exposed to more difficult questions over the course of the trial blocks they began to realise that the task was harder than they had originally anticipated and that they may be getting some answers wrong, and generation of reasons why the answer may be incorrect may have increased. This longer term perception of the questions as time passes may have caused the reduction in overconfidence in the latter trial blocks for the hard level questions. No such mechanism operated to raise the level of underconfidence for the easy level questions because it is only the production of contradictory evidence which has been shown to affect overconfidence. Paese and Sniezek (1991) also found that subjects altered their confidence over time, without feedback, because their perception of the task difficulty and their performance changed as the task progressed.

The interpretation of the results of Chapter 5 is, however, open to question as there is no direct evidence of internal feedback occurring. Another possibility is that subjects were picking up feedback about the general level of difficulty of the questions in the form of non-verbal cues from other subjects in the same room who were also doing the experiment. Single person testing may be advisable in this type of experiment. The hypothesised mechanism of internal feedback should be explored in future research, and this experiment replicated with the suggested modifications to ensure the results are real and not due to chance alone.

The Overconfidence Bias

At the start of this thesis it was stated that confidence is the subjective probability estimate of the correctness of a belief. Over the course of this thesis it has become apparent that the confidence (subjective probability) that a person has that a judgement is correct does not always equal the objective probability of that judgement being correct, possibly because subjective probabilities are biased by cognitive problems of assessing uncertainty and motivational factors. Overconfidence probably often arises because we trust our memory more than is justified. It seems that people do not consider uncertainty or give it enough weight when assessing their confidence level, and the hard-easy effect is due to the lack of appreciation of unpredictable variables for hard judgements and a better appreciation of the not-so-unpredictable variables for easy judgements. When estimating a value for uncertainty many factors have to be considered, and the number of facts which spring to mind to support or contradict a judgement will influence the confidence in it. It also appears that when people generate their own hypotheses for a judgement they express lower confidence in it, and are better calibrated, than people who are then presented with the same hypotheses for evaluation. This was found to apply to social inference tasks, general knowledge tasks and real world predictions (Koehler, 1994). It appears that at the stage of hypothesis generation, producing an argument to support a hypothesis leads to increased overconfidence, whereas considering several viable alternatives reduces overconfidence. When, and if, a person moves on to a hypothesis evaluation stage then the tendency is for the hypothesis to be treated as if it was true and the information search is biased towards confirmatory evidence.

In Chapter 1 it was found that confidence varies more for self-predictions than for predictions for other people. This may be because people use stereotypes about other people to inform their predictions, which they do not use about themselves, and they place too much validity in those stereotypes and are too confident in them. As confidence increases the amount of information people seek in making judgements declines, as Howard and Sheth (1969) showed about buyer behaviour. Levin, Chapman and Johnson (1988) commented that this failure to seek new information may be because higher confidence makes people underestimate the uncertainty produced by gaps in their knowledge and unavailable information. Complex judgements bring in more complex factors to influence calibration as they introduce external uncertainty which may lead to biased estimates.

At extreme base rates where there is a high degree of certainty that an event will or will not occur it may be harder to generate reasons why the event should occur (low base rates) or should not occur (when the base rate is high). At mid base rates there is more information available as to why an event may or may not occur, and some degree of confusion occurs resulting in some factors not being weighted sufficiently, thus causing overconfidence.

The measurement of overconfidence across the whole range of confidence judgements was made possible when Adams and Adams (1961) proposed the concept that "of all those decisions made with confidence *p*, about *p*% should be correct" (p. 37). In Chapter 1 subjects were asked to write down the probability that they would experience the events listed. In Chapters 2 and 4 subjects were given more detailed information about what each confidence statement would mean in terms of probability. For example 90% means that you are very sure that you are right in your prediction and that you'd expect to be right 9 times out of 10, or the probability your answer is correct is 0.9. Chapters 3 and 5 asked subjects to rate their degree of confidence that their chosen answer on a general knowledge task was correct.

Informing subjects that confidence *p* should equal *p*% correct improves clarity and may make subjects use the scale in a standardized way. This instruction may, however, be changing the real process which is under examination. This instruction may induce subjects to consider their answer more deeply and answer in an atypical way in order to follow instructions. Because of the possible de-biasing nature of these instructions the instructions in Chapters 1, 3, and 5 did not specifically draw attention to the fact that confidence should equal accuracy, but the instructions were clearly phrased and the aim was to try to tap what people in everyday life understand by "confidence".

The instructions given in the experiments may, however, not have sufficiently emphasised to subjects the degree of accuracy required of them by the experimenter when using the confidence scale. Research over more than thirty years has, however, consistently found the overconfidence bias, and it is considered to be robust (Fischhoff, 1982), despite many variations in testing procedures and methods. The subjects tested for this thesis did not appear to have any difficulty understanding the instructions used in any of the experimental testing situations, and they were encouraged to ask questions if they needed any help with the questionnaires. The standardised use of the confidence scale may be important and the use of more detailed explanations of the scale should be considered. Since the question of whether or not specific instructions has any effect on subjects' behaviour has not been completely resolved, it would be informative to compare the confidence scores given by subjects under different sets of instructions.

Some researchers have argued that inferential processes do not cause overconfidence, because they argue that a task with balanced levels of task difficulty reveals good calibration (for example, Juslin, 1994; Juslin, Winman, & Persson, 1995). Looking at judgements where over- or underconfidence exists, however, is still important as this is where the process of accurately assessing one's confidence has broken down, because reality is different from what was expected. Poor judgement is not universal; it occurs under certain conditions. People are not always over- or underconfident, some judgements show good calibration and these should receive more attention in future research.

It has been argued by Schraw and Roedel (1994) and others, that the severity of overconfidence that has been reported in the literature is due to test difficulty. However, they concluded from this that people are not poor judges, and that biased selections of test questions are accountable for the overconfidence bias. However, if people were perfectly well calibrated, and good judges, then they would be well calibrated for each item and not over- or underconfident on average. What may be justifiably questioned is that there is a general bias towards *over*confidence rather than *under*confidence, and a biased selection of items, with a high proportion of difficult ones, may be responsible for giving the impression that people tend to be more confident than accurate. In this thesis underconfidence has been found for many items that people made predictions about; namely items which were easier than people expected. When these items are included in experiments very low mean overconfidence is found overall, because over- and underconfidence balance out. Experiment 5 showed overall underconfidence because of the balance of test items used; with a prevalence of easy questions. Therefore the argument that there is an *over*confidence bias can be justifiably criticised. However, the fact that biases in both directions are found shows that people are not well calibrated and that overconfidence exists.

The term *overconfidence* has been generally used by past researchers because the bias tended to be found in the positive (overconfident) direction in early research. However, the bias of confidence and accuracy is not always in the confidence-is higher-than-accuracy direction, and as such the bias could really be called a "calibration of confidence bias", or "over/underconfidence bias". Perhaps the least confusing term would be just to call it the "confidence bias". The bias can go in either direction; over- or underconfidence, but the bias itself is an underestimation of the extremity of difficulty of the problem. It can be confusing to refer to overconfidence when in fact the value is a negative amount and in the direction of *underconfidence*. It is also difficult to refer to reasonably similar levels of confidence and accuracy. In such a case it can be said that little overconfidence occurs, there is good calibration, or that there is a good match between the two. The situations which cause a group average of over- and underconfidence have been examined in this thesis and in much previous research, but which people will be underconfident and which will be overconfident in each situation, and why, has yet to be determined. The term "overconfidence bias" is generally used, recognised, and will probably continue to be used, but it should be understood that it does not always mean just overconfidence because it is a bi-directional bias which also encompasses underconfidence.

The Hard-Easy Effect in Overconfidence

The hard-easy effect occurs when the amount of overconfidence increases with the level of difficulty. The difficulty of the task is obviously a critical determinant of overconfidence, as is shown by the "by questions" analyses in Chapters 1, 2, and 3, which found high negative accuracy-overconfidence correlations. The hard-easy effect also applies to social judgements as well as to general knowledge tasks. In Chapter 1 the correlation between accuracy and overconfidence was $r = -.90$ for self-predictions and $r = -.95$ for predictions for one's partner, indicating that the difficulty of the question was responsible for 81% to 90% of the variance in overconfidence in behavioural predictions, leaving confidence to account for 10 to 19% of the variance. Chapter 3 also showed a strong hard-easy effect, as the predictions which were most in error (low accuracy) showed most overconfidence ($r = -.76$). In social judgements, such as behavioural predictions, the hardest predictions are ones with mid-range base rates, where subjects do not sufficiently anticipate and allow for the influence of the unknown on their predictions. In Chapter 2 less of the variance in overconfidence was accounted for by difficulty than in Chapter 1 probably because a wider range of item difficulties (base rates) was used, which gives a better indication of the true variance across the entire spectrum of judgements from very hard to very easy.

In Chapter 3 the relationship between accuracy and overconfidence in a general knowledge test showed a correlation of $r = -.73$. Thus it seems that the difficulty of the questions is an extremely important factor, accounting for about 53% of the variance of overconfidence in general knowledge judgements and (referring to the previous paragraph) 58% of the variance in social predictions. The relationship between accuracy and overconfidence in Chapter 5 will not be discussed as it was subjected to an experimental manipulation, and this would hide the true relationship between accuracy and overconfidence present in other natural tests of general knowledge. In conclusion, there seems to be an inertia in confidence estimates which are not sufficiently modulated to track variations in accuracy (difficulty). This may possibly be due to subjects using the confidence scale in the experiment in a more approximate way than the experiment required. People seem to be slightly more aware of variations in difficulty with general knowledge questions and they modulate their confidence more than for behavioural predictions, where they tend to maintain high confidence irrespective of actual difficulty. This may be because people feel that they have personal control over future events and thus have higher confidence in behavioural predictions. It is difficult to make comparisons between behavioural and knowledge judgements though, as the difficulty of the questions is not the same in both experiments.

Juslin (1993) argued that the hard-easy effect is only due to the post hoc division of questions into categories of hard and easy on the basis of the proportion correct. In much of the reported research the difficulty of the task is defined by how well the subjects perform, that is, a difficult judgement would be one where few subjects answer it correctly and an easy one where most subjects answer it correctly. The hard-easy effect is not, however, an artefact of biased item selection, it is a real effect, where people do not realise, and allow for, the extremity of difficulty, or ease, of the judgement (Griffin & Tversky, 1992). If people were perfectly calibrated then when the judgement was hard they would realise this and be less confident, but less well calibrated people do not realise that the judgement is hard and thus they are overconfident.

Unrealistic Optimism

When a person makes a judgement or prediction it may be correct, or alternatively it may be incorrect in which case it is unrealistic. If the prediction is incorrect (unrealistic) then the direction of the prediction is relevant: if the prediction is that a good event will happen and it does not, then the prediction is unrealistically optimistic, whereas if the prediction is that a negative event will not happen when it does happen, then the prediction is also unrealistically optimistic. The prediction can either be dichotomous (either the event will occur or not, with which the outcome can be compared for the individual) or the prediction can be stated as a likelihood of the event occurring, in which case the group's perceived likelihood can be compared to the actual proportion of events (likelihood) for that group to see if a bias is occurring.

When estimating the subjective probability of a future event occurring there is a tendency for people to overestimate the likelihood of positive-outcome events occurring and to underestimate the likelihood of negative-outcome events occurring (Irwin, 1953). This bias has been called unrealistic optimism. It has been found that people are unrealistically optimistic about their own futures, that is, they overestimate the likelihood (probability) of positive events occurring and underestimate the likelihood of negative events occurring in relation to the perceived likelihood for other people (Weinstein, 1980).

Thus it seems that people show both an inability to correctly estimate probabilities of future events occurring (unrealistic optimism), and an inability to correctly estimate the probability that their judgements are correct (overconfidence). It does not seem from Chapter 2 that people are unrealistically optimistic in their short-term self-predictions, as subjects overpredicted the occurrence of both positive- and negative-outcome events, but they did show overconfidence in those predictions. Future research could investigate whether predictions about those events which people are most optimistic about are made with more or less confidence than realistic predictions. If people cannot accurately assess the likelihood of the event occurring then they may also be unable to assess the likelihood of that prediction being correct. Thus unrealistic optimism in predictions and overconfidence in them may be positively correlated. Alternatively confidence may be reduced to compensate for unrealistic optimism, and there may be a trade-off between the two.

The research on overconfidence and how people assess the probability of their judgements being correct may throw light on findings in unrealistic optimism research. For example, the amount of information people have affects likelihood estimates and the research on generation of pro reasons (why an event will occur) and con reasons also relates to unrealistic optimism, as people are motivated to consider pro reasons for positive events and con reasons for negative events, and this affects probability estimates. This may then help to explain why lower overconfidence for negative-outcome items was found in Chapter 2.

The Effect of Time on Calibration

It seems logical that an overconfidence bias should be found for predictions, because of the uncertainty of the external environment which introduces unexpected influences which alter the outcome, and thus confidence turns out to be poorly calibrated. However, in real life confidence is not a one-time, fixed or static event. Confidence in predictions is continually reassessed in the light of new incoming information and people become more or less confident in their predictions as time goes by. There may also come a point in time just before the actual outcome where confidence does equal accuracy, because the time involved between the prediction and the outcome is small and most of the relevant information is known. For example, a person goes to a horse race and judges that the horse “Lucky Winner” will win the race. They may then go and place a bet. The size of that bet may vary according to the certainty (confidence) of the person that the chosen horse will win (leaving aside other financial considerations, such as how big their overdraft is). At the start of the race the person is confident that the horse will win, but there is still an element of uncertainty. Half way through the race the horse may be half a mile behind the front runners, at which point, in the light of the information about the horse’s position in the race, the person would feel less confident that the horse will win. Making a prediction a minute before the end of the race would undoubtedly be better calibrated than a prediction one minute after the start of the race. Half way through the race predictions may still be poorly calibrated because unexpected events could still occur, the horse may fall at a fence, or the horses in front may fall and the straggler may overtake. Thus the impact of timing of judgements is important, and poorer calibration must surely be inevitable as the time between prediction and outcome increases. The findings reported in Chapter 2 emphasise the importance of timing, because in short-term predictions of the future there was no unrealistic optimism present in predictions, contrary to that found in long-term predictions as investigated by Weinstein (1980).

Confidence in Everyday Life

The majority of the work cited in this thesis required subjects to use numerical rating scales to give written estimates of their confidence. This is a convenient method for data collection in psychology experiments but neglects the primary means of communication of confidence in everyday life: verbal and non-verbal messages. The generalisability of this research to everyday life needs to be determined through more ecologically valid studies, especially to rule out the possibility of artificial biases due to experimental measurement techniques. Many experiments have assessed internal confidence via the questionnaire method. A few experiments have studied how we communicate confidence to other people and its role in person perception. People sometimes distrust overt expressions of confidence, such as “I'm certain” “very” “absolutely”, especially if they are overused (Erickson, Lind, Johnson, & O'Barr, 1978). Non-verbal signs such as pitch of voice, speech rate, eye contact, smiling, self-touching behaviour are may also be used as indicators of confidence by an observer as much as verbal cues. Different types of speech were presented to subjects in Scherer, London and Wolf's (1973) study, which found that people infer higher confidence in other people when the voice is louder, the rate of speech is more rapid and speech pauses are shorter. There is, however, a need to do more ecologically valid studies than this one, using samples of speech from real life rather than recordings of actors speaking “confidently”. The role of verbal information in a message which conveys information about confidence is important since more confident messages were found to be more persuasive by Maslow, Yoselson, and London (1971).

Kimble and Seidel (1991) measured subjects' confidence, speed of response, and loudness of vocal answers to thirty general knowledge questions and found that more confident answers were spoken more loudly and with faster response times. Using a test of assertiveness they also showed that more assertive subjects spoke louder, regardless of their confidence in an answer. They concluded that confidence in an answer was more influential on speech loudness than assertiveness, and that speed and loudness are good indicators of people's confidence. These researchers did not however analyse their results with respect to the accuracy of the judgements made, so relationships with over- or underconfidence cannot be made. There may possibly be a relationship between the degree of overconfidence in an answer and how fast or loudly it is said. The relationship between assertiveness and general self-confidence should also be investigated.

There are interesting findings relating to how people verbally and non-verbally express confidence and decode confidence in other people, but there seems to be a gap in looking at the actual accuracy behind those judgements. Expressions of confidence may vary when the person is underconfident compared with when they are overconfident, and how and whether people can correctly determine a person's accuracy from their confidence remains to be seen. Generally speaking people attribute high confidence to high levels of experience, which have presumably reduced the inhibitions which affect confidence. Confidence develops as experience, knowledge and skills in a field grows. A person's confidence can be picked up by other people through verbal and non-verbal cues and impressions are formed about that person. People tend to assume that there exists a high correlation between a person's confidence and their accuracy. However, this association can sometimes be erroneous, as everyday experience tells us that inhibitions can be reduced by other means, such as alcohol etc. If people are unaware that something such as alcohol is the cause of a person's confidence then they may erroneously believe a person to be more capable that they really are. For example, someone driving in a confident and assertive manner may be an experienced driver, or may be inexperienced but with “Dutch courage”. This in fact may be doubly dangerous as alcohol can reduce accuracy whilst increasing confidence via reduced perceptions of risk.

If someone is confident but we do not believe them to have the relevant experience then we may choose to attribute the confidence to something such as artificial stimulants or arrogance. Some people are not well calibrated and if other people realise this it can change their perceptions of that person's judgements. If confidence and accuracy are not reliably associated within a person then that person may be seen as unreliable, as an observer never knows which confident statements are in fact true. If an individual has been shown to be well calibrated in one sphere of judgement then it could be argued that more reliance should be attached to judgements made in another unverified but related sphere (Adams & Adams, 1961). A person who is perceived to be poorly calibrated in one judgemental domain may be subsequently doubted in other domains too. Problems may arise when generalizations are made to tasks/spheres which are tenuously related or even unrelated. For example, confidence in social skills in one setting may or may not generalise to another setting, but may be presumed to, and a transfer of attributes from social to educational fields may be even more problematic.

An interesting field of inquiry is how people judge the accuracy of a confident person. In social perception an observer may be misled and the actor may in fact be behaving in an over/underconfident manner to mislead the observer about the validity of a statement/judgement; lies may be enhanced by the subtle use of confidence statements. How people perceive another person's calibration of confidence and accuracy is important. For example, Mr X, the owner of a shop, may phone a supplier, Mr Y, to order some goods and he asks if they will be ready by Friday. My Y says "yes certainly" in a confident way. If over the course of several encounters the goods are always ready on time then Mr X will trust Mr Y's confident judgements. If, however, the goods are not always ready then Mr Y's judgement will be questioned and Mr X's faith in it will decline. A lack of calibration in a person will lead to an observer not trusting that person's judgements. Realistic confidence, good calibration, breeds confidence in an observer because a high certainty of outcome is expected and likely to be achieved. A poorly calibrated person may be trusted less by other people because they are less predictable.

It is possible that people could be categorized according to their calibration and accuracy, for example, accurate-underconfident, accurate-overconfident. The level of over- or underconfidence shows how well-matched accuracy and confidence are but given the choice between a poorly calibrated but accurate person and a highly calibrated but inaccurate person which would you choose? The absolute level of accuracy is still an important factor. It is important to remember that confidence is a reflection of what the person believes, and not what actually is in reality. If we remember that, and understand the distinction, then confidence tells us a lot about the person and the situation. When we evaluate someone's confidence it only tells us what they believe and not what necessarily is in reality. The role of confidence and calibration are very interesting in social psychology, in areas such as predicting other people's behaviour, and this research also has other applications.

Future Research

People's accountability for their judgements has been shown to reduce overconfidence (Tetlock & Kim, 1987). It would be interesting to see how people present themselves, in terms of confidence, when they believe that people observing their behaviour are unaware of the difficulty of the task and will not be aware of how well they performed. In this scenario, failure could be falsely attributed to the task or not admitted at all by the subject. If there is no way to "lose face" then the benefits of appearing confident would seem to be high. People may present themselves much more favourably and be more confident than when the task difficulty and their performance are common knowledge. It would be interesting to find out what sorts of situations lead people to inaccurately portray their confidence level so as to mislead other people.

Some indications about this come from a study by Kebbell and Wagstaff (1995) who found that in a task of eyewitness facial recognition, the group of subjects who could not have their answers verified expressed significantly higher confidence in their answers than the verifiable group. They also included a variable of hypnotically high/low susceptible subjects, which did not have a main effect on confidence. A significant interaction was found, however, as highly hypnotizable subjects became much more confident in the unverifiable condition than in the verified condition. Conversely, verifiability did not seem to affect the confidence of subjects of low hypnotizability. It seems that there are many influencing variables affecting the presentation of confidence when conditions of verification differ. It may be that hypnosis increases people's confidence when their answers cannot be verified and it may lower people's inhibitions and resistance to reporting things which they are not certain about.

Other questions also arise such as: do judgements about the future respond differently to feedback than ones about knowledge? Feedback may have less effect on calibration of predictions because people have to change schemas about the way that the world works, rather than just some facts about the world. Are people who are well calibrated in their knowledge about past events also well calibrated in their judgements about the future? The relationship between inhibitions and confidence should also be researched more in future, as people who can easily imagine future consequences, or who worry a lot, may be increasing their inhibitions and reducing confidence. Are introverted people more inhibited and less confident? Are extroverted people less inhibited and more confident and does this relate to self-monitoring behaviour? If so, does this only apply to social interactions or is it also meaningful for other areas of life? A general disposition towards low confidence may be due to excessive inhibitions and fear, which may spring from childhood and attachment to parents. Further investigation in these areas would be useful.

Experimental Limitations

There are limitations with extrapolating results from experimental situations to real life situations. For example, people often fail Wason's (1966) four card problem where they are given four cards showing A, T, 4 and 7 and asked to test the rule "If a card has a vowel on one side, it has an even number on the other". When, however, the same problem is rephrased as "If a letter is sealed it has a five cent stamp" (Johnson-Laird & Wason, 1977; Wason & Shapiro, 1971) subjects understand the task and answer it correctly. Thus experimental situations may indicate a deficit or bias in judgement which may not occur to such an extent in real life. I believe that the social-type tasks used in this thesis avoid some of the problems of experimental applicability to real life because events are used which are subsequently verified in real life rather than in the laboratory. The use of more naturalistic judgements is to be encouraged in this area of research.

It seems to be that confidence and accuracy are not well calibrated when the subject does not know how well they are doing, that is, they do not have good feedback. Thus in unrealistic experimental settings poorer calibration may occur than in real life as the subject does not know how well they are expected to perform. In realistic tasks there may be better calibration especially when feedback is given. In everyday life lack of feedback seems to cause problems such as stress and loss of motivation as people do not know how well they are performing and how to improve.

Conclusions

Is overconfidence good or bad for us? If someone is overconfident then researchers have suggested that they will use fewer precautionary measures, but is this true? They could be overconfident and also have taken insurance measures in case of failure. On the other hand, underconfident people may be taking too many insurance measures (which is costly) and/or avoiding risks. It could be argued that sometimes overconfidence can be beneficial, for example to mental health, but Griffin and Tversky (1992) concluded that "these benefits, however, may be purchased at a high price. Overconfidence in the diagnosis of a patient, the outcome of a trial, or the projected interest rate could lead to inappropriate medical treatment, bad legal advice, and regrettable financial investments. It can be argued that people's willingness to engage in military, legal and other costly battles would be reduced if they had a more realistic assessment of their chances of success. We doubt that the benefits of overconfidence outweigh its costs" (p. 432).

Does a small miscalibration really matter in everyday life? Will it influence the decisions taken and the resulting outcomes? Overconfidence does have implications for real life situations, as people can overestimate their personal abilities and this can affect how much effort and attention is put into a task (Stone, 1994). This thesis has shown that people are not always good at estimating uncertainty, and they are most in error when the judgement is very hard. For example, a judgement such as "how long will it take to write up your thesis?" requires an estimate of the amount of time required, and there is an associated level of confidence in that estimate, shown by the statement "I'm sure that I can finish it by the end of September". The influence of situational and dispositional variables which affect the time taken can, however, be underestimated; resulting in an incorrect and overconfident judgement, when the thesis is submitted later than anticipated. Sometimes overconfidence matters, but at other times it may not matter too much according to the consequences resulting from the error of judgement, which vary according to the task.

Overconfidence arises because people do not realise the true difficulty of a task or judgement. When the judgement is hard people are optimistic as they believe that it is easier than it in fact is; and overconfidence results. Conversely, when the judgement is easy people tend to be cautious as they believe that it is harder than it in fact is; and underconfidence results. People seem to be making estimates of task difficulty which are not extreme enough, they realise if something is hard or easy but are less able to say if something is extremely hard or extremely easy.

People who can anticipate the future may be less confident because they can see many of the influencing factors. Alternatively those who can anticipate many situational factors may feel more confident because they believe they have anticipated them. The truly well calibrated person is one who recognises the true difficulty of the judgement, through life experiences or a good appreciation of what factors influence behaviour, and also allows for the unknown. How and when people learn from their errors of overconfidence and become better calibrated in everyday life is yet to be determined. Becoming aware of the times we are right and wrong in our past judgements may be one step towards more appropriate confidence.

Returning to the definition in the introduction reminds us that confidence is a strength of a person's belief or degree of certainty in something. Thus it must always be borne in mind when we observe confidence in ourselves and in other people that the beliefs that the confidence relates to may be wrong, over- or underconfidence may exist, and it is not a foregone conclusion that reality will tally with the person's perception of it.

REFERENCES

Adams, J. K. (1957). A confidence scale defined in terms of expected percentages. *American Journal of Psychology*, *70*, 432-436.

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review*, *68*, 33-45.

Ajzen, I. (1977). Intuitive theories of events and the effect of base rate information on prediction. *Journal of Personality and Social Psychology*, *35*, 303-314.

Allport, G. W. (1954). *The nature of prejudice*. Reading, Mass.: Addison-Wesley.

Allwood, C. M., & Björhag, C. G. (1991). Mood and realism of confidence judgements of one's own answers to general knowledge questions. *Scandinavian Journal of Psychology*, *32*, 358-371.

Allwood, C. M., & Montgomery, H. (1987). Response selection strategies and realism of confidence judgments. *Organizational Behavior and Human Decision Processes*, *39*, 365-383.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, *39*, 133-144.

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, *37*, 93-110.

Armelius, B. A., & Armelius, K. (1976). Confidence and performance in probabilistic inference tasks with interrelated cues. *Umea Psychological Reports*, *96*, 10.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211-233.

Bar-Hillel, M. (1983). The base rate fallacy controversy. In R. W. Scholtz (Ed.), *Decision making under uncertainty*, (pp. 36-91). Amsterdam: North-Holland.

Baumgardner, A. H. (1990). To know oneself is to like oneself: Self-certainty and self-affect. *Journal of Personality and Social Psychology*, *58*, 1062-1072.

Bieri, J. (1971). Cognitive structures in personality. In H. M. Schroder, & P. Suedfeld (Eds.), *Personality theory and information processing*. New York: Ronald.

Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes*, *58*, 386-405.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception and Psychophysics*, *54*(1), 75-81.

Blackmore, S., & Troscianko, T. (1985). Belief in the paranormal: Probability judgments, illusory control, and the 'chance baseline shift'. *British Journal of Psychology*, *76*, 459-468.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, *49*, 188-207.

Boucher, J., & Osgood, C. E. (1969). The Pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, *8*, 1-8.

Bower, G. H., & Gilligan, S. G. (1979). Remembering information related to one's self. *Journal of Research in Personality*, *13*, 420-432.

Brown, T. A., & Shuford, E. H. (1973). *Quantifying uncertainty into numerical probabilities for the reporting of intelligence* (Report R-1185-ARPA). Santa Monica: The RAND Corporation.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116-131.

Cambridge, R. M., & Shreckengost, R. C. (1978). *Are you sure? The subjective probability assessment test*. Langley, VA: Office of Training, Central Intelligence Agency.

Campbell, J. D. (1990). Self-esteem and the clarity of the self-concept. *Journal of Personality and Social Psychology*, *59*, 538-549.

Dawes, R. M. (1980). Confidence in intellectual judgements vs. confidence in perceptual judgements. In E. D. Lantermann, & H. Feger (Eds.), *Similarity and choice*. Vienna: Hans Huber.

deFinetti, B. (1962). Does it make sense to speak of "good probability appraisers"? In I. J. Good (Ed.), *The scientist speculates*: *An Anthology of partly-baked ideas*. London: Basic Books Inc.

Dubois, T. E., & Cohen, W. (1970). Relationship between measures of psychological differentiation and intellectual ability. *Perceptual and Motor Skills*, *31*, 411-416.

Dunning, D., & Story, A. L. (1991). Depression, realism, and the overconfidence effect: Are the sadder wiser when predicting future actions and events? *Journal of Personality and Social Psychology*, *61*, 521-532.

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology*, *58*, 568-581.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review, 92*, 433-461.

Erickson, B., Lind, E. A., Johnson, B. C., & O'Barr, W. M. (1978). Speech style and impression formation in a court setting: The effects of "powerful" and "powerless" speech. *Journal of Experimental Social Psychology*, *14*, 266-279.

Feather, N. T., & Simon, J. G. (1971). Attribution of responsibility and valence of outcome in relation to initial confidence and success and failure of self and other. *Journal of Personality and Social Psychology*, *18*, 173-188.

Ferrell, W. R. (1994). Calibration of sensory and cognitive judgments: A single model for both. *Scandinavian Journal of Psychology*, *35*, 297-314.

Ferrell, W. R. (1995). A model for realism of confidence judgments: Implications for underconfidence in sensory discrimination. *Perception and Psychophysics*, *57*, 240-254.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fischhoff, B. (1975). Hindsightπforesight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology*: *Human Perception and Performance*, *1*, 288-299.

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

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty*: *Heuristics and biases*, (pp. 422-444). Cambridge, England: Cambridge University Press.

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

Fischhoff, B., & Slovic, P. (1980). A little learning...: Confidence in multicue judgment tasks. In R. Nickerson (Ed.), *Attention and performance VIII*, (pp. 779-800). Hillsdale N.J.: Erlbaum.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology*: *Human Perception and Performance*, *3*, 552-564.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75-90.

Furnham, A., & Steele, H. (1993). Measuring locus of control: A critique of general children's health- and work-related locus of control questionnaires. *British Journal of Psychology, 84*, 443-479.

Garb, H. N. (1986). The appropriateness of confidence ratings in clinical judgment. *Journal of Clinical Psychology*, *42*, 190-197.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506-528.

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraint: Interpretations of self-generated reality. *Journal of Personality and Social Psychology*, *50*, 269-280.

Ginossar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgements about another person. *Journal of Experimental Social Psychology*, *16*, 228-242.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology*: *Learning*, *Memory and Cognition*, *11*, 702-718.

Goldstein, K. M., & Blackman, S. (1978). *Cognitive style*. New York: Wiley.

Goodenough, D. R., & Witkin, H. A. (1977). *Origins of the field-dependent and field-independent cognitive styles*. Princeton: Educational Testing Service.

Gould, S. J. (1993). Assessing self-concept discrepancy in consumer behavior: The joint effect of private self-consciousness and self-monitoring. *Advances in Consumer Research, 20*, 419-424.

Griffin, D. W., Dunning, D., & Ross, L. (1990). The role of construal processes in overconfident predictions about the self and others. *Journal of Personality and Social Psychology*, *59*, 1128-1139.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*, 411-435.

Harris, D. M., & Guten, S. (1979). Health protective behavior: An exploratory study. *Journal of Health and Social Behavior*, *20*, 17-29.

Harris, P., & Middleton, W. (1994). The illusion of control and optimism about health: On being less at risk but no more in control than others. *British Journal of Social Psychology*, *33*, 369-386.

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

Hawkins, J. M., & Allen, R. (Eds.). (1991). *The Oxford encyclopedic English dictionary*, Oxford: Oxford University Press.

Hazard, T. H., & Peterson, C. R. (1973). *Odds versus probabilities for categorical events* (Tech. Rep. No. 73-2). McLean, VA: Decisions and Designs, Inc.

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty*, *4*, 5-28.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology*: *Learning*, *Memory and Cognition*, *11*, 719-731.

Howard, J. A., & Sheth, J. N. (1969). *The theory of buyer behavior*. New York: Wiley.

Howell, W. C. (1971). Uncertainty from internal and external sources: A clear case of overconfidence. *Journal of Experimental Psychology*, *89*, 240-243.

Howell, W. C., & Burnett, S. A. (1978). Uncertainty measurement: A cognitive taxonomy. *Organizational Behavior and Human Performance*, *22*, 45-68.

Humble, J. E., Keim, R. T., & Hershauer, J. C. (1992). Information systems design: An empirical study of feedback effects. *Behaviour and Information Technology*, *11*, 237-244.

Hynes, M., & Vanmarcke, E. (1976). Reliability of embankment performance predictions. *Proceedings of the ASCE Engineering Mechanics Division Specialty Conference*, Waterloo, Ontario, Canada: University of Waterloo Press.

Irwin, F. W. (1953). Stated expectations as functions of probability and desirability of outcomes. *Journal of Personality*, *21*, 329-335.

James, W. (1890). *Principles of psychology*. New York: Holt.

Janis, I. L., & Mann, L. A. (1968). A conflict-theory approach to attitude change and decision making. In A. Greenwald, T. Brock, & T. Ostrom (Eds.), *Psychological foundations of attitudes*. New York: Academic Press.

Johnson, J. T., Jemmott, J. B., & Pettigrew, T. F. (1984). Causal attribution and dispositional inference: Evidence of inconsistent judgments. *Journal of Experimental Social Psychology*, *20*, 567-585.

Johnson-Laird, P. N., & Wason, P. C. (1977). A theoretical analysis of insight into a reasoning task. In P. N. Johnson-Laird, & P. C. Wason (Eds.), *Thinking*. Cambridge: Cambridge University Press.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*, 1-24.

Jones, E. E., & Nisbett, R. E. (1971). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner, *Attribution*: *Perceiving the causes of behavior*. Morristown, N.J.: General Learning Press.

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence of one's general knowledge. *European Journal of Cognitive Psychology*, *5*, 55-71.

Juslin, P., Winman, A., & Persson, T. (1995). Can overconfidence be used as an indicator of reconstructive rather then retrieval processes? *Cognition*, *54*, 99-130.

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

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *TIMS Studies in Management Science*, *12*, 313-327.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty*: *Heuristics and biases*, (pp. 201-208). Cambridge, England: Cambridge University Press.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty*: *Heuristics and biases*, (pp. 509-520). Cambridge, England: Cambridge University Press.

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

Keasey, K., & Watson, R. (1989). Consensus and accuracy in accounting studies of decision making: A note on a new measure of consensus. *Accounting*, *Organizations and Society*, *14*, 337-345.

Kebbell, M. R., & Wagstaff, G. F. (1995). Does the knowledge that an answer cannot be verified alter 'hypnotically' elicited eyewitness' testimony? *Proceedings of the British Psychological Society*, *3*(1), 60.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*(1), 1-24.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, *67*, 95-119.

Kimble, C. E., & Seidel, S. D. (1991). Vocal signs of confidence. *Journal of Nonverbal Behavior*, *15*(2), 99-105.

Kirscht, J. P., Haefner, D. P., Kegeles, S. S., & Rosenstock, I. M. (1966). A national study of health beliefs. *Journal of Health and Human Behavior*, *7*, 248-254.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, *110*, 499-519.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 461-469.

Kogan, N., & Wallach, M. (1964). *Risk taking*: *A study in cognition and personality*. New York: Holt, Rinehart & Winston.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology*: *Human Learning and Memory*, *6*, 107-118.

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay inferences: Effects of impressional primacy, ethnic stereotyping and numerical reasoning. *Journal of Experimental Social Psychology*, *19*, 448-468.

Lenney, E. (1977). Women's self-confidence in achievement settings. *Psychological Bulletin*, *84*, 1-13.

Levin, I. P., Chapman, D. P., & Johnson, R. D. (1988). Confidence in judgments based on incomplete information: An investigation using both hypothetical and real gambles. *Journal of Behavioral Decision Making*, *1*, 29-41.

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution, and monotonicity. *Psychological Bulletin*, *114*, 162-173.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*, 159-183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*, 149-171.

Lichtenstein, S., & Fischhoff, B. (1981). *The effects of gender and instructions on calibration* (Decision Research Report 81-5). Eugene, Oreg.: Decision Research.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty*: *Heuristics and biases*, (pp. 306-334). Cambridge, England: Cambridge University Press.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231-1243.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effect of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098-2109.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, Calif.: Stanford University Press.

Major, B. (1980). Information acquisition and attribution processes. *Journal of Personality and Social Psychology*, *39*, 1010-1023.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, *35*, 63-78.

Maslow, C., Yoselson, K., & London, H. (1971). Persuasiveness of confidence expressed via language and body language. *British Journal of Social and Clinical Psychology*, *10*, 234-240.

May, R. S. (1991). Overconfidence in overconfidence. In A. Chikan (Ed.), *Progress in decision*, *utility and risk theory*, (pp. 67-75). Netherlands: Kluwer Academic Publishers.

Mayseless, O., & Kruglanski, A. W. (1987). What makes you so sure? Effects of epistemic motivations on judgmental confidence. *Organizational Behavior and Human Decision Processes*, *39*, 162-183.

McCarty, P. A. (1986). Effects of feedback on the self-confidence of men and women. *Academy of Management Journal*, *29*, 840-847.

McFarlin, D. B. (1985). Persistence in the face of failure: The impact of self-esteem and contingency information. *Personality and Social Psychology Bulletin, 11*, 152-163.

McKenna, F. P. (1993). It wont happen to me: Unrealistic optimism or illusion of control? *British Journal of Psychology*, *84*, 39-50.

McMahon, I. D. (1973). Relationships between causal attributions and expectancy of success. *Journal of Personality and Social Psychology*, *28*, 108-114.

Moore, P. G. (1977). The manager's struggle with uncertainty. *Journal of the Royal Statistical Society*, *140*, 129-165.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595-600.

Murphy, A. H., & Winkler, R. L. (1971). Forecasters and probability forecasts: Some current problems. *Bulletin of the American Meteorological Society*, *52*, 239-247.

Murphy, A. H., & Winkler, R. L. (1974). Subjective probability forecasting experiments in meteorology: Some preliminary results. *Bulletin of the American Meteorological Society*, *55*, 1206-1216.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest*, *2*, 2-9.

Naditch, S. F. (1976). Sex differences in field dependence: The role of social influence. In *Determinants of gender differences in cognitive functioning*, Symposium presented at the meeting of the American Psychological Association, Washington, D.C.

Neale, M. A., & Bazerman, M. H. (1990). *Cognition and rationality in negotiation*. New York: The Free Press.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation*, *Vol*. *26*, (pp. 125-173). New York: Academic Press.

Newman, R. G. (1984). Children's numerical skill and judgments of confidence in estimation. *Journal of Experimental Child Psychology*, *37*, 107-123.

Nicholls, J. G. (1975). Causal attributions and other achievement-related cognitions: Effects of task outcome, attainment value, and sex. *Journal of Personality and Social Psychology*, *31*, 379-389.

Nisbett, R., & Ross, L. (1980). *Human inference*: *Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Norem, J. K., & Illingworth, K. S. S. (1993). Strategy-dependent effects of reflecting on self and tasks: Some implications of optimism and defensive pessimism. *Journal of Personality and Social Psychology, 65*, 822-835.

Osberg, T. M., & Shrauger, J. S. (1986). Self-prediction: Exploring the parameters of accuracy. *Journal of Personality and Social Psychology*, *51*, 1044-1057.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs*, *76*, (28, Whole No. 547).

Oskamp, S. (1965). Overconfidence in case-study judgments. *The Journal of Consulting Psychology*, *29*, 261-265.

Paese, P. W., & Sniezek, J. A. (1991). Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision-making. *Organizational Behavior and Human Decision Processes*, *48*, 100-130.

Peterson, D. K., & Pitz, G. F. (1986). Effects of amount of information on predictions of uncertain quantities. *Acta Psychologica*, *61*, 229-241.

Peterson, D. K., & Pitz, G. F. (1988). Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cognition* , *14*, 85-92.

Phillips, L. D., & Wright, G. N. (1977). Cultural differences in viewing uncertainty and assessing probabilities. In H. Jungermann, & G. deZeeuw (Eds.), *Decision making and change in human affairs*. Amsterdam: D. Reidel.

Robertson, L. S. (1977). Car crashes: Perceived vulnerability and willingness to pay for crash protection. *Journal of Community Health*, *3*, 136-141.

Rogers, C. R., & Dymond, R. F. (1954). *Psychotherapy and personality change.* Chicago, IL: University of Chicago Press.

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, *35*, 677-688.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*, 193-218.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* Vol. 10, (pp. 173-220). Orlando, FL: Academic Press.

Ross, L. D. (1987). The problem of construal in social inference and social psychology. In N. E. Grunberg, R. E. Nisbett, J. Rodin, & J. E. Singer (Eds.), *A distinctive approach to psychological research*: *The influence of Stanley Schachter*, (pp. 118-149). Hillsdale, NJ: Erlbaum.

Ross, L. D., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control and biases in social-perception processes. *Journal of Personality and Social Psychology*, *35*, 485-494.

Ross, L., & Lepper, M. R. (1980). The perseverence of beliefs: Empirical and normative considerations. In R. A. Schweder, & D. Fiske (Eds.), *New directions for methodology of behavioral science*: *Fallible judgment in behavioral research*, (pp. 17-36). San Francisco: Jossey-Bass.

Ross, L., & Nisbett, R. E. (1990). *The person and the situation*. New York: McGraw-Hill.

Ryback, D. (1967). Confidence and accuracy as a function of experience in judgment making in the absence of systematic feedback. *Perceptual and Motor Skills*, *24*, 331-334.

Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, *7*, 31-44.

Schneider, S. L. (1995). Item difficulty, discrimination, and the confidence-frequency effect in a categorical judgment task. *Organizational Behavior and Human Decision Processes*, *61*, 148-167.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory and Cognition*, *22*, 63-69.

Sen, T., & Boe, W. J. (1991). Confidence and accuracy in judgements using computer displayed information. *Behaviour and Information Technology*, *10*, 53-64.

Shaklee, H., & Fischhoff, B. (1982). Strategies of information search in causal analysis. *Memory & Cognition*, *10*, 520-530.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes*, *42*, 271-283.

Shepperd, J. A., Fernandez, J. K., & Ouellette, J. A. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Journal of Personality and Social Psychology, 70,* 844-855.

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology*, *39*, 211-221.

Shrauger, J. S., & Osberg, T. M. (1982). Self-awareness: The ability to predict one's future behavior. In G. Underwood (Ed.), *Aspects of consciousness* Vol. 3, (pp. 267-314). London: Academic Press.

Shrauger, J. S., & Osberg, T. M. (1985). *The effects of predictions on the reported occurrence of behaviors*. Unpublished manuscript.

Sieber, J. E. (1974). Effects of decision importance on ability to generate warranted subjective uncertainty. *Journal of Personality and Social Psychology*, *30*, 688-694.

Sieber, J. E. (1979). Confidence estimates on the correctness of constructed and multiple-choice responses. *Contemporary Educational Psychology, 4*(3), 272-287.

Sniezek, J. A., Paese, P. W., & Switzer, F. S. (1990). The effect of choosing on confidence in choice. *Organizational Behavior and Human Decision Processes*, *46*, 264-282.

Staël von Holstein, C. A. S. (1972). Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance*, *8*, 139-158.

Stein, A. H., Pohly, S. R., & Mueller, E. (1971). The influence of masculine, feminine, and neutral tasks on children's achievement behaviour, expectancies of success and attainment values. *Child Development*, *42*, 195-207.

Stone, D. N. (1994). Overconfidence in initial self-efficacy judgments: Effects on decision processes and performance. *Organizational Behavior and Human Decision Processes*, *59*, 452-474.

Swindell, L. K., & Walls, W. F. (1993). Response confidence and the delay retention effect. *Contemporary Educational Psychology*, *18*, 363-375.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193-210.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology*, *52*, 700-709.

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

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* Vol. 1, (pp. 49-72). Hillsdale, NJ: Erlbaum.

Vallone, R. P., Griffin, D. W., Lin, S., & Ross, L. (1990). Overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology*, *58*, 582-592.

Wachtel, P. L. (1971). Cognitive style, attention, and learning. *Perceptual and Motor Skills*, *32*, 315-318.

Wagenaar, W. A. (1988). Calibration and the effects of knowledge and reconstruction in retrieval from memory. *Cognition*, *28*, 277-296.

Wagenaar, W. A., & Keren, G. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts. In E. Hollnagel, G. Maneini, & D. D. Woods (Eds.), *Intelligent decision support in process environments*, (pp. 87-107). Berlin: Springer.

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology*. Middlesex, England: Penguin.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, *23*, 63-71.

Webster, D. M. (1993). Motivated augmentation and reduction of the overattribution bias. *Journal of Personality and Social Psychology*, *65*, 261-271.

Weiner, B. (1985). "Spontaneous" causal thinking. *Psychological Bulletin*, *97*, 74-84.

Weiner, B., Frieze, I., Kukla, A., Reed, L., Rest, S., & Rosenbaum, R. (1971). *Perceiving the causes of success and failure*. Morristown, N.J.: General Learning Press.

Weinstein, N. D. (1977). Coping with environmental hazards: Reactions to the threat of crime. *Paper presented at the meeting of the American Psychological Association*, *San Francisco*.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*, 806-820.

Weinstein, N. D. (1984). Why it won't happen to me: Perceptions of risk factors and susceptibility. *Health Psychology*, *3*, 431-457.

Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgements: Two different accounts. *Scandinavian Journal of Psychology*, *34*, 135-148.

Witkin, H. A. (1950). Individual differences in ease of perception of embedded figures. *Journal of Personality*, *14*, 1-15.

Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. (1962). *Psychological differentiation*. New York: Wiley.

Witkin, H. A., Lewis, H. B., Hertzman, M., Machover, K., Meissner, P. B., & Wapner, S. (1954). *Personality through perception*. New York: Harper.

Witkin, H. A., Oltman, P. K., Raskin, E., & Karp, S. A. (1971). *A manual for the embedded figures tests*. Palo Alto: Consulting Psychologists Press.

Wolitzky, D. L. (1973). Cognitive controls and person perception. *Perceptual and Motor Skills*, *36*, 619-623.

Wright, G. (1982). Changes in realism of probability assessments as a function of question type. *Acta Psychologica*, *52*, 165-174.

Wright, G. N., & Phillips, L. D. (1976). *Personality and probabilistic thinking*: *An experimental study* (Tech. Rep. 76-3). Uxbridge, England: Brunel Institute of Organisational and Social Studies.

Wright, G., & Wisudha, A. (1982). Distribution of probability assessments for almanac questions and future event questions. *Scandinavian Journal of Psychology*, *23*, 219-224.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132-156.

Zajonc, R. B. (1968). Cognitive theories in social psychology. In G. Lindzey, & E. Aronson (Eds.), *The handbook of social psychology*. Reading, Mass.: Addison-Wesley.

Zakay, D. (1985). Post-decisional confidence and conflict experienced in a choice process. *Acta Psychologica*, *58*, 75-80.

Zakay, D., & Tsal, Y. (1993). The impact of using forced decision-making strategies on post-decisional confidence. *Journal of Behavioral Decision Making*, *6*, 53-68.

Zimmer, A. C. (1983). Verbal vs. numerical processing by subjective probabilities. In R. W. Scholz (Ed.), *Decision making under uncertainty*, (pp. 159-182). Amsterdam: North-Holland.

Zimmer, A. C. (1986). What uncertainty judgements can tell about the underlying subjective probabilities. *Uncertainty in Artificial Intelligence*, 249-258.