Some influences upon revisions of judgment.

ANDREW FRANKLIN MCCOLL

Some influences upon revisions of judgment.

Andy McColl

Thesis submitted for the award of PhD

Durham Business School

University of Durham, UK

2010

Contents

List of Tables

Chapter 4

List of Figures.

Chapter 1

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Declaration

If approved for the award of PhD, the author (i) permits the University of Durham Librarian to make the deposited copy of this thesis available for consultation for *bona fide* scholars either without delay or after a stated period not exceeding 5 years (ii) be photocopied when it appears to the Librarian reasonable that consultation should be allowed outside Durham, but preferable that the original work should not be leant.

Acknowledgements

With many thanks to George Wright and Fergus Bolger

Abstract

Influences upon judgment revision are issues of both theoretical and applied interest. Many studies in the extant literature have been categorized as Judge Advisor Systems (JAS) research, and algorithmic decompositions of estimation problems. JAS researchers acknowledge the differentiated social roles of advisor(s) and decision-makers; and seek to isolate the influence of advice from *advisor(s)*, upon the deliberations of decision-makers or *judges*. JAS research commonly operationalizes advice in solely numeric terms, which undermines the JAS paradigm's claims of ecological validity. Algorithmic decompositions of estimation problems provides judges with knowledge of the process of advice generation, and differs from advice provided by advisors in JAS studies, as advice is self-generated by users of algorithmic decompositions. The current work sets out why both the JAS paradigm, and algorithmic decompositions are limited (particularly in terms of single judge-advisor information exchange episodes), as means to aid beneficial judgment revision. Six studies are reported that frame, and operationalize research questions that extend understanding of potentially beneficial judgment revision. 'Conformity to advice' emerges as an important explanatory factor in judgment revision. Chapter 4 examines participants' preferences for solely numeric or reasons-based advice, and explores process measures of depth of information search. Participants report an overwhelming preference for reasons-based advice. Chapter 5 investigates the cognitive weighting strategies participants utilize when considering reasons-based or solely numeric advice. Here, participants are insensitive to the type of advice, and discount advice to the same extent - irrespective of type. Chapter 6 investigates the influence of algorithmic decomposition upon beneficial judgment revision. Here, participants were provided with a step-by-step process for solving seemingly intractable estimation problems, or given advice constituted as a testimonial assertion. Results highlight conformity to advice, and the limitations of experimenter generated algorithmic decompositions of estimation problems

of unknown effectiveness. Chapter 7 and 8 sought to develop the idea that algorithmic decompositions *should* influence judgment revision (both for extremely large, and small numerical quantities). Results show that algorithmic decompositions did not facilitate beneficial judgment revision. Instead conformity to advice, irrespective of the quality of advice, was observed only for estimates of large numerical quantities. Chapter 9 was framed as a final attempt to establish if people are able to successfully distinguish between objectively beneficial, and spurious advice. Results indicate that people are unsuccessful in doing so, and find such a task cognitively demanding. Methodological limitations of both the current state of the JAS paradigm, and research involving algorithmic decompositions of estimation problems are identified, in addition to the limitations of the work presented here. Ultimately, methodological suggestions are formulated that may improve understanding of advice giving and taking, in the context of JAS research.

Chapter 1: Introduction

It is uncontroversial to observe that human judgment is sometimes fallible, and that methodologies for enhancing appropriate judgment change are of enduring interest to researchers from a variety of disciplinary backgrounds. The problem of human judgment then, viewed from a psychological perspective, has received much attention in the last 30 years. The insights that have emerged from such a programme of research are valued both in terms of the development of theory and in an applied sense. It is becoming increasingly clear, for example, that human judgment can be modelled upon the idea of the mind being a limited processor of information, and hence sub-optimal when compared to normative judgments based upon logic and statistics. Further, people tend to use cognitive shortcuts, or 'heuristics', that generally work well in simplifying the demands of informational input and cognitive processing, to arrive at a judgment appropriate to the environmental context. However, psychological research has suggested that the use of heuristics may simultaneously introduce systematic biases into human judgment in some circumstances (Tversky and Kahneman, 1982). Given the preceding discussion of the problems of human judgment, it is encouraging to discover that the development of various judgmental aids and strategies has been effective in improving the reliability and accuracy of human judgment. In this thesis, I focus upon two approaches that potentially may be beneficial in improving human judgment – Judge Advisor Systems (JAS) research, and knowledge of the *process* of advice generation. Here knowledge of the process of advice generation is operationalized by providing people with step-by-step algorithms for solving seemingly intractable estimation problems (algorithmic decomposition). First, however it is necessary to look at the insights psychological research has provided about human judgment.

1.1 The problem of human judgment

This thesis concerns human judgment, and some potentially useful ways in which human judgment can be enhanced. Judgment can be defined as inferences people make about the world utilising logic, but based upon imperfect and uncertain informational inputs (Weber and

Johnson, 2006, p55). Judgment can be distinguished from choice, as choosing between dichotomous, or multiple choices, is qualitatively different from exercising judgment. This distinction is based upon good evidence from the extant decision making literature - the mode of response in empirical studies has influenced the decision making process (Payne *et al.* 1993; Gigone and Hastie, 1993, 1997; Hinsz, 1999). This implies that the cognitive mechanisms of the human mind that facilitate choice may be different from the cognitive structures that underpin judgment (Billings and Scherer, 1988). This distinction is important, as what follows in this thesis is concerned with the judgments individual's make. The next section considers the issue of the parameters of human judgment, based upon the analogy of the human mind as a limited information processor.

1.2 Bounded rationality

Empirically, psychologists have studied human judgment by comparing 'actual' judgments to normative benchmarks. In other words, human judgment has often been monitored and compared to an optimum case determined by logic and statistics. Observed deviations from optimum judgments suggest that human judgments are not always facilitated by the application of formal logic and statistics. Often this is because judges may not have access to all the relevant information about a particular issue, or have the time or resources to collate and deliberate upon such information. Herbert Simon (1982) suggested that this behaviour should not necessarily be regarded as 'irrational' (despite the axioms of classical economic theory e.g. von Neumann and Morgenstern, 1944; Savage, 1954). Rather, Simon proposed the notion of 'bounded rationality', whereby people behave rationally within certain cognitive and task imposed constraints. In this view of judgmental behaviour, people tend to 'satisfice' (accept the first satisfactory solution to an issue of judgment), or alternatively utilise 'rules of thumb' or cognitive heuristics in order to make judgments (Gleitman *et al.* 2004; Tversky and Kahneman,

1982). Instead of applying complex cognitive processing of information, only a limited number of simple cognitive heuristics are employed on judgmental tasks, in order to make judgments quickly and easily (Tversky and Kahneman, 1982; Hollyoak and Morrison, 2005). Often these heuristics are sufficient to provide adequate judgments to meet the requirements of a given cognitive task; however, as Kahneman and Tversky demonstrated, the use of cognitive heuristics can also introduce systematic bias into human judgment.

1.3 Heuristics and biases

In a series of insightful studies in the 1970's, Daniel Kahneman and Amos Tversky identified three important heuristics which could account for some aspects of sub-optimal human judgment – representativeness, availability, and anchoring. The idea of 'representativeness' is that people judge the likelihood of an event based on how 'representative' such an event is of a particular category of events or other similar examples (Tversky and Kahnmeman, 1982). Kahneman and Tversky operationalized this idea experimentally by tasking participants with judging the probability that individuals had particular occupations (see below).

> "Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to feel little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense. This personality description has been chosen, at random, from those of 30 engineers and 70 social scientists. What is your probability that Tom W is an engineer?"

Participants typically concluded that Tom W was an engineer – despite the information provided in the description (that only 30% of the population from which the description of Tom W was drawn were indeed, engineers). Hence, Kahneman and Tversky concluded that people are typically systematically insensitive to base-rate information when making

judgments - leading to judgmental error.

A second heuristic that potentially introduces systematic error into the process of human judgment is the 'availability' heuristic (Tversky and Kahneman, 1982). Here, the probability of the occurrence of an event is determined by the ease by which similar instances can be recalled from memory. Events that are perceived to be vivid, recent, or unusual, are easily and quickly recalled from memory and hence are assigned a high likelihood of occurrence. Events that are less easily recalled from memory are assigned a lower probability of occurrence. Therefore a key factor that emerged from this study (and others) is that whilst the availability heuristic serves as an effective strategy in many situations, it can also lead individuals to make systematic errors, particularly when judging frequency (Tversky and Kahneman 1982). In part, this is because some events may be more easily recalled from memory merely because they are unusual or rare; but this implies that such events are inherently *improbable*. Lichtenstein *et al*. (1978) showed that although people in the US made moderately accurate probability estimates about the likelihood of various causes of death, there were some serious judgmental errors. These errors of judgment appeared to have been influenced by contemporary newspaper reports of deaths caused by animal bites and stings. Hence, Lichtenstein *et al*. found that participants seriously overestimated the probability of death by animal bites and stings. Similarly, people regarded death by accident or disease to be equally probable, when in fact death by disease is greatly more probable than a fatal accident.

A third heuristic argued by Kahneman and Tversky (1982) to account for judgmental error is called 'anchoring and adjustment'. Here, Kahneman and Tversky observed that when making estimates of how long a job may take to complete, or when forecasting next month's sales, or

estimating some numerical quantity, individuals tend to fixate upon an initial value and subsequently adjust their estimate from this value. Experimentally, Kahneman and Tversky, manipulated their participants' perceptions by tasking them with making an estimate of the percentage of African nations who were members of the United Nations - by asking participants if the percentage was above or below a randomly generated value. The randomly generated value had a clear effect upon participants' probability estimates – lower random values were associated with lower probability estimates.

However, Gigerenzer has been critical of the work of Kahneman and Tversky, and has argued that it is inappropriate to regard human judgment to be riddled with systematic cognitive biases - instead judgment should be conceived as an adaptive system not bound by formal logic, or the probability calculus (Gigerenzer, 2006). Human rationality, in this view is not measured by its consistency with formal logic and statistics, but rather by its adaptive 'fit to reality'. Gigerenzer strongly argues that measuring human judgment in terms of probabilistic reasoning is mistaken, as probability theory cannot apply to *single events.* Hence, experimentally tasking participants in 1986 with judging the probability that Saddam Hussein would invade Kuwait within five years is a pointless exercise, because this event is singular. Judging the probability that a 17-year old motorcyclist will make an insurance claim for an accident however, is an entirely appropriate question from a probabilistic perspective; as such claims have a historic relative frequency. Gigerenzer then, makes an important distinction between single-event probabilities and relative frequencies. In this view, judgments about the environment are arrived at in a 'fast and frugal' fashion. Judgments are *fast* because they are *frugal* in terms of what information is taken into account.

1.4 Can human judgment be influenced?

Whilst it is important to stress that unaided human judgment is largely 'fit for purpose', the biases and heuristics literature is one way of approaching the issue of how human judgment might be influenced/improved. An important contribution to this debate was first articulated by Meehl in 1954. Meehl pointed out that Statistical Prediction Rules (SPRs) consistently outperformed clinicians who made holistic unaided diagnostic judgments. A great deal of supporting evidence has accumulated in support of Meehl's view over the last half century (Dawes, Faust and Meehl, 1989; Grove, Zald, Lebow, Snitz and Nelson, 2000; Grove and Meehl, 1996). Despite the clear superiority of SPRs and actuarial tables in clinical diagnosis, some clinicians still prefer a holistic approach (Meehl, 1993). Given that unaided holistic clinical judgment is inferior to SPRs on average, health researchers have attempted to design procedural interventions to enhance human judgment (Green and Mehr, 1997). These authors observed decision making procedures in two Michigan hospitals in the US where patients were admitted with chest pains. In these circumstances, physicians must decide whether to send the patient to a coronary care unit, or to a hospital ward where the patient can be observed via electrocardiographic (ECG) telemetry. In approximately 90% of these types of hospital admissions, patients were referred to the coronary care unit. Unfortunately, the consequences of such a high rate of referrals was that the coronary care unit was often overcrowded, resulting in a reduction in the quality of care, and potentially greater health risks for patients.

Health researchers developed the Heart Disease Predictive Instrument (HDPI) in an attempt to relieve the pressure on the coronary care unit. The HDPI consisted of a logistic formula that combined approximately fifty probabilities to arrive at a single probability that could reliably determine whether a patient should be referred to the coronary care unit or not. Should the output value of the HDPI exceed a criterion value, then the patient was admitted to the

coronary care unit, if not the patient was admitted to an observation ward. Unfortunately, this system was avoided by physicians, as it was alien to their intuitive thinking, and in addition, was not well understood. This finding suggests that clinicians are not intuitive 'Bayesian' statisticians (i.e. from the work of Thomas Bayes [1702-61] - clinicians did not identify some probability of a patient's coronary problems *a priori* and update this probability in the light of new information). Indeed, these clinicians may be susceptible to systematic biases in clinical decision-making (Kahneman and Tversky, 1982). Holistic clinical judgment in the two hospitals identified by Mehr (1997) clearly resulted in sub-optimal admission decisions. The introduction of the HDPI, whilst improving the reliability and accuracy of diagnosis, was unpopular with clinicians possibly because it was non-intuitive and demanded cognitive effort in terms of the inputting of probabilities to a formula, and the subsequent calculation of a criterion value.

Based upon the heuristics and biases research findings discussed previously, and specifically upon the idea that people make judgments in a 'fast and frugal' fashion, health researchers developed a further intervention designed to enhance the reliability and accuracy of clinical admission judgments in the two US hospitals discussed previously. A 'fast and frugal' decision tree was produced (see Fig 1), that allowed clinicians to make admission judgments based upon a series of dichotomous choice questions.

Fig 1 Fast and frugal decision tree for coronary care unit allocations (adapted from Green and Mehr, 1997).

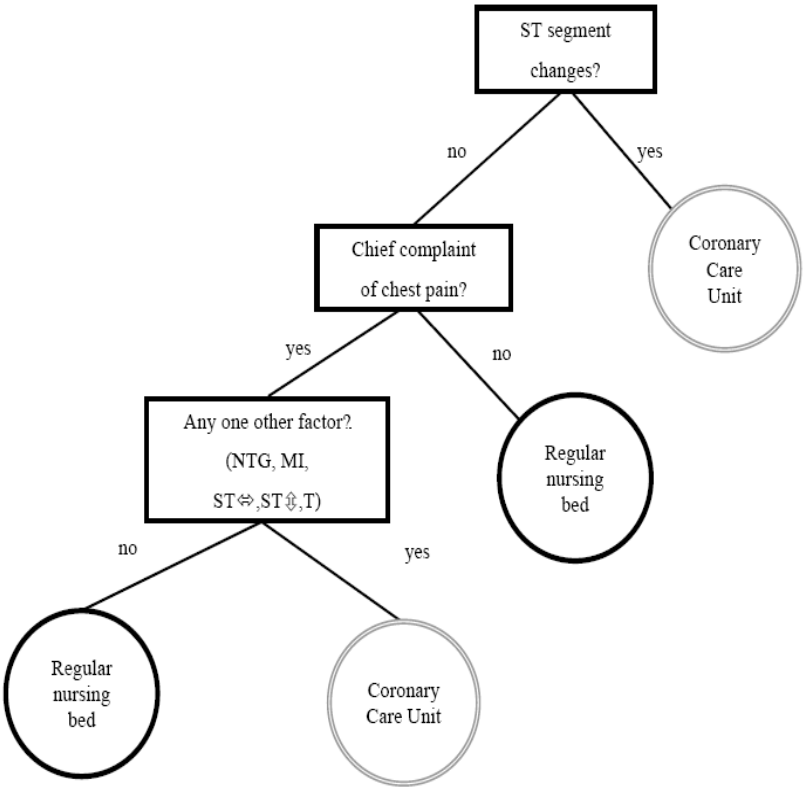


Fig 1 shows that the decision tree is a simple device made up of a series of yes/no questions. The first stage of the decision tree shows the ST segment changes (the ST segment refers to the portion of the ECG trace that can indicate a restriction of the blood supply [*ischemia*]). Here, if a patient has a reading from an ECG that is outside the normal range of typical readings, then the patient is immediately referred to the coronary care unit. No further enquiry is necessary at this stage under this diagnostic system. If the ECG reading is not anomalous, a further cue is considered from the decision tree – whether the patient is suffering from chest pains or not. Should the patient not be suffering from chest pains, then the patient is immediately assigned to a bed in an observation ward. No further diagnostic information is required at this stage.

However, should the patient be complaining of chest pains then a third diagnostic cue is considered from the decision tree (any one other factor?). Here, should the presence of additional factors (in addition to chest pain) be detected, then the patient would immediately be assigned to the coronary care unit. If no additional factors be detected, then the patient would be assigned to a bed in an observation ward. The results of this intervention were encouraging; the proportion of patients correctly assigned to the coronary care unit was greater than chance, but also greater than either holistic clinical judgment, or the HDPI. Holistic clinical judgment (perhaps unsurprisingly) was the least accurate method of differentiating between patients who should be admitted to the coronary care unit, and those who should be assigned a bed in an observation ward. These results suggest that decision aids *can* be devised that are effective in enhancing human judgment.

1.5 Advice giving and advice taking

The preceding discussion has highlighted some of the cognitive limitations of human judgment, and has shown that interventions designed to enhance human judgment can induce appropriate judgment change in certain circumstances. Inducing appropriate judgment change through interventions designed to enhance human judgment is clearly of applied utility, and can also be thought of in terms of *advice*. The preceding discussion of the work of Green and Mehr, (1997), can be thought of as an example of *advice* from a single anonymous *advisor* on how to think about a decision to admit a patient to either a coronary care unit, or an observation ward. The advantages (or otherwise) of utilising advice from other people is central to advice giving and advice taking research (Dalal and Bonaccio, 2006). Here, judgments are perceived as the product of an interaction between a decision-maker and advisor(s). The extant literature in advice giving and taking is relatively recent and underdeveloped, and no comprehensive theoretical account currently exists. This is

unfortunate, as the focus of this thesis is individual judgment change facilitated by consideration of advice. However, two approaches from the extant advice giving and advice taking literature emerge as potential influences upon individual judgment revision, and hence will inform the empirical studies reported later in this thesis - Judge Advisor Systems (hereafter JAS) research, and knowledge of the process of advice generation (algorithmic decomposition). JAS research is significant in this discussion as it is premised upon the idea that few decisions of import are made solely autonomously by individuals. This implies that human judgment - and aids to human judgment - must acknowledge the social context of decisions. On another level, knowledge of the process of advice generation allows people to judge the veracity and validity of advice; if people view the process of advice generation as valid and coherent, then advice generated by this process is likely to be influential. Subsequently, people can either choose to incorporate such advice in their deliberations, or not. This idea is operationalized as 'algorithmic decomposition' in this thesis, and is based on the idea that deconstructing an estimate of some quantity into more tractable sub-component estimates, and mechanically combing the sub-component estimates to produce a final estimate, should result in a less error prone estimate (than an estimate derived from holistic estimation). First, the findings from JAS research are considered.

1.6 Overview of JAS Research

People make numerous important decisions in their everyday lives. However, many important decisions are not made by individuals without the benefit of *advice*. Examples of typical decisions that involve the solicitation of advice are easy to bring to mind – strategic decisions in the context of organizations, decisions by companies to employ people, or the decisions of students over which university they should apply to. Hence, the influence of advice upon decision-makers (hereafter *judges*) has important implications for the quality of decision

outcomes. Of considerable interest to researchers therefore, are the potential improvements in the quality of human judgments potentially facilitated by judgment revision in response to advice (i.e. beneficial judgment revision). This is particularly so where an individual judge may not possess complete decision relevant knowledge, and where advice is formulated in the context of a social interaction (and information is exchanged in a structured manner). Unfortunately, the social context of advice giving and advice taking is an area of enquiry that has not been systematically investigated by researchers in the field of judgment and decision-making (Dalal and Bonaccio, 2006).

Acknowledging the social context of advice giving, and advice taking, is an important development in JAS research, and clearly differentiates the area from other nominal group techniques (such as Delphi – which will be discussed later). This is because JAS research recognises that the ultimate arbiter of advice is the judge, in many typical advice giving and taking situations in the real world; and it is the judge who weights different informational inputs from advisor(s), prior to a decision, and remains accountable for the outcome of any decision. Research into nominal group decision techniques, such as Delphi, has not differentiated between the different contributions that a judge and advisor make to a particular judgment. Hence, JAS researchers, such as Sniezek and Buckley (1989), have claimed that JAS research is of greater ecological validity than other similar group decision techniques. Acknowledging the different roles that judges and advisor(s) occupy in the decision making process places individual (and group) judgment revision in a social context. JAS theorists argue that this adds explanatory power to their analysis of judgment revision. This is because one robust finding in JAS research is that individual judges are disposed to revise their judgments even when advisor(s) are less expert than the people they are advising (Harvey and Fischer, 1997; Yaniv and Kleinberger, 2000; Yaniv, 2004). If it is accepted that judges are

motivated to estimate accurately, then it follows that such judgment revision is due, at least in part, to judges conforming to social factors - such as not wishing to appear to reject advice that is freely given (Sniezek and Buckley, 1995), and sharing responsibility for the outcome of the decision (Harvey and Fischer, 1997; Yaniv, 2004a, 2004b).

Conforming to advice then, may be a powerful influence upon judgment change, but up till now has not been the topic of systematic investigation. I argue that social conformity is a significant and robust influence upon judgment change across a variety of estimation tasks in several task domains – people are constrained in their abilities to distinguish poor advice from good, and in these circumstances, tend to conform to any available advice. This highlights the constraints, and limitations of, current knowledge and understanding of advice giving and taking, in terms of JAS research. Largely, this is due to the hetereogeneity of approaches, concepts, and methods, employed by researchers within the JAS paradigm. Most importantly, researchers have almost unanimously operationalized advice in solely numerical terms - the role of qualitative variables in the process of judgment change is less than fully acknowledged. It is this consideration that motivates the current work, largely because conceptualizing advice in solely numeric terms, in a social context does not appear ecologically valid. This is because it does not appear parsimonious to suppose that people are able to determine the quality, or utility, of advice on the basis of solely numeric advice.

Some of the experimental manipulations in the area of JAS studies have included appeals to expertise (Harvey and Fischer, 1997), agreement and disagreement amongst advisors (Sniezek and Buckley, 1991), and varying the degree of confidence that judges perceive advisors to possess (Sniezek and Buckley, 1995). I argue that such manipulations are really a proxy for

advice quality – and it would probably be more useful (in terms of facilitating beneficial judgment revision) to provide judges with intrinsic information regarding the quality of available advice, in terms of supporting arguments or rationales. As an alternative to the operationalization of advice articulated in the extant JAS literature, I have incorporated qualitative variables within a prototypical JAS experimental framework. These qualitative variables include reasons in support of numerical estimates of varying accuracy, where the true answer was a known quantity.

The issues identified in the preceding discussion are addressed in the first two (of six) empirical studies reported here. Chapter 4 is an exploratory study that frames and operationalizes research questions that concern the preferences people have for different types of advice in a single judge-advisor(s) information exchange episode. Specifically, the study addresses an individual's preference for solely numeric advice, or numeric advice with additional reasons. Moreover, this study explores whether people can separate out 'poor' from 'good' advice when revising judgment or, are prone to the influence of social conformity. Intuitively, reasons-based advice is an important determinant of beneficial judgment revision, and Chapter 5 examines whether people conform to advice (in a single judge-advisor information exchange episode), to the same degree, (when they revise judgment), when in receipt of solely numeric advice, or reasons-based advice. Should people revise their judgements to the same degree, on average, irrespective of whether they receive solely numerical, or reasons-based advice, this would show that people are constrained in their abilities to evaluate advice in a single judge-advisor episode, whilst simultaneously highlighting the limitations of the current JAS paradigm as a lens for studying these issues, and lend support to the idea that conformity is a powerful influence upon judgment revision. Moreover, this study also distinguishes between *accurate* and *inaccurate* numerical advice –

should people not easily be able to recognise 'poor' or 'good' numerical advice - then conformity to accurate advice will result in potentially beneficial judgment revision on average, whilst conformity to inaccurate advice will result in judgment revision that is not beneficial, on average. The second approach that I examine, in terms of potentially beneficial judgment revision, is research premised upon 'algorithmic decomposition'.

1.7 Knowledge of the process of advice generation (algorithmic decomposition)

I argue that providing judges with information about the *process* of advice generation, conveys a means by which judges are able to evaluate advice. Should judges deem that the process of advice generation is logical, and valid, such information should be a more powerful influence upon judgment revision, than factors such as appeals to expertise, or the perceived confidence of advisor(s). The empirical work in Chapters 6, 7 and 8 concentrates upon the idea that knowledge of the *process* of advice generation should be influential in participants' deliberations over the potential benefits of judgment change. Chapter 6 examines the idea that providing participants with knowledge of the process of advice generation (step-by-step algorithms offering possible solutions to four estimation problems provided by an anonymous advisor), could result in greater beneficial judgment revision, than another group of participants in receipt of numerical advice supported by a testimonial assertion from an anonymous advisor (responding to the same estimation problems as the former group). Should the result of this manipulation be that the degree of judgment revision is not significantly different between experimental conditions, it would support the idea that conformity to advice is an important determinant of individual judgment revision. A potential limitation of this study might be that the algorithms provided to the participants were not known *a priori* to be effective. This methodological limitation is addressed in Chapter 7.

Chapter 7 addresses the main methodological limitation of Chapter 6 by providing participants with algorithms that were known to be effective in improving estimation of numerical quantities (MacGregor *et al.* 1988, 1991). These studies show that participants in receipt of algorithmic decompositions of estimation problems beneficially revise their judgments. However, the data can also be explained by the idea that people consistently underestimate large numerical quantities, and consistently overestimate small numerical quantities. Hence, it is not clear whether algorithmic decomposition facilitates beneficial judgment revision, or that conformity - due to the relative positions of the values of pre-advice initial estimate, output-of-algorithm (advice), and the true (but unknown answer) - is influential in terms of beneficial judgment revision.

Chapter 8 examines a further methodological issue. Knowledge of the process of advice generation should enable judges to evaluate the utility of the output-of-algorithm (advice), against the same judge's pre-advice initial estimate. Subsequently, judges should be in a position to decide whether judgment revision will be beneficial. This reasoning should not just apply to estimates of extremely large numerical quantities, but also to estimates of extremely small numerical quantities (relative frequencies). A relative frequency can be defined as the incidence of one instance of an event, from a target population of events (e.g. the chance of winning the UK National Lottery can be expressed as 1/40,000,000). Proponents of the algorithmic approach (MacGregor *et al.*) argue that beneficial judgment revision for estimates of extremely small quantities follows the same process as that of estimates of extremely large numerical quantities – deconstructing the problem into more tractable sub-component problems and mechanically combining these estimates, leads to beneficial judgment revision, on average. Chapter 8 tests this proposition by presenting participants with estimation tasks that involve calculating the relative frequency of rare events, and subsequently providing

participants with algorithmic decompositions (of unknown effectives *a priori*) of these tasks as an estimation aid. Hence, this study adds to an understanding of the circumstances in which knowledge of the process of advice generation can influence judgment revision.

1.8 Discriminating between objectively 'good' and 'poor' advice

Chapter 9 investigates whether people are able to adequately discriminate between objectively 'good' (i.e. numerical advice that is correlated with the target quantity to be estimated), and 'poor' advice (i.e. numerical advice that is not correlated with the target quantity to be estimated). Should people appear constrained in their abilities to distinguish between 'good' and 'poor' numerical advice, then beneficial judgment revision could be expected where people conform to 'good' numerical advice, illustrating the robust influence of conformity.

The final chapter of this thesis will consider how the results of the studies that have been described in the preceding sections, in sum show enhanced understanding of the potential of beneficial judgment revision through the lens of JAS research and research involving algorithmic decompositions of estimation problems. By contextualizing these findings in the extant literature, implications for future JAS studies, and studies involving algorithmic decomposition, will be drawn – both for laboratory based experimentation, and field studies in the real world. Limitations of the current work will be scrutinized, and some methodological improvements will be suggested.

It is to the literature of advice giving and taking that I now turn in order to contextualize the empirical work reported in Chapters 4, 5, 6, 7, 8, and 9.

Chapter 2: Theoretical background

In this section I will review the extant advice giving and taking literature, focussing upon the insights JAS research, and then studies involving algorithmic decomposition, offer to an understanding of aspects of human judgment. Initially, this will involve a discussion of the development of the JAS paradigm. Next, I consider how advice has been operationalized in JAS research, before discussing the merits of incorporating qualitative variables into prototypical JAS experimental manipulations. One theme that emerges from these deliberations is the importance of supporting reasons, or rationales, in addition to solely numeric advice for people tasked with estimating numeric quantities. I next discuss the central findings from JAS research. In this section I draw on research evidence that considers if, and to what extent, people are disposed to follow numeric advice (advice utilization and advice discounting). Subsequently, I consider the degree of confidence (both prior to the receipt of advice, and post-advice) held by judges in typical JAS studies. These findings are relevant to the current discussion as it seems plausible that judges faced with an estimation problem for which a solution is not immediately available, may have little faith in their own intuition. In these circumstances, judges may consider judgment revision (which may be sub-optimal), whilst simultaneously feeling increased confidence. This lack of calibration between ultimate estimation accuracy and confidence is discussed in terms of the findings of 'overconfidence' in the extant literature. Ultimately, beneficial judgment revision results in outcomes that are more accurate, than the accuracy of pre-advice estimates. Clearly then, ultimate estimation accuracy is an issue of interest to JAS researchers, and I discuss some variables that are theorized to mediate judge's post-advice accuracy.

A second theme of interest is the potential of beneficial judgment revision where estimators have access to knowledge of the process of advice generation. This issue draws upon several

studies that have examined the circumstances in which algorithmic decomposition has been shown to be a successful method of estimating intractable quantities. I next consider research that may suggest that people are constrained in their abilities to recognize objectively 'good' and 'poor' numerical advice, before examining some of the methods of enquiry prevalent in JAS research. Finally, this discussion considers research concerned with the Delphi technique of group decision making, and makes clear how insights from Delphi research inform understanding about what factors influence human judgment.

2.1 JAS research and ecological validity

JAS research is not the only approach to acknowledge the social context of decision-making - one stream of research, in the existing extant decision-making literature, focuses on the social context of human decision making in 'small groups' (Kerr and Tindale, 2004). Here, it is acknowledged that few decisions of any import are reached by autonomous individuals; rather the decision-making process is modelled upon the interactions of members of small groups. However, this research is based on the assumption that all group members are 'undifferentiated', (i.e. group members are equally involved, and equally responsible for the group decision). One criticism of the 'undifferentiated' assumption is that it is uncontroversial to observe that in many social structures and organizations, social roles are formalized, which may imply that individual contributions to decisions are often unequal (Katz and Kahn, 1966). This leads to the conclusion that many important decisions are difficult to model upon the idea of individuals acting autonomously, or on the idea that contributions in a group decision task are always undifferentiated. Of interest to decision researchers, therefore, is the growing number of studies that consider how combining one's own estimate(s), or forecasts, with that of others results in changes in judgment. The search for an ecologically valid approach to this issue has lead to the development of the JAS paradigm (Sniezek, 1992; Sniezek and Buckley,

1995; Budescu, and Rantilla, 2000; Budescu, *et al.* 2003).

2.2 The JAS paradigm

The JAS paradigm represents collaborations in which a decision-maker or *Judge,* receives information and recommendations from others in the role of *Advisor*. It is argued that this representation of the decision process is more ecologically valid than models of individual decision making as few decisions of any import are arrived at solely on an individual basis. JAS research starts from the premise that the naturalistic environment of decision makers is inherently determined by their social role (e.g. in organizations, or the marketplace etc.). JAS research models the decision making process in a social context where the person responsible for a decision or forecast (the Judge) seeks or receives some form of input from one or more persons in the role of Advisor. Although there are various motives for using advice from others, the motive of greatest interest is the desire to improve the quality of the decision (Harvey and Fischer, 1997). Whilst acknowledging that both Judge and advisor(s) are involved in the decision-making or forecasting process, JAS assigns the judge as the final arbiter and accountable for the decision or forecast. Hence, the rules for combining individual contributions are at the individual level of the judge and individual contributions from members in the advisor role are known as *advice.* Hence, the JAS paradigm differentiates the social role of judges and advisors in the decision-making process. In sum, the general thrust of JAS research is to isolate the influence of advice upon judges' final decisions.

JAS studies are typically operationalized so that participants are randomly assigned to either the role of the judge, or the role of advisor. Participants are informed that the final decision about a given experimental task rests with the judge, but that advice is solicited from

advisor(s). It is up to the judge whether to accept advice or not, and what weight to attribute to any advice that is accepted, prior to making a final determination. Often (but not always) judges make an initial decision about a given task prior to receiving advice. Advice is then made available to the judge, who is then afforded the opportunity to revise their unaided judgment (if the judge wishes to do so). Next, the issue of how advice has been operationalized in JAS research is addressed.

2.3 Operationaliizing 'advice'

The operationalization of exactly what constitutes 'advice', in the extant JAS literature, is dependent upon the variety of tasks that researchers have utilized to investigate decision behaviour. Advice was operationalized in Sniezek and Buckley (1995), for example, as an advisor's recommendation to a judge; where the participants in an experimental task made a dichotomous choice (i.e. in the United States, 41% of all money spent on food is spent in (A) supermarkets or (B) restaurants?). Here, the advisor's recommendation of choice option (A), or choice option (B), constitutes the advice available to a judge. However, this is not the only way in which advice is operationalized in the extant JAS literature - advice has also been operationalized as the confidence advisors express in their intuitive recommendations (Salvadori *et al.* 2001; Sniezek and Van Swol, 2001), or where advisors provide confidence intervals that include a point estimate (Yaniv and Kleinberger, 2000). Conceptualizing advice as a point estimate (i.e. solely numerical) is also prevalent in JAS research (Sniezek, 1992; Harvey *et al.* 1997, 2000; Budescu, and Rantilla, 2000; Yaniv and Kleinberger, 2000; Budescu, *et al.* 2003; Yaniv 2004b), and is understandably attractive for reasons of mathematical tractability. Yet, this operationalization of advice in solely numeric terms is problematic for reasons that are discussed in the following section.

Clearly, in many everyday situations solely numeric advice is uncommon (i.e. a judge seeking advice upon whether to appoint a particular job applicant, may not find the advice of '77.08' useful in any way). It is perhaps more likely in this scenario that advisors would provide a cogent rationale in support of appointing or not appointing a particular job applicant. Whilst such a rationale, in this scenario, does not necessarily exclude numeric information, it is unlikely to *only* include numeric information. Hence, it is clear that limited progress can be made in isolating the influence of advice, on a decision maker's final judgment, where solely numerical information is the total advice package.

Operationalizing advice in purely numeric terms obscures two main avenues of interest from consideration. First, it is parsimonious to suppose that both in mundane everyday matters, and also in matters of some import, that ostensibly 'quantitative' advice often incorporates reasons that speak to the validity and veracity of the advice, and hence, it is at least possible, for judges in receipt of advice, to evaluate the quality of the advice. If one accepts this premise, then claims made for the superior ecological validity of JAS research are hard to sustain. This is because up till now, the majority of JAS studies have modelled the advice exchange in an impoverished way – in this framework advice has been almost exclusively operationalized in numerical terms. Secondly, purely numerical advice disguises the reasoning process that resolves in a quantitative output. Ultimately, the reasons that support quantitative advice are components of the persuasiveness of the advice package; and the inclusion of qualitative variables (such as reasons), in addition to solely numeric information, is necessary to enhance the claims of ecological validity, advocated by JAS researchers. The potential merits of the inclusion of qualitative variables into experimental manipulations in the context of JAS research is considered next.

2.4 Qualitative advice

Few JAS studies have considered the *qualitative* effects of 'reasons' in addition to solely numerical advice, upon a quantitative estimate of some quantity (as discussed in the preceding section, this is because studies loosely grouped under the label of JAS research, largely conceptualizes advice in solely numerical terms). Even in the few studies where non-numerical advice has been made available to participants, such advice has been impoverished and superficial (Van Swol and Sniezek, 2001). These authors were interested in the relationship between trust, confidence, and expertise, in a typical JAS system. Their design included a 40 item multiple-choice set, that tested their participants' knowledge about computing, and IT issues. Participants were allocated to the role of either advisors or judges, where one party had the power to determine the division of financial rewards. In addition to their recommendations over which of the multiple-choice items was correct, advisors were free to elaborate on their recommendations, by writing comments to the judge. In the minority of instances where advisors elaborated upon their recommendations, such elaborations amounted to little more than comments such as 'I'm definite about this', or 'This is a guess' (p297). This is unfortunate, because a substantial literature outside the area of decision research, points to numerous qualitative variables that can potentially make advice more or less persuasive (Reinard, 1988; Flugstad and Windschitl, 2003). A conceptualization of what exactly constitutes 'advice', then, is an important issue in developing a fuller understanding of advice giving, and advice taking, at the level of individual interactions (Bonaccio and Dalal, 2006). Next I consider the significant findings reported by JAS researchers.

2.5 Central findings of JAS research

Perhaps the most unifying theme in the extant JAS literature is the study of advice utilization

or discounting. Here, advice utilization refers to the extent that judges attend to, and 'follow' the advice of advisor(s), whereas advice discounting refers to the degree to which judges do *not* follow advice. Empirical studies have identified potential advantages to judges in incorporating the informational inputs from advisors prior to making a final determination. Utilizing advice from advisors can enable judges to consider a decision issue in a different way (Schotter, 2003), and can allow judges to consider new information, or alternatives previously unknown (Heath and Gonzales, 1995). Researchers also assert that potential accuracy benefits, in certain decision tasks, are available to judges by integrating advice that originates from multiple uncorrelated sources (Soll, 1999). Further, there are diminishing returns, in terms of accuracy improvements, for judges when the number of uncorrelated sources of advice exceed five or six (Budescu and Rantilla, 2000: Yaniv, 2004a, Yaniv and Kleinberger, 2000). However, JAS authors who considered advice utilization, established that judges incorporate advice in their deliberations not just to improve the quality of judgment, but also to share accountability for the outcome of the decision(s) (Harvey and Fischer, 1997; Yaniv, 2004a, 2004b), and for social reasons (avoiding the impression of rejecting freely offered advice) (Sniezek and Buckley, 1995).

Advice discounting is a particularly robust finding in the JAS literature (Dalal and Bonaccio, 2006). Here, people are either unable, or unwilling, to benefit optimally from advice that is made available to them. Instead people are predisposed to overweigh their own judgment (irrespective of its veracity or accuracy); in contrast to the weight (if any) they place upon advice. Moreover, consistent with the social context argument discussed previously (Sniezek and Buckley, 1995), a 'token shift' (20-30%) of judgment revision on the part of judges, towards that of the advisor(s) has been observed (Harvey and Fischer, 1997; Soll and Larrick, 1999).

Yaniv (2004a, 2004b; Yaniv and Kleinberger, 2000) argue that one explanation of 'egocentric discounting' of advice by judges, can be explained through a differential reasoning perspective. In a situation where a judge is only able to access numerical advice from an advisor, the reasoning behind such advice will be unavailable. Hence, a judge's own initial judgment will dominate any numerical advice. This is because one's own judgment implicitly contains the reasons for a stated numerical judgment, whereas, numerical advice from advisors does not contain the reasons underlying the advice. Egocentric discounting of advice by judges has also been viewed from an 'anchor and adjust' perspective (Harvey and Fischer, 1997; Lim and O'Connor, 1995). Here, it is argued that egocentric discounting of advice occurs because judges 'anchor' upon their initial unaided judgment, or estimate, and insufficiently 'adjust' this estimate post-advice (Kahneman and Tversky, 1974). However, Krueger (2003), and Harvey and Harries (2004) criticize the anchor and adjust account of egocentric discounting, on the grounds that individuals applying the anchor and adjust strategy, are utilizing a temporary and short-term cognitive process to evaluate new stimuli. Instead, these authors point to the concept of 'egocentrism' (a long term belief that one's own judgments are superior to those of others – including those of advisors), as the mechanism that facilitates egocentric discounting of advice on the part of judges. Krueger (2003) argues that people exhibit egocentrism when (i) making judgments about novel situations and (ii) when advice is made available prior to a judge making an unaided estimate. This implies that judges are unable to adequately evaluate advice, and adjust judgment accordingly; and also that a judge's initial unaided estimate is unavailable to serve as an anchor (Clement and Kreuger, 2000). Hence, there are conflicting accounts of how the cognitive mechanisms of egocentric discounting of advice by judges occur in the extant JAS literature. I now turn to the issue of the degree of confidence held by judges about their estimates pre-advice, and post-advice.

Commonly, confidence is defined as the degree to which either the judge expects their decision, or the degree to which an advisor expects their advice, to be correct, in JAS studies. This expectation has been expressed as a probability estimate, in some form of rating on a Likert-type scale, or as confidence interval within which the true answer is contained (Klayman, *et al.* 1999). The levels of confidence reported by advisors will be discussed first. It is clear that judges regard expressions of confidence by advisors, as proxies for the knowledge, expertise, ability and accuracy of advisors (Sniezek and Buckley, 1995: Sniezek and Van Swol, 2001). Moreover, the narrowness, or width of such confidence intervals are argued to be a cue to the advisor's self-perceived task relevant knowledge (Yaniv and Foster, 1997; Yaniv 1997). Judges have been found to more often follow the recommendations of advisors, where advisors express levels of confidence in excess of that held by judges (Lawrence and Warren, 2003; Phillips, 1999; Sniezek and Buckley, 1995; Sniezek and Van Swol, 2001; Van Swol and Sniezek, 2005; Yaniv, 1997). Further, judges have been reported to prefer overconfident advisors to appropriately confident advisors (Price and Stone, 2004). However, there are limits to the extent that confidence can dominate advice quality. Judges perceived advisors' expressions of confidence that were close to 100%, in a probabilistic forecast (dichotomous choice) task, as reckless, rather than confident (Yates *et al.* 1996). Whilst there is evidence of a relationship between the confidence expressed by advisors, and the accuracy of their advice (Sniezek and Van Swol, 2001: Van Swol and Sniezek, 2005); judges' attribution that confident advisors are the source of accurate advice is not always found to be appropriate (Gibbons *et al.* 2003; Phillips, 1999).

The confidence of judges is also of interest to JAS researchers. Less confident judges have been observed to seek advice to a greater extent, than confident judges, prior to making a final determination in a decision task (Cooper, 1991). Post-advice, the degree of confidence held by

judges appears to be influenced by the accuracy of advisors, the extent of information available to advisors, and the degree to which advisors have access to the *same* information (Budescu *et al.* 2003). Agreement between advisors also appears to be a determinant of the level of confidence held by a judge post advice, where disagreement amongst advisors is evident to a judge, the level of post-advice confidence held by the judge is low (Budescu *et al.* 2003; Savadori, Van Swol and Sniezek, 2001). The degree to which a judge processes advice information is a further factor that may influence post-advice confidence, which may explain the common finding that judges' post-advice levels of confidence exceed pre-advice levels of confidence (Heath and Gonzales, 1995; Savadori *et al.* 2001). Such findings suggest that judges may be overconfident in their judgments post-advice.

JAS research has not systematically investigated inappropriate levels of confidence in either judges or advisors. This is regrettable, as there are many studies in the realm of judgment and decision-making that examine both 'overconfidence' and 'under-confidence'. Overconfidence can be defined as the relationship between judgmental accuracy, and the confidence of a judge. If the relationship between judgmental accuracy, and an individual's confidence in their own judgment was perfect, then judgments with 100% confidence would be correct 100% of the time (judgments with 90% confidence would be correct 90% of the time etc.). However, research findings suggest that a perfect match between judgmental accuracy and confidence is not the default position of human cognition (Lichtenstein, Fischhoff, and Phillips, 1982; Alpert and Raiffa, 1982; McClelland and Bolger, 1994; Brenner, *et al.* 1996; Teigen and Jorgenson, 2005). Indeed, these authors report that, participants in their studies were *overconfident* in their judgments, relative to the degree of accuracy they were able to achieve (i.e. an individual participant might report 100% confidence in a judgment, and yet achieve only 75% accuracy, on average). An exemplar of overconfidence is presented by Adams and Adams, (1960).

Here, participants in a spelling task were only correct about 80% of the time, when they reported their confidence as being “100% certain". In other words, participants expected 0% error, when the actual error rate was 20%, on average. Further, confidence exceeds accuracy; markedly so, where judges have incomplete domain knowledge. Alpert and Raiffa (1982) tasked their participants with estimating quantities and providing confidence intervals that included the correct answer (i.e. an upper bound, and a lower bound, within which the true answer can be found), for quantities such as the total egg production of the US economy, or the total number of surgeons and physicians in the Boston telephone directory. Not only did participants report overconfidence in their estimates for these uncommon quantities, but, in addition, this finding was robust, even when participants were explicitly warned of the detrimental effects of overconfidence upon accurate estimation. It is clear that overconfidence is observable where people attempt to make judgments, but have incomplete domain knowledge. The few JAS studies that incorporate some analysis of overconfidence are not inconsistent with the findings reported from the judgment and decision-making literature. Advocates of the JAS paradigm, such as Sniezek and Buckley (1995), for example, observed that judges with little access to task specific information, and who subsequently relied to a large extent on the advice made available from advisors, were overconfident in their judgments. Yet judges who received conflicting advice from two or more advisors did not report overconfidence. Further, interactions between judges involving the exchange of advice, may not necessarily increase the ultimate quality of judgments, but may increase confidence in those judgments, and hence overconfidence (Heath and Gonzales, 1995). The quality of judges’ judgments post-advice is the issue to which I discuss next.

The post-advice decision accuracy of judges is of interest to researchers working within the JAS paradigm, primarily, to ascertain if judges that utilize advice are able to improve upon the

accuracy of estimates, or decisions, made prior to the receipt of advice. Clearly, for beneficial judgment revision to occur judges must have some means to distinguish between 'good' (advice that potentially enhances the quality of judgment), and 'bad' (advice that potentially degrades the quality of judgment). Dalal and Bonaccio (2006) propose three 'core JAS –level variables' (p19) that mediate judges post-advice decision accuracy (i) the amount of task relevant information available to advisors (ii) the accuracy of advice (iii) the weight judges attribute to advice in their deliberations. This articulation of how judges may discriminate between good and poor advice pre-supposes that advisors who possess greater task relevant information are more accurate on average, which subsequently influences the post-advice accuracy of judges, to the extent that they are able to distinguish between good and poor advice. However, judges may be limited in their abilities to distinguish between good and poor advice for several reasons (constraints upon the task relevant knowledge held by the judge; limitations upon human cognition in terms of processing multiple decision relevant cues simultaneously; and the level of motivation held by a judge to achieve optimal accuracy in the decision outcome). However, one variable that does influence the accuracy of decision outcomes is 'feedback', which is discussed next.

Feedback has been shown to influence judges' post-advice decision accuracy in the JAS literature. Where judges receive feedback on the accuracy of their quantitative estimates across multiple trials, decision accuracy has been reported to improve (Fischer and Harvey, 1999). Moreover, where feedback is provided about the accuracy of an advisor who is untypical of a group of advisors (i.e. their advice can be described as a statistical outlier), judges learn to discriminate between poor and good (outlying) advisors (Harries *et al.* 2004). Harvey and Fischer (2005) have proposed that feedback can potentially improve the post-advice decision accuracy of judges, by providing judges with information about how

performance can be improved in a format that is cognitively undemanding. Next I consider the potential of knowledge of the advice generation process (algorithmic decomposition) for judges in terms of beneficial judgment revision.

2.6 Knowledge of the process of advice generation (algorithmic decomposition)

The second approach that is influential upon individual judgment change is argued here to be knowledge of the process of advice generation. Knowledge of the process by which advice is generated *should* enable judges to better discriminate between good and poor advice. Whilst this issue has not been systematically investigated in the advice giving and advice taking (or the JAS) literature, the ideas behind the common notion of 'divide and conquer' provide a useful analogy (Armstrong, et al. 1975; Fischer, 1977; Henrion et al.1993; (Kleinmuntz et al. 1996; Morera, and Budescu, 1998). Here, it is argued that it is less demanding (and less error prone) to produce estimates to the constituent parts of a problem - and subsequently re-combine these responses to form an overall estimate - than to produce an overall 'holistic' evaluation. The 'divide and conquer' principle is widely used to enhance estimation, or forecasting, in various contexts, from medicine to accounting. Early studies in clinical judgment showed that a linear model of clinical judgment outperformed holistic judgment (Meehl, 1957; Goldberg, 1968, 1970).

The divide and conquer principle also underpins 'algorithmic decomposition', a technique used to deconstruct seemingly intractable estimation problems into more tractable sub-component problems, and then subsequently mechanically re-combine sub-component estimates, into a final determination of some quantity (Armstrong, Denniston and Gordon, 1975; MacGregor, Lichtenstein and Slovic, 1988; MacGregor and Lichtenstein, 1991; MacGregor and Armstrong,

1994; MacGregor, 2001). Results indicate that judges benefit from using the approach; for example, auditors' assessments of conditional probabilities were improved by algorithmic decomposition, in comparison to list-type aids, in the context of accounting (Bonner, Libby and Nelson, 1996). Similarly, studies (such as JAS) that task participants to make point estimations of target quantities in almanac-type tasks (where the answer is known to the experimenter, but not to the participant), show estimation improvements for participants utilizing the decomposition approach, above that achieved by holistic evaluation (Armstrong, Denniston and Gordon, 1975; MacGregor, Lichtenstein and Slovic, 1988; MacGregor and Lichtenstein, 1991; MacGregor and Armstrong, 1994; MacGregor, 2001). The results from the preceding studies suggest that people make less accurate assessments through intuition, than they do using simple linear models (Dawes, 1979; Dawes, Faust and Meehl, 1989).

Algorithmic decomposition is theorized to facilitate improved estimation performance for judges, because the attributes of an estimation problem are broken down into various constituent parts, allowing consideration of the attributes of the problem individually, or as sub-groups of attributes. The decomposition of complex or uncertain estimation problems, aids decision making and estimation, because holistic judgments deteriorate over time due to the limits on human information processing capacity. The systematic decomposition of complex or uncertain estimation problems, however, relaxes the information processing demands upon an estimator, potentially allowing a more rigorous consideration of the attributes of an estimation problem, (Fischer, 1977; Kleinmuntz, 1990).

It is not necessary, for the purposes of the current discussion, to accept entirely the explanation of theorists who advocate the algorithmic approach as a route to improved estimation;

however, I argue that judges *should* be able to distinguish between approaches to successful estimation on the basis of knowledge of the process by which advice is generated. In sum, where people are provided with an algorithmic decomposition of an estimation problem, they are able to judge its logical coherence against their own intuition, and come to a view of the utility of the provided algorithm. If the logic of an algorithmic decomposition is accepted by a judge, and the output of the algorithmic decomposition is dissimilar to an individual's unaided holistic estimate, then it is likely that a judge will adopt the output of the algorithm, and subsequently revise their unaided pre-advice estimate of the quantity in question (i.e. appropriate judgment change is induced by the triumph of a superior rationale). People may however, effortlessly and flawlessly, recall accurate information that leads to an optimal estimate, or retrieve sufficient information to self-construct a viable cognitive algorithm that 'solves' the problem sufficient for the purposes of the environmental context, without the need for any external intervention – in such cases the output of an algorithmic decomposition is likely to be ignored. Hence, the process by which advice is formulated provides an important cue to judges over whether advice is beneficial (and should be adopted by the judge in deliberating over an estimate of an uncertain quantity), or whether advice is poor (the advice does not enable a judge to improve upon the judge's unaided estimate, or may even degrade judgment in comparison to the judge's unaided estimate). Whether judges are able to adequately discriminate between 'poor' advice and 'good' advice in a general sense is the issue discussed in the following section.

2.7 'Poor' vs. 'Good' advice

The idea that judges utilize environmental cues in a decision, or estimating task, as proxies for knowledge, has previously been referred to in the discussion of confidence levels held by advisors and judges, both pre-advice, and post-advice. Of interest to the present discussion, is

whether in addition to knowledge of the process by which advice has been generated, people are able to adequately recognize, and utilize environmental cues, to adequately discriminate between good and poor advice. Put simply, can people recognize and utilize objectively good advice, when it is 'staring them in the face'? For judges, the answer to this question depends on several factors (existing knowledge of the decision or estimation task; cognition; motivation). Moreover, where people hold little confidence in their ability to make accurate estimates of target quantities, they may be disposed to attend to *any* advice (irrespective of its quality) from a credible source in attempting to improve estimation accuracy (Yaniv and Kleinberger, 2000; Yaniv, 2004b). These findings suggest that people find distinguishing between potentially beneficial, and poor advice, cognitively demanding. Judges may seek to relax the constraints upon cognition, by recognizing environmental cues that potentially serve as substitutes for direct feedback about the veracity, and accuracy of unaided estimate(s) - recall that 'feedback' is one environmental cue that does facilitate estimation performance improvements. Previous research has shown the influence of trust (Van Swol *et al.* 2001), advisor confidence (Sniezek *et al.* 1995), and source credibility (Reinard, 1988), as some of the cues that are inferred to relax cognitive demands, and potentially influence the estimation performance of judges.

Lee (2007, unpublished) sought to develop the idea that people utilize 'proxy' cues when attempting to make quantitative estimates under conditions of uncertainty. This author reasoned that people may generate quantitative estimates on the basis of ecologically valid proxy cues, and that people will successfully distinguish between potentially beneficial and unhelpful cues. Such a rationalization was operationalized in terms of a task where participants sought to estimate the Infant Mortality Rate (hereafter IMR) of several countries. IMRs are defined as the number of children who die (irrespective of cause) prior to reaching

one year old, per every one thousand live births. Hence, an IMR of 35.8 would indicate that, on average, in every 1000 live births, 35.8 children under the age of twelve months die. Some participants were aided in this task through the provision of numerical information of each of the target countries' per capita (US$) Gross Domestic Product (hereafter GDP). It was theorized that 'national wealth' is an objectively good cue to national IMRs as, on average a nation's GDP is negatively correlated with its IMR. Other participants were similarly aided in making estimates of countries' IMRs; by the provision of accurate population estimates of each of the target countries (there was no correlation between population estimate statistics and IMR). Such an operationalization facilitated a comparison in terms of post-advice estimation accuracy between people in receipt of GDP advice, and people in receipt of Population estimate advice. Results indicated (Study 4) that people in receipt of GDP advice were able to use this knowledge to enhance estimation performance. Participants in receipt of population estimate advice were not able to use this advice to significantly improve estimation performance, on average.

Lee (unpublished 2007), interpreted these findings as evidence that people are able to utilize environmental proxy cues to successfully distinguish between good and poor advice. However, an alternative 'conformity to advice' account can also explain these findings. The 'conformity to advice' account proposes that people are unable to adequately distinguish between good and poor advice, and that it is expected that merely by 'following' advice people in receipt of GDP information (if attended to, and acted upon) would improve performance (the advice is objectively 'good'). Similarly, people in receipt of population estimate information would not significantly improve performance, (the advice is not correlated with IMR), merely by following advice. The idea that in conditions of uncertainty people who revise their judgment post-advice are disposed to incorporate *any* advice (irrespective of its

veracity or quality) (Yaniv and Kleinberger, 2000), is one that will be developed in the main body of this thesis. Methodological issues in JAS research will now be considered.

2.8 Methods of enquiry in JAS research

From a methodological viewpoint, JAS studies are difficult to compare and evaluate due to the heterogeneity of experimental designs and tasks that researchers have reported. However, there are four key areas that emerge strongly from the extant JAS literature – the variety of experimental tasks JAS researchers have utilized; the various conceptualizations of advice; characteristics of, and number of advisor(s); and the characteristics of judges. First, it is important to re-iterate the distinction between tasks involving choice, and tasks involving judgment. Experimental tasks in JAS research have been operationalized in terms of dichotomous, or multiple choice, where an advisor(s) recommendation constitutes 'Choose option A' or 'Choose option C', from an array of possible alternatives (Sniezek and Buckley, 1995; Schrah *et al.* 2004; Savadori *et al.* 2001; Sniezek and Van Swol, 2001; Koehler and Beauregard, 2005). JAS researchers have also utilized tasks that involve judgment - often point estimates, or estimates of subjective probability (Harvey and Fischer, 1997: Harvey *et al.* 2000; Yaniv and Kleinberger, 2000; Yaniv 2004a, 2004b; Budescu and Rantilla, 2000; Budescu *et al.* 2003; Budescu and Yu, 2006). The distinction between choice and judgment is an important one, as there is good evidence from the wider decision making literature that the mode of response in empirical studies has influence in the decision making process (Payne *et al.* 1993; Gigone and Hastie, 1993, 1997; Hinsz, 1999). This implies that the cognitive architecture that facilitates choice may be different from the cognitive structures that underpin judgment (Billings and Scherer, 1988). Critically, for JAS research, "many studies did not provide a rationale for selecting one type of task over another (e.g. judgment over choice, or vice versa)" (Dalal and Bonaccio, 2006, p31). This observation serves to highlight the

underdeveloped status of JAS research, in that it would seem difficult to argue that the experimental findings from JAS tasks involving choice are directly comparable to the JAS findings involving judgment.

Two secondary issues are also relevant to the discussion of the variety of tasks that characterize JAS research – the timing of advice, and task complexity. First, recall that the basic paradigm of JAS research was previously outlined (a judge makes an unaided judgment or estimate prior to receiving some advice, then subsequently makes a final determination over the issue of interest). However, some JAS studies have manipulated the timing of advice so that a judge is prevented from forming an unaided estimate prior to receiving advice (Budescu and Rantilla, 2000; Harvey *et al.* 2000; Sniezek and Van Swol, 2001). Here, judges are tasked with familiarizing themselves with a decision problem whilst simultaneously receiving advice. Whilst such a manipulation is useful where the object of study is to examine how judges aggregate multiple pieces of advice, under conditions of uncertainty (Budescu and Rantilla, 2000), it is less helpful when trying to establish both the amount, and degree of judgment change, and ultimate accuracy of post-advice judgments (Yaniv and Kleinberger, 2000; Yaniv 2004a, 2004b). A further consideration is the degree of complexity of tasks in JAS research. Few researchers have considered this variable in the extant literature, but results indicate that people overweigh advice when deliberating on 'difficult' tasks, but underweight advice when considering 'easy' tasks (Gino and Moore, 2007). Gino and Moore (2007) tasked their participants to estimate the weight of individuals from a series of photographs; task difficulty was varied by presenting clear images in the 'easy' condition, and blurred images in the 'difficult' condition.

The preceding discussion sets out an argument for why JAS studies are limited, and points towards a framework that can be utilized for incorporating qualitative variables into future JAS studies. Clearly, for any advice to be beneficial in terms of a judge's estimation performance, then it must be timely, plausible, and more useful than any estimate that the judge could achieve alone. Advice that does not fulfill these criteria is likely to degrade the quality of estimation – should a judge choose to attend to the advice. Hence, the real problem for judges, when faced with a problem for which they do not know the true answer, (and may not have the time, or resources, to find the true answer), is distinguishing between beneficial advice, and poor advice.

2.9 Delphi - an alternative paradigm in which to examine judgment revision.

This thesis focuses upon the factors that influence individual judgment change, but it is useful to examine what insights about judgment revision have emerged from studies of group decision making. Given this proviso, a similar and more widely recognized technique (than either JAS or algorithmic decomposition) for combining estimates in a forecasting or decision-making context is the 'Delphi' technique. Delphi is a forecasting technique originally developed by the RAND Corporation during the 1960s in the US. The technique is designed to counter the sub-optimal performance of face-to-face interacting groups when making forecasts (e.g. the 'loudest voice' phenomena, 'groupthink', and the formation of political alliances). It is argued that assembling a panel of heterogeneous experts to consider some issue of interest - in which the anonymity of panellists is assured by removing extraneous social factors - may minimise the effects of the 'loudest voice', and curtail the influence of majorities (and minorities) upon individual reasoning. This is facilitated by the elicitation of estimates (often numerical) in response to question(s) of interest. Individuals' responses are collated and analyzed by a facilitation team, and feedback is distributed (usually the mean, or median figure

of the panels' estimates). Subsequently, individuals re-consider their initial estimate in the light of the new information provided through feedback. In addition, those panellists whose estimates fall in the upper and lower inter-quartile range, are required to provide a reason or rationale for their estimate. The process output of several iterations is the aggregate response of the panel.

It is claimed that the success of the Delphi technique is based upon the analogy of the 'Theory of Errors' (Dalkey, and Helmer, 1963). The 'Theory of Errors' explains improvements in judgmental accuracy facilitated through judgment revision in the following way. Delphi studies have typically involved an initial estimation round, following which, an individual participant would receive feedback consisting of the median group estimate. Participants would then be asked to make a further estimate in the light of this new information. This process was repeated until participants' responses became stable - little or no judgment change between iterations. On average responses converged over rounds, and there was a tendency for participants whose estimates were furthest from the group median to change their judgments in the light of feedback. These panellists were categorized as 'swingers'; a second category of panellist was categorized as 'holdouts' as they tended not to change their judgment in the light of feedback. Hence, both the true answer (T), and the median estimate of the group (M) became poles of attraction for participants receiving numerical feedback. Improvements in accuracy could be expected - on average - as long as the pull of (M) did not overshadow the pull of (T).

The 'Theory of Errors' is a possible explanation for improvements in judgmental accuracy where advice is solely numerical. However, even within these parameters a 'Theory of Errors'

does not fully explain all the permutations of 'holdouts' and 'swingers'. It is certainly true that panellists who make accurate estimates of some quantity, and are highly confident in their judgments are desirable in any group judgment seeking accuracy. However, it is also entirely possible that some panellists make inaccurate estimates, but remain convinced of the veracity of such an estimate (i.e. *inaccurate* holdouts). This group could potentially degrade overall group accuracy. Similarly, swingers who make accurate judgments, but are predisposed to revise judgment in the light of panellists' feedback due to low confidence, may also degrade overall group judgment (i.e. *accurate* swingers). Swingers who make inaccurate judgments, but simultaneously have low confidence in such judgments are likely to move towards the judgments of holdouts. Hence, if practitioners of the Delphi technique wish to maximise the accuracy of process output from the Delphi panel, it is desirable to minimise the influence of highly confident panellists who make inaccurate judgments, and minimise the influence of unconfident panellists who make accurate judgments (as the judgments of these panellists are likely to be dominated by more confident panellists).

One way in which Delphi researchers have attempted to enhance the accuracy of the output of Delphi panels is by going beyond solely numerical feedback, and incorporating additional reasons-based advice where a participant's estimate fell in the upper, or lower, inter-quartile range of Delphi panel estimates. The addition of reasons-based advice caused 'significant improvements in accuracy' (Parente and Anderson-Parente, 1987, p136). This may imply that reasons-based advice offers enhanced diagnosticity to participants' receiving statistical feedback, facilitating improvements in accurate estimation. Further, the idea that reasons-based advice may be influential in enhancing judgmental accuracy has an evidential basis; Best (1974) found that for one of two task items a Delphi group given reasons-based advice in addition to means and medians, outperformed a group given only statistical feedback.

Furthermore, Rowe and Wright (1996) compared performance accuracy in a forecasting task between the feedback conditions of iteration, statistical, and statistical plus reasons-based advice. Their findings indicate that participants in the statistical plus reasons condition were less likely to change their estimates over rounds, but when they did change, it was in the direction of greater accuracy. One explanation for these results might be that purely statistical feedback is likely to induce indiscriminate conformity in some participants – they may change their judgment merely to conform to the majority judgment of the group. In contrast, feedback of qualitative reasons may offer greater diagnosticity and hence provide a route to enhanced accuracy. From these findings it is possible to infer that Delphi works partly through iteration (Boje and Murnighan, 1982; Parente *et al.* 1984), where panellists have the opportunity to reflect upon their estimates; and partly through the diagnosticity of the feedback provided (Parente and Anderson-Parente, 1987). This implies that the structured exchange of reasons in a Delphi-like task may lead to greater judgmental accuracy than the structured exchange of statistical information (group means or medians).

These findings indicate that where the social context of nominal group decision making is controlled, the confidence held by panellists in the veracity of their own estimates is an important factor in any subsequent propensity to revise judgment in the light of advice/feedback. Further, where supporting reasons were provided for estimates that were in the upper or lower inter-quartile range of panellists estimates, accuracy improved. Arguably then, there appears to be a case for controlling social influences in small decision making groups if the factors that effect judgment revision are to be examined. If this is so, it may extend to JAS-type manipulations and render the paradigm as an inappropriate one in which to examine the potential factors that influence judgment revision.

However, despite the claims of some researchers, participants in Delphi studies may not be immune from the robust effects of social influence – despite the anonymity of Delphi panellists and non-verbal means of communication. Harvey and Fischer (1997) have shown that people do not weight advice and form judgments - even when presented with exclusively statistical information from anonymous advisor(s) – outside of a social context. These authors examined the factors that determine the influence that (numerical) advice has on judgment utilising the JAS approach. A hypothetical scenario was developed in which participants were tasked with forecasting the potential mortality rate amongst cattle following an outbreak of an unspecified bovine disease. This assessment had further ramifications in that participants were told that these forecasts were to be used to calculate the amount of financial compensation farmers could receive (in reality this was an experimental manipulation). Following a training session, participants responded to pictorial stimuli representing the extent and severity of the disease outbreak, and made a forecast of the number of bovine fatalities. This was facilitated by the display of a circle for 0.5 seconds on a computer monitor: the colour of the circle represented the type of viral outbreak under consideration, and the area of the circle designated the extent of the disease outbreak. Subsequently, anonymous statistical advice and information of the advisor's level of training was supplied to the decision-maker prior to his/her final forecast.

One experimental finding to emerge from this study, is that people are reluctant to dismiss advice from people less knowledgeable than themselves (determined experimentally by level of training) when the advice is freely offered. Even 'experienced' judges (determined by level of training) altered their initial estimates by up to 20% when offered advice from others less knowledgeable than themselves. Normatively this may appear inconsistent with the objective of judgmental accuracy, given that experienced advisors could be expected to possess greater expertise in this particular task, and give more accurate advice facilitating judgmental accuracy

on the part of the judges. That this was not the case may imply that judges make inferences about the nature of statistical advice beyond the numerical information presented to them. Such inferences appear inconsistent with the 'Theory of Errors' that is cited to explain the success of the Delphi technique – more accurate decision makers are hypothesised to 'holdout' in the face of estimates produced by Delphi panellists of lesser accuracy. One insight offered by JAS research for the Delphi method, is that some sources of extraneous social influence persist even in scenarios where judges and advisors communicate non-verbally. This insight may explain some of the sub-optimality in judgmental accuracy that Delphi groups typically exhibit.

In sum, the insights that research into the Delphi technique of group decision making offers to a greater understanding of individual judgment revision, in the light of advice, are somewhat limited. It is true however, that both JAS and Delphi research emphasise the importance of the perceived confidence judges hold in their own judgments, and the relationship between individual confidence and judgmental accuracy, as influences upon judgment revision. Distinct from JAS, Delphi research also highlights the influence of majorities, and minorities upon judgment revision, and it is argued by Delphi researchers that these influences operate through the mechanism of iteration. However, it is clear from the extant Delphi literature that the cognitive mechanisms that facilitate judgment revision are not fully understood. Hence, the 'Theory of Errors' is an incomplete account of group judgment revision. Similar to JAS research, this may be in part due to the largely unacknowledged role of reasons in addition to solely numerical advice. The importance of reasons-based advice in Delphi research has been acknowledged in Delphi studies that have successfully manipulated reasons-based advice to enhance the accuracy of aggregate process output. This insight suggests that reasons-based advice may well be influential in instances of individual judgment revision – unfortunately the

extant JAS literature reveals few examples of such a manipulation. A further insight about judgment revision that is addressed by the extant Delphi research concerns the social nature of advice giving and advice taking – temporarily isolating individuals for the purposes of experimentation is unlikely to entirely remove the social context in which judgments are made. In this scenario, the JAS paradigm may offer greater ecological validity for researchers than Delphi, as typical JAS manipulations are designed to recognise the different social roles of judge and advisor(s).

In the next section I consider the methods that were utilized in carrying out the experimental work reported in Chapters 4 - 9, and a justification for the methods used.

Chapter 3: Common Methods

3.1 Introduction

Six studies are reported in Chapters 4-9 of this thesis, where various experimental manipulations are carried out to investigate the central issue of this work – appropriate judgment revision. The purpose of this section is to argue why the experimental methods used were centred within a philosophically positivistic paradigm, and further why an interpretivist methodological approach would be inappropriate for the purposes of the current work. Further, justification is provided for the general approach adopted in this thesis. Next, consideration is given to the size and characteristics of the population sample that participated in the current experimental work. Procedural issues are then addressed, before consideration of inferential statistical tests is discussed.

3.2 Philosophical assumptions

The research area that is the focus of this investigation concerns ways in which appropriate individual judgment revision can be induced. Several key assumptions underlie the approach taken to investigate questions of interest within the parameters of the research area. In terms of ontology, this study recognizes an objective reality independent of individuals and individual cognitions (Morgan and Smircich, 1980). However, further clear ontological distinctions are somewhat more difficult to assert. The notion of an entirely 'harmonious' reality is brought into question. 'Irrationality' in some definitions would seem to imply a degree of conflictual reality - at least at the individual psychological level. Hence, although there is a strong assumption of 'atomism' (understanding the whole on the basis of the parts) there is also recognition of a conflictual reality in so far as dissonant cognitions can be said to lead to sub-optimal judgment (Psillos, 1999). The epistemological assumptions that underlie this approach

are almost entirely positivist; whereby knowledge is represented as real and precise (Popper, 1959). The positivistic paradigm has been chosen in the context of the parameters of the area of study for several reasons. Firstly, it is consistent with the aforementioned ontological assumptions that are made about the nature of reality. The implications of the key fundamental assumptions and the consequent ontological and epistemological positions adopted for the proposed research are distinct. The scope of the research is bounded both by the area of interest and the positivist paradigm through which it is examined. In essence, the measurability of the phenomena under investigation is critical in deciding what to include in the study and what to ignore. A research design incorporating a hypothetico-deductive methodology indicates the development of testable propositions that will either be supported by data analysis or not supported (Popper 1959). Moreover, a significant amount of data from participants will be responses in numerical form. This leads directly to the use of both descriptive and inferential statistical techniques for the analysis of data. However, other forms of responses are not precluded from the analysis and form important parts of the rationale and argument developed through the study.

A fundamental ontological assumption that underpins the current work is that reality is objective, not subjective. Reality can be perceived as - at least - external to individual consciousness (Myers and Avison, 2002; Burrell & Morgan, 1979; Hirschheim, 1985; Hirschheim and Klein, 1989). An illustration might be that gravity exists whether an individual likes it or not. Although gravity is invisible, no matter where any individual person may be in the world, gravity's effects are essentially the same and not variable by an individual's subjective evaluation of it. Despite this assertion, the ontological assumption that objective reality is essentially harmonious is questioned (Chalmers, 1999). It does not appear impossible to suggest that at both the individual psychological level and as part of any

organizational culture, that a conflictual reality exists. An objective definition of reality has epistemological implications in so far as knowledge can be represented as real, certain and precise. In broad terms such an approach may be characterized as philosophically objective and positivistic in terms of a scientific paradigm (Donaldson, 2003).

Whilst acknowledging that there may be alternative ways of approaching this area of interest through different paradigms and methodologies, positivism offers some attractive qualities. It seeks to induce generalizations or laws by gathering and ordering data, developing explanatory theories, deducting and testing hypotheses to test theories, which in turn either supports theory or leads to an adjustment of theory. What positivism does not seek to do is to provide certain 'proof' of the correctness of a proposition (Donaldson, 2003). Hypothetico-deduction aims to manipulate one or more features of reality (the independent variable[s]) whilst holding everything else constant. If this manipulation results in a measurable effect, upon the dependent variable above the level of chance occurrence, then the hypothesis is said to be supported. Positing the null hypothesis - any effect is due to chance occurrence - usually tests hypotheses (Lee, 1991). By disconfirming the null hypothesis, support is given to the initial hypothesis (Popper, 1959). Research conducted through a positivist paradigm is incremental and cumulative. Studies are rigorous and replicable, demonstrating the robustness of the propositions tested and the establishment of 'facts'.

The consequences of these ontological and epistemological positions for the area of research are philosophical and pragmatic. A positivistic paradigm - by its very nature - excludes many forms of data (which may include elaborated social interactions) that may well be important research issues. However, by concentrating on real and demonstrable phenomena positivism

provides a measurable and testable basis for inferences about reality. Importantly, positivism makes a distinction that the researcher is largely independent from the phenomena under investigation (Donaldson, 2003). Pragmatically, the self-imposed boundaries that positivism sets itself, encourages a focused and achievable research project. The limitations of positivism could be described as strengths rather than weaknesses and have some bearing upon the scope of the current project. The scope of this study encompasses the development of theory from the extant literature, formulation of a series of testable (and falsible) hypotheses, experimental studies involving samples from selected populations of people, data analysis and tentative conclusions. The study is empirical, and involves (largely numerical) data analysis. Descriptive and inferential statistics are utilized as means to interpret results.

3.3 Alternative philosophical approaches

Despite the strength of the positivist position outlined above it is certainly possible to investigate the research area through the lens of another philosophical research paradigm. However, a possible caveat might be that one would have to formulate variations on the questions asked, to fully benefit from any differing approach. Interpretivism - as a research philosophy - differs in many ways from the fundamental assumptions that underpin the positivist paradigm (Denzin and Lincoln, 1994). Ontologically, interpretivism regards reality as subjective and as socially constructed. Individuals are regarded as 'sense-making' participants rather than as objects of study and as such there are multiple, local and specific 'constructed realities'. Crucially, an interpretivist paradigm views the researcher as an integral part of the socially constructed reality under investigation - and it is arguably the case that it is difficult for an observor to perceive another person entirely separate from social reality (Khlentzos, 2004). Researchers in the interpretivist paradigm view knowledge as a social constructed process produced through the interaction of enquiry and participation (Orlikowski

and Baroudi 1991). Hence, how revisions of judgment occur that arise from factors such as politics, society, culture, ethnicity, and gender are all relevant questions that could profitably be pursued through an interpretivist framework. Interpretivism could fruitfully explore the socially constructed foundations of judgment and judgment revision in terms of individual and organizational meaning. The consequence for such a project ontologically and epistemologically would be a subjective view of reality encompassing differing socially constructed multiple realities (Wenger, 1998). Analysis would involve questions of knowledge representation, and a conceptualization of multiple levels of 'three-dimensional' meaning for participants.

There are many compelling reasons why researchers might choose interpretivism as a paradigm in which to investigate influences upon appropriate judgment revision, judgmental biases, and (perceived) irrational decision-making. However, interpretivism has limitations that would exclude some avenues of enquiry. Interpretivism - almost by definition – is concerned with subjectivity, and hence seeks to understand social reality often by interacting with people better to understand the subjective meaning people may attach to judgment and appropriate judgment revision, in the environment in which these issues are addressed on a day-today basis (ecological validity). However, in practice, researchers adopting an interpretivist approach face considerable obstacles. First, the integrity of any data collected where a researcher is (i) part of a complex social interaction and (ii) perceiving and imposing meaning upon real-time and real-world events whilst simultaneously trying to collect data is open to question. Second, the practicalities of data collection for researchers using an interpretivist approach limit samples of people from any given population, to sole individuals or very small groups of people – hence it is difficult to see how findings can be generalized to a wider population. Third, given the unique quality of any specific social interaction, it is very

unlikely that findings from one interpretivist study can be replicated by the findings of another. Interpretivist researchers may possibly counter these criticisms by claiming that as their subject matter is subjective meaning, then issues of validity, generalization, and replicability do not in any way undermine the findings of their studies. Without elaborating upon the circularity of such a position, or the historical debate between the merits of positivism and interpretivism further, it is parsimonious to assume that in order to study influences upon human judgment and appropriate judgment revision, a useful starting point might be to adopt an approach where judgment can be measured and judgment revision can be verified.

A further limitation of an interpretivist project seeking to determine (some) of the influences upon judgment and appropriate judgment revision, is (in purely practical terms) its size and complexity. Such a project would probably be far wider in the nature of participant involvement and the types of data sought than a project centred in a positivistic framework. Adopting an interpretivist approach would probably involve high levels of discourse analysis following in-depth interviews with participants (Edwards and Potter, 1999). In this context any research design is likely to be multi-faceted and 'immersive' in that the researcher would become part of what one is trying to study (Agar, 1996). A further difficulty is likely to be that theoretical constructs are difficult to define to a singular meaning and would probably be subject to a form of hermeneutics as a way of coping with what is meant by meaning. The type of data arising from studies utilizing the interpretivist paradigm is likely to be qualitative in nature, and indeed may raise questions over what can and cannot be classified as research data. Data analysis is likely to be multi-layered given the ethnographic and phenomenological nature of interpretivist investigations. The interpretation of results derived from the interpretivist approach is by nature somewhat subjective, involving extensive discourse analysis and hermeneutics (Agar, 1996), and as such is difficult to analyze to determine the

influences upon judgment and appropriate judgment revision.

As has been shown, the issues of judgment and appropriate judgment revision can be approached from more than one research paradigm. However, there are both philosophic and practical reasons that dictate that a positivistic approach may be more fruitful in investigating the phenomena under discussion in this thesis. 'Rationality' in the context of judgment can be a measurable construct in terms of experimentation (i.e. the degree to which people underweight information that could potentially improve judgment). Hence, systemic deviations from normative models - biases - can also be measured. This may imply that underlying processes or causes may be discoverable through a systematic and rigorous approach encapsulated by positivism. Interpretivism, in contrast, would serve as an extremely useful tool in exploring the sociology of judgment, but would appear less useful in examining both the individual and group psychological processes that lead to sub-optimal judgmental outcomes.

3.4 Justification of the general approach

The preceding section examined the philosophical basis of the positivistic approach adopted in this thesis to the issues under consideration. In this section, the focus is narrowed to justify the general experimental approach adopted in this thesis, with specific reference to methods that are utilized widely in the advice giving and taking extant literature. As the issue of interest in this thesis is the process of judgment change, both in terms of amount and degree, that can be attributed to the provision of reasons-based advice, the studies were designed to investigate human judgment, rather than choice (as discussed in the preceding literature review the cognitive mechanisms theorized to facilitate judgment are not the same as the cognitive

architecture that enables choices to be made). Recognising the distinction between judgment and choice has implications for the design of empirical studies, in that participants in the studies reported here were *not* asked to make choice(s) from an array of potential alternatives. Instead, participants were asked to make judgments (numerical estimates) of uncertain quantities. This basic premise was useful in terms of mathematical tractability, and allowed the use of a variety of statistical techniques (discussed later) to be employed to explore relationships between variables. The issue of judgment change also implied (in terms of experimental design) repeated measures (at least two measures are necessary to measure judgment change). Here, experiments in JAS research provide a useful template - a judge makes an unaided estimate, then receives advice, and subsequently is able to decide whether to revise his/her unaided estimate with the benefit of advice (Dalal & Bonaccio, 2006). Albeit with variations that encompass the idea that knowledge of the process of advice generation (algorithmic decomposition) should influence judgment- the studies reported in the empirical chapters follow this basic JAS paradigm. Consideration is now given to characteristics of the population from which participants were drawn.

3.5 Participants

In all, 834 people participated in the six studies reported in this thesis. These participants were drawn from staff and students at Durham Business School, University of Durham, UK. Participants were mostly undergraduate students undertaking courses of study at Durham Business School, and many were non-native English speakers. It was assumed that the linguistic requirement of the University for admittance to undergraduate, and post-graduate, courses was sufficient to allow the linguistic abilities of the sample to be treated as one population.

3.6 Recruitment of participants

Participants (N = 73) in the first study reported in this thesis (Chapter 4) were made up entirely of academic and non-academic staff from Durham Business School who held a valid email organizational email address. An invitation email was sent to all Durham Business School staff, and respondents became participants in the study by self-selection (if staff wished to participate they clicked on an internet link within the invitation email that subsequently opened the online survey). This process meant that participants were recruited without the physical presence of the experimenter.

The recruitment of participants to the study reported in Chapter 6 followed a different process. Here, the sample was made up entirely of undergraduate students (N = 255) that I had access to through my teaching role at Durham Business School. Here, questionnaire booklets were prepared and distributed to large groups of students in lecture theatres and classrooms at both Durham, and Queen's Campus (Stockton). Participants completed the questionnaire booklets individually, and returned them to me on the same occasion.

Participants recruited to the study reported in Chapter 7 made their responses both online (N = 60) and on paper (N = 82). Participants responding online were recruited by the distribution of an email to Durham Business School staff that contained an internet link, clicking on which allowed staff to become participants in the study. Similarly, students responding online were recruited by the distribution of an email to students that the author had access to through his teaching responsibilities – by clicking on an internet link in the email students were able to become participants in the study. In both cases participants were 'self-selecting' in so far as respondents could either choose to become participants, or ignore the invitation. Participants

responding online were incentivized to take part in the study by receiving an Amazon voucher worth £4 for completing the online survey. Participants responding on paper received no incentive. Participants responding online made their decision over whether to take part or not, without the presence of the experimenter, participants responding on paper completed the questionnaire in the presence of the experimenter.

Durham Business School staff and students (N = 105) taking part in the study reported in Chapter 7, completed and returned questionnaire booklets in the presence of the experimenter. Staff were approached directly by the experimenter and asked to participate in the study; students were asked to complete and return the questionnaire booklets during lectures. There was a far greater number of undergraduate student participants, than staff.

The recruitment of participants in the study reported in Chapter 8 of this thesis excluded staff and only included students. Here, the experimenter received permission from colleagues to invite students who were attending lectures to become participants in the study. Questionnaire booklets were distributed and completed by individual students (whilst in the proximity of many other students) in lecture theatres in the presence of the experimenter.

Similarly, in the final study of this thesis reported in Chapter 9, Durham Business School staff and students were invited to become participants in the study by completing a questionnaire booklet. Identical to the recruitment strategy employed in the study reported in Chapter 7, staff were approached directly by the experimenter and asked to participate in the study; students were asked to complete and return the questionnaire booklets during lectures. There was a far greater number of undergraduate student participants, than staff.

3.7 Instruments

As the data to be collected from participants was predominantly numerical, and as numerical data is conveniently captured by 'closed questions', questionnaires were formulated for each experimental manipulation, so that raw numerical data was available for analysis after the completion of the experiment. Data was collected from participants 'electronically' in Chapter 4 and Chapter 6. Data was collected in the traditional 'paper and pencil' questionnaire format in Chapter 5, 7, 8 and 9. First, consideration is given to 'electronic' experimental instruments.

Participants submitted their responses 'on-line' in Chapter 4. Chapter 4 was designed to explore whether participants sought out reasons-based advice, in preference to solely numerical advice, when tasked with estimating the dates of historical events. Participants responded to the historical quiz on-line, by mouse-clicking a hyperlink in an email that invited Durham Business School staff to participate in the experiment. The on-line interface was adapted from 'MouseLabWeb' process tracing software (Willemsen and Johnson, 2004), and data was electronically recorded in a MySQL database (see Appendix I for details of the on-line web pages). The advantage of using 'MouseLabWeb' software was that it allowed construction of a distinct sequence of web pages. Once a participant had accepted an electronic invitation to participate in the experiment, the same participant was directed to a web page, where the participant input their name into a text field. By clicking on a 'Next page' icon, the participant was then presented with an estimation problem (i.e. In which year did event x take place?), and directed to enter their estimate in a text box. On clicking another 'Next page' icon, the participant was next presented with a digital information board, where participants could examine all, some, or none, of eight information boxes. Four information boxes contained solely numeric advice, the remaining

four information boxes contained numeric advice and additional supporting reasons (of four distinct types). The participant was then free to choose to revise their unaided estimate, or continue without revising their unaided estimate, in the light of the information available to them from the eight digital information boxes. The participant then clicked one from an array of nine radio buttons indicating which piece of advice (the option 'No advice useful' was available), was most useful to them in formulating an answer to the estimation problem. Participants could then revise their unaided estimate, and enter a new estimate in the provided text box. On clicking the 'Next page' icon on this page, a participant was presented with a text box in which they were invited to type in text explaining why they had, or had not, revised their unaided estimate. A further web page thanked participants and exited the experiment.

Similar to the instruments described in Chapter 4, the questionnaire in Chapter 6 was programmed in 'aspx', (by David Challener, IT Consultant, Waterstons). Here, participants were tasked with making estimates of almanac type quantities (e.g. the number of passengers that transit through Durham Tees airport annually). The questionnaire tasked participants with making an unaided estimate, advice was then made available (by mouseclicking 'next page'), and participants could then choose whether or not to revise their unaided estimate by incorporating advice into their deliberations. Data was captured and stored in an electronic database, and could be subsequently recovered for the purpose of analysis. Again, participants were able to access the questionnaire from a computer terminal by 'clicking' on a link in an email invitation. As the author is unfamiliar with many aspects of the appropriate computer programming languages necessary to fully take advantage of this media, the traditional 'pencil and paper' questionnaire form was used in Chapters, 5, 8, and 9. One issue that needs to be addressed at this juncture is whether

participant responses systematically vary according to whether people provide responses in a ‘pencil and paper’ format or online.

There has been considerable research on whether there are systematic differences in responses depending on whether people respond online, or in the more traditional ‘paper and pencil’ format. Whilst the issue has not been definitively settled there is considerable evidence that there are few systematic differences between the responses of people online, and the responses of people to an identical questionnaire presented in ‘hard copy’ (Chuah, Drasgow, & Roberts, 2006; Cronk & West, 2002; Knapp & Kirk, 2003; Pettit, 2002; Truell *et al.* 2002). Clearly online surveys offer researchers many practical advantages in terms of data collection, and where differences have been found between surveys presented online and paper-based surveys, they are often connected to an enquiry linked to sensitive personal issues which may attenuate a social desirability bias in respondents (i.e. respondents may answer questions in a way that appears socially desirable, rather than answering honestly) (Turner, Ku, Rogers, Lindberg, & Pleck, 1998). This thesis does not seek to collect sensitive personal information from participants, and hence assumes that the responses from participants to my experimental manipulations can be pooled for the purposes of analysis.

The study in Chapter 5 was conducted entirely in the traditional ‘paper and pencil’ format of questionnaire surveys. Here, a 14-item questionnaire was formulated that tasked participants to ascertain the correct date - or a best estimate of the correct date - of historical events that occurred in the last 300 years. In addition, two conditions were derived under which advice would be available to participants: First, numerical advice (accurate/inaccurate), second, numerical advice and supporting additional reasons (very

weak/weak/strong). The categorization of supporting reasons was based upon the Elaborative Likelihood Model theory (ELM) advocated by Petty and Cacioppo (1986). Here, the experimenter (or pilot group, if practical) generates convincing and specious arguments about the question(s) of interest. These reasons are then independently rated on a 5-point Likert scale of *persuasiveness* (1 = unconvincing; 5 = compelling). In ELM studies of *message persuasiveness* these ratings would then be passed to a second rating group who would write down how favourable or unfavourable these ratings are to the advocated reasons (thought listing) i.e. very weak/weak/strong. As reasons are of interest and not persuasive messages, only the first step of Petty and Cacioppo's methodology was retained. Subsequently the reasons generated by the experimenter were categorized as 'very weak', 'weak', and 'strong' by two independent coders. These categories serve as methodological proxies for the argumentative quality of the generated reasons. Participants were given no information regarding the identity of the advisor beyond the description, 'advisor *C*', or 'advisor *P*'. For each of the fourteen questionnaire items in the two advice conditions the alphabetic designation of the advisor was varied (creating the illusion of fourteen different advisors). The questionnaire items were counterbalanced across participants and conditions in order to offset any systematic presentation order effects of the questionnaire items. The experimental task was divided into two phases; initially participants were tasked with estimating the correct date (or best estimate) for a questionnaire item; secondly, participants were tasked with re-considering their estimate in the light of available advice. Each phase of the questionnaire was printed on a separate sheet of the questionnaire booklet; so that advice was not available to participants in phase one. Participants were instructed to attempt each of the 14 questions sequentially, so that on completion of the first experimental phase participant's proceeded to the second phase, where a final estimate was made in the light of the available advice.

Chapter 6 was devised to compare the estimates of people who received advice in the form of a testimonial assertion, with the estimates of people who received advice in the form of a step-by-step procedure (algorithmic decomposition) to formulate estimates to almanac questions. The almanac questions were selected on the basis of issues where a true answer existed, but is not widely known. Consistent with the manipulations in the preceding study, participants made an unaided estimate, were exposed to advice from an anonymous advisor, and could then revise (or not revise) their unaided estimate with the benefit of advice. Some participants responded on-line, by clicking a hyperlink in an invitation email, others completed a questionnaire booklet 'by hand'. One important distinction was that people in receipt of algorithmic decompositions (recommended by advisors as a method of forming an estimate) of the estimation problems, ultimately produced their own advice (output-of-algorithm), whereas participants in receipt of testimonial assertions were provided with a fictitious advisor's rationale.

Chapter 7 sought to extend the idea that knowledge of the process through which advice is generated, is a cue that may facilitate enhanced estimation performance. Participants completed one of three versions of the questionnaire booklet. Initially, participants made an unaided estimate of the number of pieces of mail handled by the US postal service in 1989 (hereafter the US Mail problem), or the number of Forested Miles in the US state of Oregon (hereafter the Forested Miles problem). Subsequently participants were randomly allocated to (i) a control condition (participants made an unaided estimate, completed a distractor task, and then were tasked with revisiting their unaided estimate) (ii) a condition where an algorithmic decomposition of the US Mail problem constituted advice (iii) a condition where an algorithmic decomposition of the Forested Miles problem constituted advice. No

participant was allocated to more than one experimental condition. The number of 'steps' in the algorithmic procedure differed between conditions.

Chapter 8 examined the issue of whether knowledge of the process of advice generation was a means by which participants could evaluate the utility of advice, where the numerical quantities to be estimated were extremely small (i.e. relative frequencies). Participants in Study 5 completed questionnaire booklets by hand. Initially all participants made unaided estimates of either the incidence of multiple maternities, or the incidence of pedestrian fatalities as the result of road traffic accidents (no participant estimated the incidence of multiple maternities and pedestrian fatalities). Participants were randomly allocated to the treatment conditions of estimating the relative frequency of multiple maternities, or the relative frequency of pedestrian deaths attributed to road accidents in the UK, in the same way that participants had been allocated to experimental conditions in Chapter 7. After inputting information into the algorithms (or completing a distractor task for those participants allocated to a control condition), participants could, if they so wished, revise their unaided estimate with the benefit of advice (output-of-algorithm).

Questionnaire booklets were completed by hand by participants in Study 6 - the same pattern of unaided estimation, provision of advice, and revision of judgment was utilized in a similar fashion to previous manipulations. Although participants in Study 6 who received advice were given advice that (i) was positively correlated with the target quantity, another group of participants was given advice that was (ii) not positively correlated with the target quantity. This was to investigate whether participants were able to adequately distinguish between potentially beneficial advice, and poor advice, on average.

3.8 Common Measures and Data analysis

Participant's estimates of target quantities were recorded pre-advice, and post-advice. An ordinal measure of 'perceived utility of advice' was recorded on a 5 point Likert scale in Chapters 5-9. In Chapter 4 'perceived utility of advice' was operationalized as the radio button that participants clicked in the on-line survey (each button was allocated a value). Participants were also tasked with providing a confidence interval, within which they were 95% certain that the true value of the target quantity could be found (Chapters 4-8), both prior to the receipt of advice, and post-advice.

Data was recorded in Chapter 4 by 'MouseLabWeb' software, and stored in a MySQL database. In addition to the measures outlined in the preceding paragraph, process measures - such as where the on-screen cursor came to rest, how long the cursor remained stationary, what on-screen items were 'clicked' upon, and the time from the start of the experiment (in milliseconds) to any of these events – were recorded. Data was captured by an electronic interface in Chapter 6, and stored in a MySQL database, where participants completed the on-line version of the questionnaire.

Both parametric and non-parametric statistical tests of significance were used in analyzing the data from the experimental manipulations. Both techniques were necessary as much of the raw data did not meet the assumption that the data was normally distributed about the mean. Where appropriate, such data was Log, or $Log_{(10)}$ transformed prior to any test of significance. Statistical tests from the General Linear Model (GLM) were applied to the data (ANOVA, Multiple Linear regression, correlations, *t* tests). In addition, Mann-Whitney U, and Kruskal-Wallis non-parametric tests of significance were utilized where

appropriate. In Chapter 9, it was necessary to compare the strength of the relationship between pre-advice rank ordering of target items, and the true rank order of target items, with the strength of the relationship between post-advice rank ordering of target items, and the true rank order of target items - this was facilitated by the use of Steiger's Z_1 statistic. Chapter 4 reports the findings of an initial exploratory investigation.

Chapter 4: Do people prefer purely numerical advice or numerical advice qualified by reasons?

4.1 Introduction

It is contended that conformity to advice is an important determinant of judgment change, and this first study sets out to explore the influence of reasons-based advice, and solely numeric advice, upon participants' deliberations, and subsequent estimation, of the dates of three historical events. The motivation for this enquiry is that up till now, advice in JAS research has been operationalized in solely numeric terms (as set out in Chapter 2). Hence, it is not clear, in the absence of adequate environmental cues such as reasons, if people are able to distinguish between accurate, and inaccurate numerical advice. In these circumstances, where people revise their judgment post-advice, social conformity to advice can be expected. Further, intrinsic information of the quality of advice is made available to participants in the form of supporting reasons for the numerical component of advice in this study. This implies that should participants be sensitive to the quality of advice available to them, then this advice could be incorporated into their deliberations; reflected in any reported evaluations of the utility of advice, and measured in terms of judgment change. For the purposes of the study, reasons-based advice is operationalized to be broadly consistent with the work of Toulmin (1958) and more specifically, Brockriede and Ehninger (1960). These authors indicate that only a limited number of arguments are available to people advocating specific claims – arguments of parallel case, analogy, motivation, and authority. Each of these argument forms is discussed in the following section.

4.2 Argumentative forms

Arguments of parallel case rely upon a warrant in which it is asserted that an instance reported

in the data is essentially similar to a second instance in the same category. A typical example of an argument of parallel case might be the media debates over US, and allied, military strategy in the Persian Gulf War of 1991, extrapolated from the negative parallel case of US military strategy in Vietnam. Here, US military strategy in the Persian Gulf is deemed to be from the same category of strategic cases as the Vietnam conflict. In a similar vein, analogous arguments rely upon the idea that, if two things have certain attributes in common, then they are likely to have one, or more, additional traits in common. Hence, if a person likes the music of Oscar Peterson, Miles Davis, and Ella FitzGerald, the same individual will probably also like the music of Amy Winehouse, Gato Balbieri, and Gil Scott Heron. In this instance, the former group of musicians pioneered the musical form of jazz, whilst the latter perform music that is overtly influenced by jazz, with additional elements (i.e. traits) of pop and soul music. Each of the two groups of musicians have traits in common, yet it cannot be said that all of the musicians perform music that can be categorized as jazz (hence the comparison is not a parallel case, but an analogy). Contrastingly, motivational arguments rely upon a warrant that motivates acceptance of the claim by appealing to inner drives or values, emotions or aspirations. So, the statement 'rich people are only wealthy because they are inherently greedy' is an argument from motivation, where the advocate links the internal motivations of rich people (greed), to the outcome of wealth. Finally, arguments of authority are based upon a statement made in the data, made credible by identifying the source of the statement. Here, the credibility of a proposition is enhanced (or diminished), by knowledge of the source that substantiates the proposition (i.e. *ipse dixit* – he, himself, said it). An example might be that, the theory of evolution explains the origins of humankind, according to Charles Darwin. The implication here is that Charles Darwin is an expert on the origins of humankind, and as such, his views in this area should carry more weight than people less expert.

Each of the argument forms discussed in the preceding section *can* be persuasive in a given set of circumstances, but a comprehensive taxonomy of such arguments is unlikely, given the difficulties of delineating argumentative form from content. This is because it is entirely possible to have a perfectly reasonable *form* of argument (e.g. a syllogism), but variations in the quality of the content of the argument. The quality of the content of an argument can be gauged against its truthfulness. An example might be that 'I have never seen a black swan, therefore all swans must be white' is a reasonable syllogistic argument, yet the argument fails as soon as a black swan appears. Hence, the persuasiveness, or otherwise, of a particular argument form is difficult to disentangle from its contents. Nevertheless, there is some evidence to suggest that arguments of analogy are more persuasive than messages without any comparative content (McCroskey and Combs, 1969). Here, McCroskey and Combs presented their participants with three analogy messages critical of US foreign aid policy towards Brazil. A 'literal' analogy compared US foreign aid policy to Brazil, with the US urban renewal programme; a 'figurative' analogy message compared US foreign policy towards Brazil to a snowstorm - whilst a 'no-analogy' condition (control) contained no comparative messages (instead the message contained descriptions of the failures of the US foreign aid program in Brazil). Further, source credibility was manipulated by attributing the statements (literal analogy, figurative analogy, no analogy), to either a highly credible source (Charles. L. Wilson, former US ambassador to Brazil), or a low credibility source (Lin Tai, ambassador to Brazil from Red China). Prior to the experimental manipulations, participants completed an attitude pre-test. This test tasked participants with evaluating statements such as, "The U.S. has a well administered foreign aid program for Brazil". Participants were subsequently asked to evaluate this statement on scales of "Right-Wrong", "False-True", "Yes-No", "Incorrect-Correct", "I agree-I disagree". After completing the attitude measure, participants received the experimental message manipulations, and completed an attitude post-test. Measures of attitude change scores, and source credibility ratings, indicated that both the 'literal' analogy message condition and the 'figurative' analogy message condition produced greater attitude change than messages without comparative content (no-analogy). Similarly, Smith (1972) established that

desirability evidence (motivational arguments in our terminology) was more persuasive than controls and truth evidence (evidence that establishes that something is actual). Further, Reinard (1988) indicates that source credibility can enhance or diminish the persuasiveness of an argument; whilst Stanchi (2006), asserts that arguments of parallel case, or precedent, are persuasive in a legal context.

For the purposes of this study, it is not helpful to attempt to hypothesize, and predict, *a priori,* which of the four arguments types discussed in the previous section (analogy, parallel case, authority, and motivation), is *more* persuasive, rather it is preferable to formulate research questions exploring whether an advice package that includes numerical advice in addition to one of these argument types, will be regarded as more helpful, or useful by people, than solely numerical advice, when people are faced with estimation problems. Exploration of this issue may reveal that participants value solely numerical advice (recall this is the common operationalization of advice in JAS research) *less* so than reasons-based advice. That JAS research has so far not systematically considered this issue underscores the limitations of the paradigm. Hence, the main thrust of the initial empirical study reported in Chapter 4 is to explore the preferences that people hold towards different types of advice. This approach allows consideration of several issues that would either be excluded, or not add to the clarity of a hypothesis driven experimental design. The next section discusses the exploratory issues addressed in Chapter 4.

One issue worthy of further investigation concerns whether people are sensitive to the asymmetries in the quality of available advice. Should participants be able to adequately distinguish between potentially beneficial advice, and poor advice, then reported perceived ratings of the utility of advice should reflect the sensitivities participants report regarding the asymmetries of the quality of available advice. Alternatively, people may *not* be sensitive to

the asymmetries of the quality of available advice, and hence regard all advice as approximately equal in terms of its potential to enhance estimation performance. Under this latter scenario, people may conform to advice irrespective of its quality when they choose to revise their judgments post-advice. Subsequently, individuals' expressed preferences for particular advice packages (i.e. a preference for solely numerical advice over reasons, or a preference for analogous advice, over advice based upon some authority etc.) can be considered. One issue that this study explores, is whether people exhibit a preference for one particular type of advice, when responding to an estimation problem, and subsequently consistently prefer the same advice type, when faced with further estimation problems. The issue of what advice 'package' individuals may prefer, also potentially speaks to the question of whether people are able to discriminate between potentially beneficial advice (advice that if followed, enhances estimation performance), and poor advice (advice that potentially degrades estimation performance). These exploratory issues are explored in the following section.

4.3 'Poor' vs 'Good' advice

In part, it is possible that an individual's preference for reasons-based advice over solely numeric advice, might be because the former simply contains more information; and it is parsimonious to suppose that people might find the former more 'helpful', than the latter, irrespective of the numerical accuracy of the advice (this is not to say that the perceived utility of reasons, however elaborate those reasons might be, correlates with advice that is beneficial). This implies that either reasons, or solely numerical advice, will be beneficial for participants whose unaided estimates are more inaccurate than the available advice. However, where participants' initial unaided estimates are more accurate than the advice, any post-advice revision of judgment that results in an estimate that is *less* accurate, than a participant's initial unaided estimate, maybe be attributable to the influence of poor advice (i.e. revising one's judgment in the 'direction' of poor advice is detrimental to estimation performance, in this

scenario). If people are successfully able to distinguish between potentially beneficial advice, and poor advice, then it might be expected that beneficial advice would be incorporated into an individual's deliberations (provided individuals are motivated to estimate accurately), resulting in enhanced estimation performance. However, should individuals be unable to successfully distinguish between beneficial and poor advice, then people are likely to be attracted to reasons-based advice - as it contains more information - in preference to solely numerical advice. However, in these circumstances, if people do revise their judgment, an individual's estimation performance is unlikely to be any more accurate than the advice. One way to observe if people do not easily discriminate between potentially beneficial and poor advice; is to simultaneously present people with several pieces of advice that varies in accuracy; so that if people are able to successfully distinguish between beneficial and poor advice, they might at least reject the poor quality advice. Investigating this question is appropriate, as where advice is solely numeric it is possible to discover if improvements in estimation performance occur, because participants are merely conforming to advice. In sum, one of the research question that is explored here is to discover whether people can discriminate between beneficial advice, and poor advice, in circumstances of incomplete domain knowledge.

4.4 Confidence

The role that judges' confidence may play in judgment revision is an issue discussed in Chapter 2. This discussion serves to motivate research questions that are formulated here. First, it is of interest to discover if people who are highly confident pre-advice, are less disposed to change their judgment, post-advice, than people who are less confident, pre-advice. People who are less confident pre-advice may be disposed to change judgment post-advice because (i) the judge appreciates the superior logic of the advice given to him/her by the advisor where the advice is reasons-based or (ii) the judge is able to bring so little 'world

knowledge' to bear upon the issue of interest, that any advice (either reasons-based or solely numeric) is potentially useful. Second, it is of interest to discover if people report greater degrees of confidence post-advice, compared to the degrees of confidence they report pre-advice (i.e. whether people become more overconfident). Third, whether people in receipt of reasons advice will become more confident in their intuitive estimates post-advice, than people in receipt of solely numeric advice. In order to address the research questions formulated in the preceding discussion, a simple 3-item historical quiz was devised.

4.5 Method

Participants

The sample was drawn from staff at Durham University (UK). A total of 74 participants attempted the on-line questionnaire – three participants were omitted from the analysis due to incomplete data entry. One participant only attempted Question 1, entered no data for a pre-advice estimate and entered '1920' for a pre-advice upper confidence bound, the post-advice lower confidence and upper confidence bound, without entering a post-advice estimate. This participant was excluded from further analysis. A second participant only partially completed Question 1 and omitted to enter values for a pre-advice estimate, lower and upper confidence bounds, and a post advice estimate, but did enter values for post-advice lower and upper confidence bounds. Hence, this participant was excluded from further analysis. A third participant attempted to complete Questions 1 and 2. For Question 1 this participant provided no initial estimate or confidence bounds, and where data was provided for a post-advice estimate the lower confidence bound was numerical greater than the post-advice estimate. For Question 2, the same participant provided data for a pre-advice estimate, lower and upper confidence bounds, but where data was provided for a post advice lower and upper confidence

both entries were identical and outside the permitted range of values (i.e. '2050' has not yet occurred) - hence this participant was excluded from any further analysis. Hence, a total of 71 participants were tasked with completing three estimation problems resulting in 213 responses for analysis (regrettably not all respondents provided all the requested data)..

Stimuli and Materials

A 'historical quiz' was devised consisting of three questions –

Question 1 – In what year was the first Japanese railway opened?

Question 2 – In what year did Britain receive a mandate to govern Palestine?

Question 3 - In what year was the first Camp David summit between Egypt and Israel hosted by the United States?

These questions were piloted on 10 individuals (subsequently not included in our experiment), to check that people could *not* easily recall the true answer to each of these questions, whilst still reporting some familiarity with the historical events. Participants responded to the historical quiz on-line, by mouse-clicking a hyperlink in an email that invited Durham Business School staff to participate in the experiment. The on-line interface was adapted from the 'MouseLabWeb' process tracing software (Willemsen and Johnson, 2004), and data was electronically recorded in a MySQL database. See Appendix 1 for details of the on-line web pages.

Procedure

Participants responded to the historical quiz on-line, by mouse-clicking a hyperlink in an email that invited Durham Business School staff to participate in the experiment. The presentation order of the historical quiz questions was randomized (a 'php' script that randomized question presentation was activated when participants clicked on the hyperlink in the invitation email). Subsequently, participants were tasked with estimating the date at which each of the three historical events occurred (see Appendix 1 for details of the on-line web pages). Participants were also tasked with providing a range of values that bracketed their estimate, within which they were 95% certain that the true answer (whatever that might be) could be found. Participants estimated a number that provided a lower estimate bound, and also a number that provided an upper estimate bound, as the parameters of their range of values. The range of values (that contained both a participant's estimate and the true answer) was operationalized as a dependent measure of our participant's confidence, (where greater confidence was indicated by a smaller range of values that bracketed an individual's estimate and the true answer).

On completion of this first phase, participants clicked a 'Next page' icon and proceeded to the following web page of the survey. Here, participants were presented with solely numeric advice, and reasons advice. There were four advice boxes containing solely numerical pieces of advice, and four boxes containing pieces of advice that consisted of a number with an additional reason. Participants could view the advice contained in the advice boxes by simply moving the on-screen cursor over the box and 'clicking' the computer mouse on the box. When participants moved the on-screen cursor out of a particular advice box, the advice became concealed once more. Participants were free to click on all, or none of the advice boxes, and could re-visit any, or all, of the advice boxes at any time. Where an advice box contained a number and a reason, the type of reason was one of the following - a parallel case,

an analogy, evidence by an authority, or justified by motivation. Explanation of, and examples of, each reason type (see Appendix I, Screen 4 for details) were presented to participants as they completed the first historical quiz question; when completing the remaining two historical quiz questions, participants could access this information by clicking, and holding a 'Show help' icon. The accuracy of the advice was varied so that some advice was more accurate than others (see Table 1 below). With the exception of Question 2, the accuracy of the advice was balanced in terms of its sign relative to the true value. Hence, Question 1 contained four underestimates of the true value, and four overestimates of the true value. Question 3 contained six underestimates of the true value, and two overestimates of the true value (three underestimates were in the solely numeric advice condition, and three in the reasons advice condition), whilst Question 2 contained four underestimates of the true value, three overestimates of the true value, and the true value. Further, the position of the advice boxes was varied between participants, to counteract potential presentation order effects.

Table 4.1. Advice accuracy variations – numbers in parentheses indicate the variation of the advice around the true value

Question	True answer	Advice box	Solely numerical advice	Advice box (including reason type)	Reasons-based advice
1	1872	A	1861 (- 11)	E - analogy	1890 (+18)
		B	1849 (- 23)	F – parallel case	1869 (- 3)
		C	1883 (+ 11)	G - authority	1835 (- 37)
		D	1895 (+ 23)	H - motivation	1902 (+30)
2	1920	A	1899 (- 21)	E - analogy	1946 (+ 26)
		B	1955 (+ 35)	F – parallel case	1918 (- 2)
		C	1911 (- 9)	G - authority	1902 (- 18)
		D	1933 (+ 13)	H - motivation	1920 (0)
3	1978	A	1955 (- 23)	E - analogy	1960 (-18)
		B	1979 (+1)	F – parallel case	1971 (-7)
		C	1967 (-11)	G - authority	1950 (-28)
		D	1945 (-33)	H - motivation	1982 (+4)

Once participants had considered the available advice, they were asked to click one of an array of nine 'radio-type' buttons to indicate which piece of advice they had found most helpful in forming an estimate to the historical quiz question ('0' was included as an option where 'no advice was useful'). Next, participants were asked to indicate if they wished to revise their estimate. If participants clicked the radio button 'Yes' then they were provided with a text box

in which to enter their revised estimate (post-advice). A mouse-click on a 'Next page' icon then allowed participants to proceed to the following web page. Here, a text box was provided, so that participants were able to enter an explanation as to why they either revised their estimate, or did not revise their unaided estimate. Another 'Next page' icon, when clicked allowed participants to access the penultimate web page, where they were asked to provide numerical estimates of the upper, and lower bounds, within which they were 95% certain contained the true answer to the historical quiz question. These upper and lower bound estimates also bracketed our participant's revised (or un-revised) estimate. Once completed, a 'thank you' message was displayed on the participant's computer monitor, allowing the browser to be closed. As each page of the web site was loaded onto a participant's computer monitor screen, the participant's data entries from previous completed web pages were simultaneously displayed for the purposes of clarity (i.e. unaided estimate, confidence range, revised estimate, and revised confidence range). The 'MouseLabWeb' interface is designed to capture a wealth of data, which is temporarily stored in an HTML form, and saved to the MySQL database when a participant clicks on 'submit' (see Fig 4.1 below).

Fig 4.1. ‘MouseLabWeb’ data capture

Form field	Description
expname	Name of MouseLabWeb page
subject	Name of subject
condnum	Number of the counterbalancing condition
IP	Address of participant’s computer
choice	Choice made by the participant (should choice buttons be added to the MouseLabWeb table)
procdata	Process data – a serial list of events
Event	Type of event (mouseover, mouseout, click)
Name	Name of the event (Name of MouseLabWeb cell, or form element that is captured)
Value	Value of the event
Time	Time in milliseconds measured from the moment the page was started
OnLoad	Time the page was fully loaded and displayed
Order	Exact order in which columns and rows were presented
nextURL	The name of the next page that the browser will go to after submission of the current page.
addvar	Additional form elements
Adddata	Values of additional form elements

The inputs participants made (i.e. unaided estimate, confidence range, revised estimate, comments, and revised confidence range), were programmed into MouseLabWeb as 'addvar' (additional form elements), and their values recorded in the 'adddata' form field, prior to submission to the MySQL database.

4.6 Results

The following results section is primarily focused upon the main argument contended in this thesis - that conformity to advice is an important determinant of judgment change. This implies that people are constrained in their abilities to distinguish between potentially beneficial advice, and poor advice. As a measure of perceived utility of advice was taken, it was possible to determine if people were sensitive to the variations in the quality of advice made available to them. This is because it might be expected that if people were sensitive to the quality of advice, then ratings of the perceived utility of advice should reflect this sensitivity. If this is found to be so, then the argument that people are able to adequately discriminate between potentially beneficial advice, and poor advice, would be supported. However, if (as is contended here) people are constrained in their abilities to distinguish between potentially beneficial advice and poor advice, then reported perceived utility of advice ratings will not reflect the asymmetries in the quality of the available advice. Further evidence for the notion that people do not successfully discriminate between potentially beneficial advice and poor advice will be examined in terms of the influence of advice upon judgment change. Here, it is possible to determine the absolute amount of judgment change, and determine if there are significant differences between (i) people who report that they found no advice useful and (ii) people who report that they found solely numeric advice useful (iii) people who report that the found reasons-based advice useful. Further, the data of only participants who found reasons-based advice useful is analyzed to determine if there are

significant differences between argument types (parallel case, analogy, authority and motivation) in terms of the absolute amount of judgment change. These analyses will determine if participants were sensitive to the asymmetries in the quality of reasons-based advice. Subsequently, analyses on the direction of judgment change are performed to determine whether people revise their judgment post-advice, in the direction of that advice, where they report that they value it. Moreover, analyses are performed on the data of participants who revised their judgments post-advice to ascertain if such judgment was beneficial in terms of changes in estimation accuracy, and ultimate estimation accuracy. Ultimately, it is necessary to determine if conformity to advice is an adequate explanation for the observed data, before considering further exploratory analyses.

Further exploratory analysis are considered in this study, and can be broadly characterized by analyses upon the process data generated by the 'MouseLabWeb' software, and analyses concerning participants' confidence both prior to receipt of advice, and post-advice. First, process measures are considered. One issue that is investigated is whether participants were consistent in their preferences for advice. As participants clicked upon a radio button indicating which type of advice they found most helpful when responding to each of the three estimation questions, it was possible to calculate the number of instances in which participants chose the same type of advice (i.e. as indicated by clicking the radio button that such advice was most useful), on subsequent estimation questions, as they did when responding to the first estimation question. This analysis determines whether participants were consistent, or inconsistent, in their preference for particular types of advice. Consistency in participants' preferences for advice might indicate that participant's information search was not random. A second analysis, based upon the process measures generated by 'MouseLabWeb', concerned whether participants held preconceptions of the utility of the available advice. The

'MouseLabWeb' software recorded time in milliseconds (calculated as the absolute difference from the start of the questionnaire until a particular event). Given that both question order, and advice box presentation was randomized across participants, it is possible to determine the order in which people first accessed advice. Should participants hold preconceptions of the utility of any particular advice type, this should be reflected, on average, in the speed in which advice is first accessed. A third analysis based upon the results of the process data generated by 'MouseLabWeb' concerned what advice participants examined prior to revising (or not revising their judgment). The 'MouseLabWeb' software recorded which advice boxes participants 'clicked' on as they completed the questionnaire, and subsequently it was possible to determine the total number of advice 'acquisitions' and 're-acquisitions' participants made, prior to revising (or not revising), their ultimate estimate. It might be inferred from this analysis, that greater cognitive processing of advice is indicated by the number of times particular advice types are acquired, and re-acquired by participants prior to any judgment revision.

Finally, measures of confidence, and changes in confidence, are considered. Confidence was operationalized here as the interval between an upper and lower bound within which participants were 95% certain that the true answer to the three estimation problems fell. The degree of confidence held by participants is inferred by the 'width' of the confidence interval (i.e. the greater the absolute difference between the lower and upper bound of the confidence interval, the less confidence participants are inferred to hold). This operational definition of confidence allows exploration of the issue of whether attending to advice increases or decreases confidence in participants' estimation abilities. Second, it is possible to determine whether participants low in confidence, are more or less likely to revise their judgment post-

advice. Thirdly, whether attending to advice results in changes in the degree of confidence held by participants is an issue addressed here.

The first issue that needs to be addressed in analyzing the data generated by the MouseLabWeb software concerns the response rates of participants in the study. Seventy-one participants provided data, in response to three estimation problems, optimally this provides 213 responses (71 x 3). The observed number of questions answered was 163. Hence, participants responded to 76.5% of questions, on average. This sub-optimal rate of response can be attributed to the failure of the author, and the limitations of the MouseLabWeb software, to ensure that all data entry fields were subject to 'validation' before a participant could complete the online questionnaire (validation is a line of software programming code that ensures that data entry is present and within certain bounds, before a participants can complete the online questionnaire). The author was able to successfully validate some, but not all, of the data entry fields in the online questionnaire. This limitation had an unexpectedly large influence upon the data that was collected from participants. Recall that each participant could potentially provide seven pieces of data for each estimation question (initial unaided estimate, pre-advice lower confidence bound, pre-advice upper confidence bound, perceived utility of advice rating, post-advice estimate, post-advice lower confidence bound, and post-advice upper confidence bound). Hence, optimally, each participant should have entered twenty-one pieces of data (3 questions x 7 data entries) to the online interface. Hence, 71 participants should have entered (21 x 71) 1491 pieces of data. The observed number of data entries was 1124, giving an overall response rate of 75.39% (see Table 4.2).

Table 4.2 Number of responses

N = 71	Unaided estimate	Pre-advice Lower bound	Pre-advice Upper bound	Advice rating	Post-advice estimate	Post-advice Lower bound	Post-advice Upper bound	Total out of 497
Question 1 (Date of Japanese Railway)	60 (84.5%)	60 (84.5%)	60 (84.5%)	57 (80.3%)	60 (84.5%)	59 (83.1%)	58 (81.7%)	414 (83.3%)
Question 2 (Date of GB mandate over Palestine)	55 (77.5%)	55 (77.5%)	55 (77.5%)	51 (72%)	55 (77.5%)	53 (74.6%)	53 (74.6%)	377 (75.9%)
Q3 (Date of the first Arab-Israeli Camp David)	48 (67.6%)	48 (67.6%)	48 (67.6%)	45 (63.4%)	48 (67.6%)	48 (67.6%)	48 (67.6%)	333 (67%)
Totals (out of 213)	163 (76.5%)	163 (76.5%)	163 (76.5%)	153 (71.2%)	163 (76.5%)	160 (75.1%)	159 (74.6%)	1124/ 1491 = 75.39%

Given that the dataset collected from participants in this study contains many missing responses, the figures and tables that follow report *N* responses, rather than *N* participants.

Initially, the distribution of the data (for 163 valid responses) was examined, and it was found that participants' estimates were not normally distributed. Log transformation of these data was successful in satisfying the assumptions of normally distributed data. Hence, parametric statistical tests were used where appropriate in the analysis that follows.

Can participants discriminate between good and poor advice?

The question of whether people are adequately able to distinguish between potentially beneficial advice, and poor advice, is the first issue addressed in this analysis. Initially, it is important to clarify that reasons-based advice was more numerically accurate, than solely numeric advice. The absolute difference between the numerical value of each piece of advice, and the true value, was calculated, and subsequently aggregated over the three estimation questions, and it is clear that reasons-based advice is less error prone (35 years, on average), than solely numerical advice (48 years, on average) (see Table 4.1). Superficially, this may appear as an unfortunate bias that privileges participants who attend to reasons-based advice in terms of estimation accuracy. In retrospect, stricter control over the variation of the quality of numerical advice around the true (but unknown to participants) value could (and perhaps should) have been applied for solely numeric advice types, and reasons-based advice types. Despite the exploratory nature of this study, it is important to acknowledge this methodological shortcoming, as aggregating the average accuracy of advice types over the three estimation questions may suggest that participants were subject to a demand characteristic of the situation. What this means is that participants may have initially been attracted to the 'richness' of reasons-based advice in preference to solely numeric advice from the outset, and subsequently benefited from the greater aggregate accuracy of reasons based advice across the three estimation questions. Potentially, this is a serious confound of the study. However, this line of reasoning was not supported by the data. Aggregating the mean post-advice accuracy scores of participants across questions (see Table 4.10), and performing a one-way ANOVA, revealed no significant differences between conditions ($F_{(2, 114)} = 1.28$, $p < 0.28$). An identical result was found when the aggregated mean post-advice scores from only participants who revised their judgments post-advice (see Table 4.11), were entered into a one-way ANOVA - no significant differences were found between conditions ($F_{(2, 50)} = 1.10$, $p < 0.34$). These findings suggest that despite the author's failure to fully acknowledge the importance of control over the

variation of the quality of advice between advice types, participants were insensitive to the potential benefits of reasons-based advice in the aggregate. However, the purpose of the manipulation was to assess the preferences participants exhibited towards advice in an exploratory fashion. Recall that the accuracy of the numerical component of advice presented to participants was not precisely counterbalanced in terms of the number of years (greater or less), than the true value for each estimation question. Rather, two of the estimation questions were counterbalanced in terms of the number of pieces of advice that, if followed, would lead participants to either overestimate, or underestimate the true value. One of the estimation questions contained the true (but unknown) answer. This issue will be discussed further when the post-advice accuracy of participants' estimates are considered.

In addition to the preceding findings, the idea that participants attended reasons advice because they recognized that such advice was more accurate, than solely numeric advice, is not completely satisfactory. This is because reasons-based advice may superficially appear more useful to participants due to the greater amount of information such advice contains, in comparison to solely numeric advice. Further, it is possible that participants may have realised that reasons-based advice was constructed by the author, rather than being real advice from appropriate sources. This may have been particularly true of advice evidenced by an authority. Should participants evaluate advice substantiated by an authority as being bogus, then such advice is likely to be perceived as of poor quality and unlikely to be persuasive. Participants did not evaluate advice evidenced by authority as being from persons other, than the named authority. This is evidenced by the comments participants entered into a text box within the online survey. Of the thirty-two comments that specifically mention advice evidenced by authority, only two were particularly sceptical of the credibility of the source of the advice (e.g. 'see my previous answer - the flawed authority' and 'none of the advice was persuasive - the

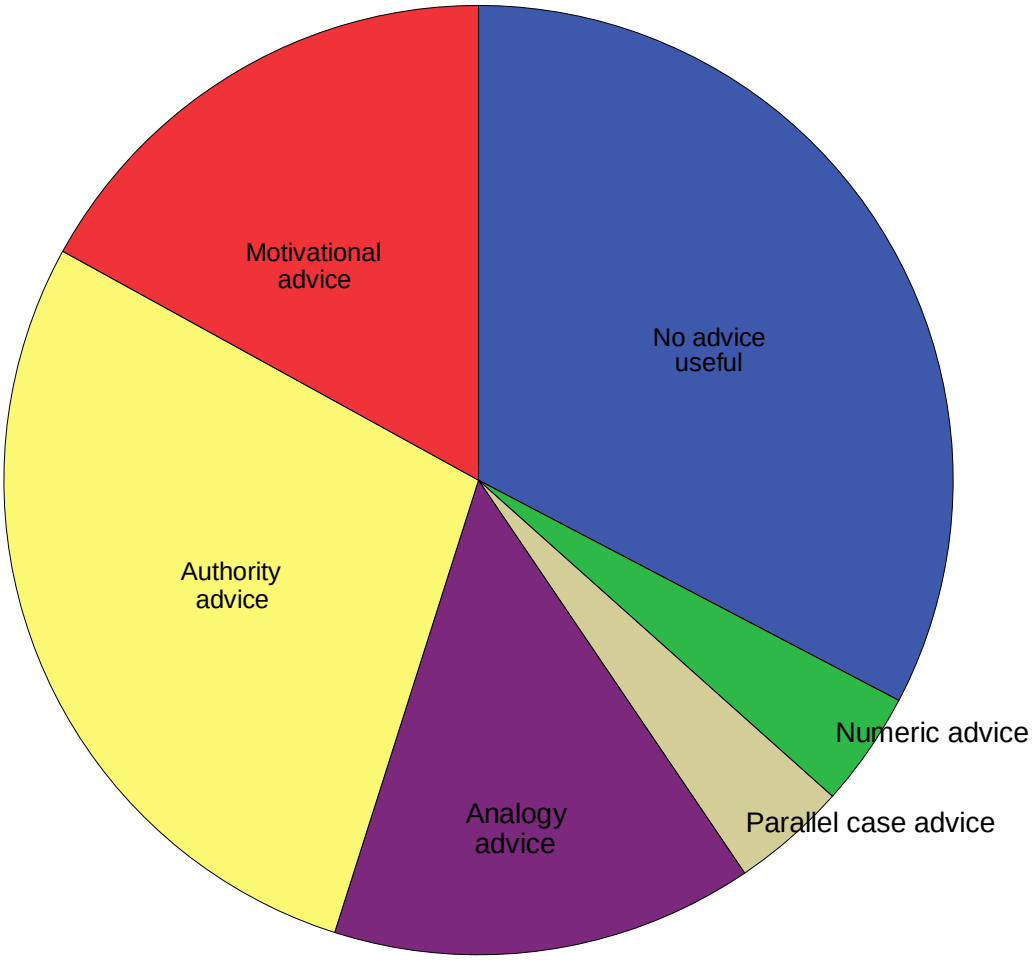
authorities behind the explanation are unknown to me'). Instead the remaining thirty comments did not question that the authorities behind the advice were real (e.g. 'he's an expert and I have so little knowledge of the subject that his advice seems quickest and best to accept').

If participants are able to discriminate between beneficial, and poor advice, in circumstances of incomplete domain knowledge, then it follows that participants should be able to differentiate between the utility of solely numeric advice, and reasons-based advice. In addition, participants in receipt of solely reasons-based advice should be able to distinguish between beneficial, and poor advice. It was possible to determine the perceived utility of advice reported by participants, by aggregating individual rating scores. Inspection of Fig 4.2 and Table 4.3 shows that, across the three estimation problems presented to participants, advice that was most accurate, on average (i.e. parallel case advice) was not perceived to be the most useful by participants (in fact it was perceived as the least useful of the four advice types, on average). In fact, parallel case advice was only the most accurate advice on offer to participants in Question 1 (estimate the date of the first Japanese railway). Providing one assumes that participants were motivated to estimate accurately, then these reported perceptions of the utility of the available advice, do not reflect the *actual* accuracy of advice.

Seventy-one participants provided data for analysis in this study. However, only seventy participants clicked on a radio button in the online questionnaire indicating what advice (if any) was most useful to them. In addition, nine participants (of the seventy that provided perceived utility of advice ratings) did not provide a value for the perceived utility of advice on one of the estimation questions that were attempted. Hence, 153 responses were available for analysis. One issue might be that participants were only permitted to indicate one source of

advice as being 'most useful' to them which prevented a response when more than one piece of advice might have been 'most useful'. However, participants in these particular circumstances did have the option of indicating that no one piece of advice was 'most useful' by clicking on the 'no advice useful' option. Fig 4.2 and Table 4.3 show the advice participants found most helpful in the aggregate, and at the level of each estimation question.

Fig 4.2 Aggregate (all estimation questions) perceived utility of advice ratings (responses *N* =153)



Inspection of Table 4.2 shows that, in the aggregate, 32.68% of participants did not regard any of the advice as useful to them when formulating an estimate to the three estimation problems

presented. Where participants did find advice useful, 3.92% regarded solely numeric advice as useful, whilst 63.24% regarded reasons-based advice as useful, on average (14.23% perceived analogous advice as most useful, 3.92% reported parallel case advice most useful, 28.10% reported advice substantiated by authority as most useful, and 16.99% reported advice supported by motivational arguments as most useful).

Table 4.3. Perceived usefulness of advice (frequencies)

N = 70	Number of valid responses	No advice useful	Numeric advice useful	Parallel case advice useful	Analogy advice useful	Authority advice useful	Motivation advice useful
Q1	57	12 (21.05%)	2 (3.51%)	1 (1.75%)	9 (15.79%)	22 (38.60%)	11 (19.30%)
Q2	51	23 (44.23%)	2 (3.85%)	3 (5.88%)	10 (19.23%)	9 (17.31%)	4 (7.7%)
Q3	45	15 (33.33%)	2 (4.4%)	2 (4.4%)	3 (6.7%)	12 (26.7%)	11 (24.4%)
Total	153 (100%)	50 (32.68%)	6 (3.92%)	6 (3.92%)	22 (14.23%)	43 (28.10%)	26 (16.99%)

It is noteworthy that of the participants who prefer reasons-based advice, the most popular advice type is numerical advice evidenced by some authority, in two out of three estimation questions. In this study, numerical advice evidenced by some authority, was operationalized as advice consisting of an estimate of the date of the historical event in question, and its source (i.e. a fictitious historian – e.g. Dr. Smith). These findings strongly suggest that conceptualizing advice in solely numeric terms poorly reflects the preference participants here have reported for reasons-based advice, and in particular, reasons-based advice evidenced by authority.

If participants were sensitive to the accuracy of reasons-based advice, then it might be expected that the advice reported as useful reflected such sensitivity. However, there is little evidence that participants were sensitive to the asymmetries of advice quality in either the aggregate level of analysis, or at the level of each estimation question (see Table 4.4). Table 4.4 shows that participant's perceptions of the utility of the available reasons-based advice, and its *actual* accuracy, were dissimilar. This may imply that our participants' belief that they were able to distinguish between beneficial advice, and poor advice, was not well calibrated.

Table 4.4 Reasons advice: Perceived utility of advice scores, and observed accuracy of advice (rank order).

Question	Accuracy: rank order of advice type (1 = most accurate)	Perceived utility: rank order of perceived utility (1 = most useful)
1	1. Parallel case	1. Authority
	2. Analogy	2. Motivation
	3. Motivation	3. Analogy
	4. Authority	4. Parallel case
2	1. Motivation	1. Analogy
	2. Parallel case	2. Authority
	3. Authority	3. Motivation
	4. Analogy	4. Parallel case
3	1. Motivation	1. Authority
	2. Parallel case	2. Motivation
	3. Analogy	3. Analogy
	4. Authority	4. Parallel case
Aggregate of Questions 1 - 3	1. Parallel case	1. Authority
	2. Analogy	2. Motivation
	3. Authority	3. Analogy
	4. Motivation	4. Parallel case

Table 4.4 shows that the perceived utility of the advice that participants indicated was most helpful to them bore little resemblance to the actual rank order accuracy of the observed advice types, both at the level of individual questions and in the aggregate.

It has so far been established that, in circumstances of incomplete domain knowledge, when participant's perceived advice as useful, they overwhelmingly preferred reasons, to solely numeric advice. Further, this preference was obtained irrespective of the accuracy of the perceived useful advice. In sum, participants were constrained in their abilities to distinguish between beneficial and poor advice. The finding that people are constrained in their abilities to distinguish between beneficial and poor advice is perhaps not completely surprising, considering the 'almanac' nature of the three estimation questions, the 'one-shot' nature of each estimation question and that feedback to participants was unavailable. This meant that participants were prevented from learning how accurate their unaided estimates were prior to examining available advice, and given the 'one-shot' nature of the estimation scenario also prevented from learning which advisor(s) were more accurate than others. However, these considerations may not be the most influential factors in determining the abilities of people to differentiate between potentially beneficial advice and poor advice.

For the findings discussed previously to have any general utility it is necessary to ascertain whether people are, on average, constrained in their abilities to differentiate between potentially beneficial, and poor advice beyond the narrow confines of the manipulations reported in this study. A striking example of how people with expertise in a particular area are simultaneously constrained in their abilities to differentiate between potentially beneficial advice and poor advice is provided by Ayres (2007). Ayres cites the example of economist

Orley Ashenfleter who claimed to be able to predict the quality, and ultimately the price, of Bordeaux wine on the basis of a simple statistical model that acknowledges the influence of winter, harvest rainfall, and growing season temperatures. Ashenfleter's methodology was derided by many experts in the wine industry including the wine critic Robert Parker (Ayres, 2007). However, it was Ashenfleter's predictions that proved to be more accurate about the 1986, 1989 and 1990 vintage Bordeaux wines, than Parker's. This example shows that even people with domain knowledge expertise may not always recognise potentially beneficial advice, and be able to separate it out from poor advice.

In a more general sense, it is entirely possible that people may not be able to discriminate between potentially beneficial and poor advice, on the basis of individual preferences – irrespective of the quality or veracity of advice, people may choose not to accept it. Such a position acknowledges individual differences perhaps, but makes the idea of advice giving no less complex. Many other examples of both the limits of human cognition, and decision-making guided by individual preferences displayed by people when trying to discriminate between potentially beneficial and poor advice can be found in the extant literature of the management of organizations. Unsurprisngly, this literature is replete with many types of advice (Checkland, 1981; Warfield, 1990). Indeed, one area of the organizational literature that seeks to improve advice to managers is 'action research' Denzin, and Lincoln, 1994; Reason, 1988). The results of this program of research has not conclusively shown either that people are able to adequately discriminate between potentially beneficial and poor advice in any global cognitive sense, or that interventions designed to improve the quality of advice available to decision-makers has any generic utility.

Does attending to advice influence judgment change?

Given participants' preferences for reasons, the degree of judgment change across advice types (i.e. no advice perceived as useful, solely numeric advice perceived as useful, and reasons perceived as useful) was examined. As only 70 participants had entered a value for perceived utility of advice, 153 responses were available for analysis. Moreover, analyzing the data at the level of each estimation question is made problematic by the fact that only two responses in each estimation question favoured solely numeric advice. Judgment change was measured as the absolute numerical difference between participants' unaided estimates, and their estimates post-advice (see Table 4.5 for means and standard deviations). As these scores violated the assumption of normality, $Log_{(10)}$ transformed values were used in tests of significance.

Table 4.5. Degree of judgment change scores (years)

N = 70	Number of valid responses (*N* = 153)	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed)	*sd*
Solely numeric advice perceived as useful	6	12.83	14.96	0.69	0.76
Reasons-based perceived as useful	97	17.52	30.91	0.67	0.75
No advice perceived as useful	50	4.42	12.79	0.19	0.51

The $Log_{(10)}$ transformed degree of judgment change scores were input to a one-way ANOVA test of significance to ascertain if the differences between groups reached significance. A main effect of advice group was found ($F_{(2, 150)} = 8.33$, $p < 0.0005$). Follow-up planned comparisons were performed on the data and the only conditions that differed significantly were where participants indicated that no advice was useful, and people who indicated that they found reasons useful ($t = 4.56$; $df = 134.21$; $p < 0.0005$). Perhaps unsurprisingly, this finding

indicates that people in receipt of reasons-based advice changed judgment to a greater degree than people in who did not find any advice useful. However, caution should be used when drawing inferences from this analysis due to the low number of participants in the solely numeric advice condition.

I next examined the absolute difference scores only for people who perceived reasons-based advice (analogy, parallel case, authority, motivation) to be useful (see Table 4.6). The $\text{Log}_{(10)}$ transformed values of these scores met the assumption that the data was normally distributed. A one-way ANOVA was performed on these data, but no main effect of advice type was found ($F_{(3, 93)} = 2.16$, $p < 0.10$).

Table 4.6. Judgment change scores (years) for participants who perceive reasons to be useful.

N = 70 Reason type	Number of valid responses *N*=97	$\bar{x}$	*sd*	$\bar{x}$ ($\text{Log}_{(10)}$ transformed)	*sd*
Analogy	6	14.67	14.23	0.86	0.69
Parallel case	22	13.32	23.17	0.46	0.75
Authority	43	25.77	40.19	0.85	0.79
Motivation	26	8.08	15.41	0.49	0.60

However, follow-up planned comparisons reveal a significant difference in the degree of judgment change between people who preferred numerical advice substantiated by parallel case advice and participants who preferred advice evidenced by authority ($t = -2.03$; $df = 93$; $p < 0.05$). Also, a significant difference was found in the degree of judgment change between people who preferred reasons-based advice substantiated by some motivation, and participants who preferred reasons-based advice substantiated by an authority ($t = 2.01$; $df = 93$; $p < 0.05$). People in receipt of advice substantiated by authority changed judgment to a greater degree,

than people in receipt of advice evidenced by a parallel case, and people in receipt of advice evidenced by authority changed judgment to a greater degree than participants who found motivational advice useful.

Do participant's who value advice, follow it?

The preceding analysis has shown the influence of advice upon the degree of judgment change. I now turn to the *direction* of judgment change in relation to advice. The measure of judgment change relative to advice was calculated as the absolute difference between participants' unaided estimates and the advice, subtracted from the absolute difference between participants' post-advice estimates and the advice. As these data met the assumptions that the data was normally distributed, it was appropriate to use parametric methods in tests of significance. Table 4.7 provides means and standard deviations of participants judgment change relative to advice scores.

Table 4.7. Mean judgment change relative to advice scores

N = 37	Number of valid responses	$\bar{x}$	*sd*
Solely numeric advice	4	- 0.75	23.84
Reasons-based advice	51	10.92	56.51

Table 4.7 shows that only four valid responses of people who indicated that solely numeric advice was most helpful to them were available for analysis. Hence, comparing the responses of people who indicated that solely numeric advice was most useful to them with the responses of people who indicated that reasons-based advice was most useful to them, is not likely to be meaningful given the low *N* in the former condition. Instead, it is possibly more appropriate to determine whether the mean judgment change relative to advice scores of participants in the

reasons-based advice group are significantly different from zero (t = 1.38, df = 50, p < 0.17). This analysis shows that where participants who indicated that they found reasons-based advice most useful and subsequently revised their judgments post-advice, such judgment revision reduced the absolute numerical difference between a participant's pre-advice estimate and advice, and the same participant's post-advice estimate and advice, on average. However, the amount of this judgment revision is not significantly different from no judgment revision.

Next, the preceding analysis was repeated only for participants who had clicked on a radio button in the online survey, indicating that they had found one of the four types of reasons-based advice (i.e. analogy, parallel case, authority, motivation) to be most useful to them in formulating an estimate to the historical quiz questions (see Table 4.8).

Table 4.8. Mean judgment change relative to advice scores for people who perceived reasons to be useful.

Reason type	Number of valid responses (N = 51)	$\bar{x}$	*sd*
Analogy	4	22.00	11.17
Parallel case	10	29.10	53.71
Authority	25	5.12	70.72
Motivation	12	4.17	27.18

Table 4.8 shows the responses of participants' (N = 37) who revised their judgment post-advice and who indicated that they found reasons-based advice most useful in their deliberations. The judgment change relative to advice scores, were entered into a one-way ANOVA. No significant main effect of reason type was found ($F_{(3, 47)}$ = 0.53, p < 0.68). Follow-up planned comparisons were performed on the judgment change relative to advice data for participants

who indicated that reasons-based advice was most useful to them, and no significant differences were found between advice types. It is difficult to have confidence in this finding however, due to the low *N* for participants who indicated that they found analogous advice most useful in their deliberations.

In sum, the preceding analysis of judgment change has shown that participants in receipt of reasons-based advice, change judgment more often, than participants who did not perceive any advice to be useful. Moreover, participants in receipt of reasons-based advice tend to revise their judgments in the 'direction' of advice, whilst participants in receipt of solely numeric advice tend not to revise their judgments to any great extent. Of the participants who perceived reasons to be most useful, no differences between advice types was found in terms of judgment change relative to advice.

Consideration was next given to whether people who attributed some utility to the available advice, changed judgment post-advice, as a consequence. People who had indicated that they had found 'no advice useful' were excluded from this analysis. Recall that participants indicated which piece of advice was useful to them in formulating an estimate, by clicking on a radio button in the online survey. Each piece of advice was assigned a different whole integer of (1-8), which was unknown to participants. The four pieces of solely numeric advice were assigned integers 1-4; whilst the four pieces of reasons-based advice were assigned integers 5-8. Hence, it was possible to determine the median value of the reported utility of advice measure for each participant, responding to the three historical quiz questions (see Table 4.9).

Table 4.9. (N = 37) Median utility of advice ratings

Historical quiz question		Perceived utility of advice rating (Median)	Range	Scores > Median	Scores <= Median
Date the first Japanese railway opened (*N* = 45)	Participant revised judgment (*N* = 18)	7	7	4	14
	Participant did not revise judgment (*N* = 27)	7	7	7	20
Date of the British mandate over Palestine	Participant revised judgment (*N* = 16)	6	7	7	9
	Participant did not revise judgment (*N* = 12)	6	7	6	6
Date of the first Camp David summit	Participant revised judgment (*N* = 21)	7	7	6	15
	Participant did not revise judgment (*N* = 9)	8	7	0	9
All questions (aggregated) (Participants *N* = 61)	Participant revised judgment (*N* responses = 55)	7	7	12	43
	Participant did not revise judgment (*N* = responses 48)	7	7	14	34

Table 4.9 shows that the median scores of the utility attributed to advice by participants are very similar. The aggregated scores show that irrespective of whether participants changed judgment or not, post-advice, there were more utility of advice scores reported below the median than above it. In other words, there appears to be little evidence that the perceived utility of advice scores – considered in isolation from other variables – were related to subsequent judgment revision.

Does attending to advice ultimately lead to more accurate estimation?

Finally, the accuracy of our participants' estimates was examined. Prior to examining participants' accuracy scores, it was necessary to revisit the issue of counterbalancing the numerical value of advice discussed earlier. Recall that the counterbalancing scheme employed in this study was to provide participants with equal numbers of advice that overestimated, and underestimated the true value in two out of three questions. The remaining historical quiz question contained equal numbers of underestimates, and overestimates in the solely numeric advice boxes, and two underestimates, one overestimate, and the true answer, in the reasons advice boxes. As a manipulation check, the responses of participants who did not change judgment (and hence are inferred not to attend to advice), were compared with the responses of participants who did change judgment post-advice (see Table 4.10).

Table 4.10 Overestimates and underestimates of participants who did/did not change judgment post advice

Historical quiz question	Participants who did not change judgment post-advice	Participants who did change judgment post-advice
Question 1	Post-advice estimate is less than true value	Post-advice estimate is less than true value
Question 2	Post-advice estimate is greater than true value	Post-advice estimate is greater than true value
Question 3	Post-advice estimate is less than true value	Post-advice estimate is less than true value

Table 4.10 shows that the provision of numerical advice that varied around the true answer of the three historical quiz questions, appears to have had little influence upon whether participants, either underestimate, or overestimate, the true answer post-advice. This is because the same pattern of overestimation, and underestimation, obtained irrespective of whether participants changed judgment, or not, post-advice, on average. Whilst there may well have been alternative ways to counterbalance numeric advice, the scheme utilized here at least did not confound the results of the analysis.

Next, the accuracy of participants' estimates was examined post-advice. Absolute accuracy was calculated as the absolute difference between the correct historical date in each estimation problem, and participant's post-advice estimates. To satisfy the assumption of normality in the data, the Log_{10} transformed values were used in tests of significance. Recall that participants clicked on a radio button during the experiment indicating which piece of advice they had found most useful in formulating an estimate to each of the estimation problems, this data was aggregated into the conditions of (i) no advice perceived to be useful (ii) solely numeric advice

perceived to be useful (iii) and reasons advice perceived to be useful (see Table 4.11). Seventy participants provided data, and although 163 responses were recorded by the MouseLabWeb software, only 153 post-advice estimates were available for analysis as ten data entries were omitted from the advice utility radio button field. The absolute accuracy scores for participants in each advice type condition were entered into a one-way ANOVA, and no significant differences between conditions were found ($F_{(2, 150)} = 0.17$, $p < 0.85$).

Table 4.11. Mean absolute accuracy scores (years) post-advice

Reason type	*N*= 153 responses	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed	*sd*
No advice perceived to be useful	50	45.74	27.29	1.52	0.47
Solely numeric advice perceived to be useful	6	38.83	20.80	1.52	0.31
Reasons-based perceived to be useful	97	41.04	22.57	1.47	0.45

Despite a non-significant main effect, follow-up planned comparisons were performed on the data so that any differences between specific conditions should not be missed – no significant differences were found.

The distribution of data points for the post-advice accuracy scores for participant's who changed judgment post-advice, approximated a normal distribution. Hence, a one-way ANOVA was performed on the untransformed post-advice accuracy scores of participants (*N* = 43) who revised their judgment post-advice (see Table 4.12). Sixty-three responses were

available from participants who revised their judgment post-advice, however one response could not be included in the analysis as this participant had not provided a advice utility rating – leaving a total of 62 responses from 43 participants. For participant's who revised their judgment post-advice, there were no significant differences found between conditions ($F_{(2, 59)} = 2.69$, $p < 0.08$). However, follow-up planned comparisons revealed a significant difference in the post-advice accuracy scores of participants in receipt of reasons based advice and participants who did not find any advice to be useful ($t = 2.29$, $df = 59$, $p < 0.03$). This finding is difficult to accept at face value, as the condition in which participants indicated that they found no advice to be useful, only contained seven responses. Hence, this finding is undermined by low statistical power.

Table 4.12 Participants who changed judgment post-advice: Mean absolute accuracy scores

Reason type	*N*= 62 responses	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed	*sd*
No advice perceived to be useful	7	26.43	24.52	1.10	0.71
Solely numeric advice perceived to be useful	4	39.50	13.33	1.58	0.13
Reasons perceived to be useful	51	46.29	21.56	1.58	0.35

Hence, although these results are not inconsistent with a conformity effect (participants in receipt of solely numeric advice, are ultimately no more accurate post-advice, than participants in receipt of reasons-based advice), it is difficult to accept this finding as definitive, due to the low N in the no advice perceived a useful condition.

Does attending to a particular advice type, advantage participants in terms of changes in accuracy?

Consideration was next given to the absolute accuracy scores of only participants ($N = 50$) who perceived reasons to be most useful (see Table 4.13). The distribution of data points of these scores satisfied the assumption of normality, so the untransformed scores were used in analysis of significance. The absolute accuracy scores of participants who perceived particular advice types to be useful (analogy, parallel case, authority, motivation), were entered into a one-way ANOVA, but no significant main effect of group was found ($F_{(3, 93)} = 0.39$, $p < 0.76$).

Table 4.13. Absolute accuracy scores post-advice

	$N = 97$ responses	$\bar{x}$	*sd*
Analogy	6	36.33	16.32
Parallel case	22	37.36	24.22
Authority	43	42.74	23.95
Motivation	26	42.42	20.57

Follow up comparisons were performed on the data in Table 4.13 in order to check that the lack of a main effect of group did not mask significant differences between pairs of advice types. However, no significant differences between advice types were found.

Similarly, no significant main effect of group (in terms of absolute accuracy scores) was found when the preceding analysis was repeated, only for those participants who changed judgment, post-advice ($F_{(3, 47)} = 1.86$, $p < 0.15$) (see Table 4.14).

Table 4.14. Participants who change judgment: absolute accuracy scores post-advice

	N	$\bar{x}$	*sd*
Analogy	4	38.00	4.00
Parallel case	10	34.80	27.07
Authority	25	52.50	12.41
Motivation	12	46.33	19.36

In order to determine that the non-significant main effect of group did not mask significant differences between pairs of advice types, follow-up planned comparisons were performed on the data in Table 4.14. This analysis revealed a significant difference in post-advice accuracy scores between participants who reported a preference for reasons-based advice evidenced by an analogy, and participants who reported a preference for reasons-based advice evidenced by an authority (t = - 3.12, df = 25.27, $p < 0.004$). However, given that only four responses were available for analysis in the analogy condition, it is difficult to have confidence in this finding.

Next, participants' accuracy change scores were examined. The measure of accuracy change was calculated as the absolute difference between the true value and a participant's post-advice estimate, subtracted from the absolute difference between the true value and a participant's pre-advice estimate. Negative scores indicate that participants, where they do change judgment, make less accurate estimates, than they were able to formulate unaided. Low scores (i.e. scores that are close to zero), describe participants, who, where they do change judgment, do not make estimates that are very much more accurate, than estimates participants made unaided. High (positive) scores indicate participants who revise their judgment, so that their post-advice estimate is more accurate than their pre-advice estimate. The scores of participants

who did not revise their judgments were excluded from the following analysis, as was the response of a single participant who had not indicated if any advice or no advice had been useful. This meant that 63 responses from 43 participants were available for analysis. These scores were aggregated into the conditions of (i) no advice perceived to be useful (ii) solely numeric advice perceived to be useful (iii) and reasons advice perceived to be useful (see Table 4.15). In the interests of clarity the sign of the means scores were reversed – positive scores indicate accuracy improvements.

Table 4.15. Accuracy Change (Participants N = 43)

Reason type	*N* = 63 responses	$\bar{x}$	*sd*
No advice perceived to be useful	7	20.14	32.01
Solely numeric advice perceived to be useful	4	1.75	26.42
Reasons-based advice perceived to be useful	52	13.87	46.76

The distribution of data points from Table 4.15 approximated a normal distribution and hence the necessary assumptions for the use of parametric statistical tests were met. A one-way ANOVA was performed on the accuracy change data and no significant main effect was found ($F_{(2,\ 60)} = 0.22$, $p < 0.81$). The analysis was repeated only for participants ($N = 38$) who perceived reasons to be most useful (see Table 4.16).

Table 4.16. People who perceive reasons to be useful: mean accuracy change scores

	N = 51 responses	$\bar{x}$	*sd*
Analogy	4	22.00	11.16
Parallel case	10	8.70	40.03
Authority	25	18.48	60.33
Motivation	12	8.83	24.67

A one-way ANOVA was performed on the data in Table 4.16 and no main effect of group was found ($F_{(3, 47)} = 0.19$, $p < 0.90$). Follow-up planned comparisons were performed to check for significant differences between advice types and none were found to be significant. These findings suggest that where participants did change judgment, post-advice, there was no clear advantage to having attended to any one of the four types of reasons advice. Some caution should be exercised in the interpretation of this finding however, due to the low numbers of participants in some of the cells of Tables 4.15 and 4.16.

Are participants in this study merely conforming (to a degree) to any available advice?

Finally, analyses were performed to establish the effect upon accuracy, of the relative positions of participants' pre-advice estimates, the true answer, and the available advice. Recall that there are six possible permutations of pre-advice estimate, advice, and true answer (see Table 4.17). If, as I have argued, participants are constrained in their abilities to discriminate between potentially beneficial advice and poor advice, then we might expect improvements in accuracy in permutations A, B, C, and F, but increases in error in permutations D, and E.

Should this pattern of results be observed, this would provide evidence that our participants were indiscriminately conforming to any available advice.

Table 4.17. Likely judgment revisions in the possible permutations of true answer and advice, in relation to an intuitive pre-advice estimate

Permutation	*Increase in* → *numerical value*			judgment revision due to conformity effect
A	Estimate	True answer	Advice	towards 't'
B	True answer	Advice	Estimate	towards 't'
C	Advice	True answer	Estimate	towards 't'
D	Advice	Estimate	True answer	opposite to 't'
E	True answer	Estimate	Advice	opposite to 't'
F	Estimate	Advice	True answer	towards 't'

To explore this suggestion, the 'permutations' from Table 4.17 can be thought of where (i) permutations B and F describe the scenario where in numerical terms, the advice lies between an unaided estimate, and true answer (ii) permutations A and C describe the scenario where in numerical terms, the true answer lies between the advice, and an unaided estimate (iii) permutations D and E describe the scenario where in numerical terms, an unaided estimate lies between the advice, and the true answer. Support for the idea that participants indiscriminately conform to any available advice, would be found if improvements in estimation accuracy were

observed for the scenario where the advice lies between an unaided estimate, and true answer, and also if improvements in estimation accuracy were observed for the scenario where an unaided estimate lies between the advice, and the true answer. Simultaneously, estimation accuracy will decrease, where an unaided estimate lies between the advice and the true answer. The accuracy change scores from Table 4.18 were entered into a one-way ANOVA and a significant main effect of condition was found ($F_{(2, 51)} = 4.30$, $p < 0.02$).

Table 4.18. Accuracy change: relative positions of true answer, advice and pre-advice estimate (positive mean scores indicate accuracy improvements).

Advice position	*N* = 99 number of responses	$\bar{x}$	*sd*
(1) Advice between estimate and true answer (*N* participants = 27)	34	23.82	45.75
(2) True answer between advice and estimate (*N* participants = 33)	38	1.89	25.06
(3) Estimate between advice and true answer (*N* participants = 24)	27	- 2.15	16.59

Follow-up planned comparisons showed a significant difference between advice position (1) and advice position (3) in terms of participants accuracy change scores ($t = - 2.54$, $df = 51$, $p < 0.014$). Significant differences were found in accuracy change scores, between advice position (1), and (2) ($t = - 2.89$, $df = 51$, $p < 0.006$). However, the accuracy change scores of participants ($N = 24$) whose unaided estimate fell between the advice and the true answer, did not differ significantly from zero ($t = 0.67$, $df = 26$, $p < 0.51$). Hence, limited support is found

here for the idea that participants indiscriminately conform to advice. This is because improvements in estimation accuracy were observed for the scenario where the advice lies between an unaided estimate, and true answer. Simultaneously, estimation accuracy decreased marginally where an unaided estimate lies between the advice and the true answer. These results are consistent with the idea that people are constrained in their abilities to discriminate between beneficial and poor advice, in circumstances of incomplete domain knowledge.

Are participants consistent in their preferences for a particular type of advice?

Consideration was next given to whether a participant's preference for (i) their own intuition, or (ii) any particular advice type, was a consistent feature of his/her deliberations prior to revising (or not revising) their unaided estimate. A pattern of consistent advice preferences may indicate something more than mere curiosity about the nature of the various pieces of advice. Recall, that participants clicked a radio button in the online survey that indicated which of the nine pieces of advice (four pieces of solely numeric advice, four pieces of reasons advice, and the option of 'no advice useful') was useful to them in formulating an estimate to each historical quiz question. Hence, the number of instances in which participants chose the same type of advice (i.e. as indicated by clicking the radio button that such advice was most useful), on subsequent estimation questions, as they did when responding to the first estimation question, was calculated. The data is aggregated and tabulated below (see Table 4.19).

Table 4.19. Consistent advice preferences

Participants who initially cite reasons-based advice as useful, and continue to cite this advice type as useful when responding to subsequent estimation questions	Participants who initially cite numerical advice as useful, and continue to cite this advice type as useful when responding to subsequent estimation questions	Participants who initially cite no advice as useful, and continue to cite no advice as being useful , when responding to subsequent estimation questions	Participants who are not consistent in their preference for advice.
19/45 (42.2%)	0	7/45 (15.6%)	19/45 (42.2%)

A significance test was performed, in order to determine whether the proportions tabulated above, were significantly different from what could have occurred by chance, finding that the observed proportions were significantly different from the expected proportions ($\chi^2 = 22.72$; $df = 3$; $p < 0.001$). Given that people who cited solely numeric advice as most useful to them, in formulating an estimate to the first historical quiz question, did not subsequently find solely numeric advice useful on the other two historical quiz questions, this finding is perhaps not surprising. However it is plausible that this finding provides some indication, that participants' preferences for advice were guided by something other than mere curiosity that declined over time. If so, this provides further evidence of the limited ecological validity of conceptualizing advice in solely numeric terms. Further consideration of participants' preferences for different advice types is developed in the following section.

Do participants hold preconceptions of the utility of any particular advice type?

In order to further investigate participants' preferences for different types of advice, the time of first acquisition of advice was calculated. Whilst participants interacted with the on-line questionnaire, the 'MouseLabWeb' software recorded the time in milliseconds (calculated as the absolute difference from the start of the questionnaire until a particular event). Given that both question order and advice box presentation was randomized across participants, it is possible to determine the order in which people first accessed advice. However, despite $\text{Log}_{(10)}$ transformation the time data violated the assumption of normality (in addition, the assumption that our data was normally distributed could not be satisfied despite attempting arcsine, square root, $\text{Log}_{e,}$ and arctan transformations). Hence, it was necessary to use non-parametric techniques to test for significance.

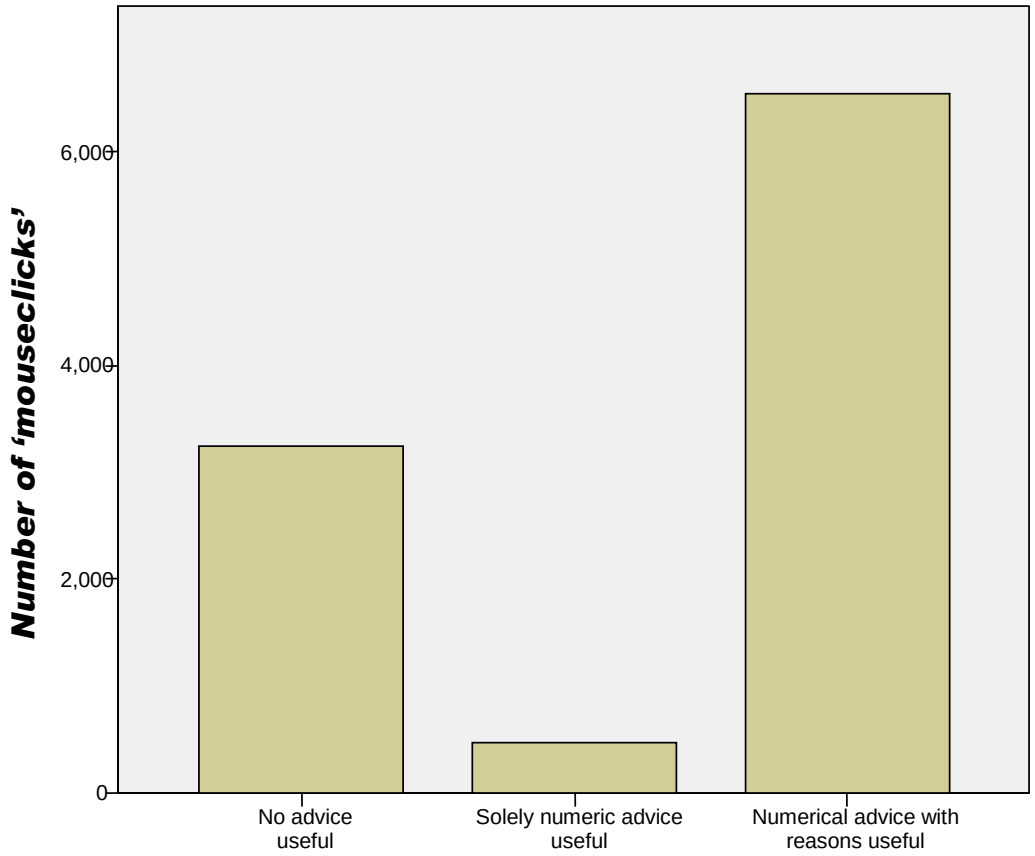
First, the data was aggregated across estimation problems, and collapsed into the advice categories of (i) the time of first acquisition of solely numeric advice, and (ii) the time of first acquisition of reasons (iii) the time when participant's first clicked on the 'no advice useful' option. A Kruskal-Wallis test, comparing the mean times of first acquisition of advice for participants who first chose solely numeric advice, for participants who first chose reasons advice, and for participants who first chose 'no advice', was performed. On examining this data, it was clear that, on average, there was no significant effect of advice type upon time of first acquisition of advice ($\chi^2 = 0.54$, $df = 2$, $p < 0.76$). A similar result obtained when the data of only participants who subsequently changed judgment post-advice was examined ($\chi^2 = 0.64$, $df = 2$, $p < 0.73$), and also for people who did not change judgment post-advice ($\chi^2 = 0.24$, $df = 2$, $p < 0.89$). Moreover, there were no significant differences in the time that participants first accessed advice, when the consistency of their advice type preferences was taken into account. Comparison was made between participants who consistently reported that no advice was

useful, participants who reported that reasons advice was useful, and where participants were inconsistent in their preference for advice types ($\chi^2 = 0.36$, $df = 2$, $p < 0.83$). These findings suggest that participants did not hold a pre-conceived idea of what advice might be useful to them, prior to accessing and processing advice messages.

What advice did participants look at prior to revising (or not revising) their judgment?

Next, consideration was given to what advice (if any) our participants elected to examine prior to revising (or not revising) their initial estimate. The 'MouseLabWeb' software recorded which advice boxes participants 'clicked' on as they completed the questionnaire, and subsequently it was possible to determine the total number of advice 'acquisitions' and 're-acquisitions' participants made, prior to revising (or not revising), their ultimate estimate. These data were aggregated in order to ascertain whether participants found no advice useful, attended to solely numeric advice, or attended to reasons. These data are displayed in Fig. 4.3 below.

Fig 4.3. Total number of advice acquisitions and re-acquisitions.



Subsequently, a significance test was performed to determine whether the observed number of advice box acquisitions for the categories of no advice perceived useful, solely numeric advice, and reasons-based advice, was different from what might be expected by chance. A significant result of this test would provide evidence that participants' search for information was not random – this proved to be the case ($\chi^2_{\text{Friedman}} = 1808$, $df = 2$, $p < 0.0005$). This finding demonstrates that participants gave greater consideration to reasons-based advice. Next, the analysis was repeated, this time comparing the observed number of advice box acquisitions for people who subsequently did not change judgment ($N = 5856$), and participants who did subsequently change judgment ($N = 4564$), with the expected frequency of advice box acquisitions ($N = 5210$). It was found that the pattern of observed frequencies was

significantly different from what might be expected by chance (χ^2 = 160.20, *df* = 1, *p* < 0.0005). However, this pattern of non-random observations did not support the idea that participant's who did not revise their judgment post-advice, conducted a less rigorous information search pre-advice, than participants who did revise their judgment post-advice. People who did not change judgment post-advice, tended to acquire (and re-acquire), no less information than people who did subsequently change judgment (U = 397575.5, N_1 = 853, N_2 = 955, *p* < 0.38). In sum, there was no difference between conditions in terms of the *depth* of information search that participants undertook, but participants gave greater consideration to reasons-based advice, than either solely numeric advice, or their own intuitions.

Does attending to advice bolster participants' confidence in their estimation abilities?

Recall that the measure of confidence proposed here is the absolute difference between a lower bound, and an upper bound within which participants are 95% certain that the true answer (whatever that might be) could be found. Hence, it is possible to infer that a smaller absolute difference (between upper and lower bounds) is indicative of a greater confidence. The confidence scores reported by participants' were not normally distributed, which necessitated the use of the $Log_{(10)}$ transformed values in order to meet the requirements of the assumption of normality. It follows then, that if participants were more confident in the accuracy of their estimates post-advice; participants would report a significant difference between pre-advice confidence levels, and the confidence levels reported post-advice (see Table 4.20). This is exactly what was observed (*t* = 5.12, *df* = 158, *p* < 0.0005).

Table 4.20. Pre-advice and post-advice levels

N participants = 66	*N* = number of responses	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed	*sd*
Pre-advice confidence	159	54.45	54.76	1.47	0.56
Post-advice confidence	159	36.92	35.24	1.30	0.57

Overall, participants appear to be significantly more confident after considering advice, than they were prior to considering advice. However, in order to determine whether this effect masked differences between groups in terms of pre-advice confidence levels a one-way ANOVA was performed on the pre-advice confidence scores of participants (see Table 4.21).

Table 4.21. Pre-advice confidence

Reason type *N* participants = 71	*N*= 153 responses	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed	*sd*
No advice perceived to be useful	50	38.62	37.82	1.32	0.57
Solely numeric advice perceived to be useful	6	47.00	45.76	1.49	0.45
Reasons-based advice perceived to be useful	97	65.46	60.81	1.60	0.49

The ANOVA performed on the data in Table 4.21 revealed significant differences between conditions ($F_{(2, 150)} = 4.70$, $p < 0.01$). It was then necessary to carry out follow-up planned comparisons to ascertain which conditions differed. The results of the analysis of pair-wise

comparisons shows that only participants who perceived reasons-based advice to be useful, and participants who perceived no advice to be useful to them significantly differed in terms of their pre-advice levels of confidence (t – 3.07, df = 150, p < 0.003). This finding is perhaps not that surprising, given that the number of participants indicating that they found reasons-based advice useful was almost double the number of participants, who indicated that they found no advice useful to them in their deliberations. Further, participants who did not find any advice useful in their deliberations appeared to be most confident in their deliberations pre-advice, and remained confident in their estimates post-advice (t = 1.30, df = 49, p < 0.20) (i.e. these participants did not significantly revise their level of confidence post-advice). Hence, participants who were least confident in their pre-advice estimates became more confident after they had considered advice.

Next, consideration was given to whether there were differences in levels of post-advice confidence dependent upon the type of advice our participants found most useful (none, solely numeric, or reasons). Table 4.21 reports means and standard deviations for participants' post-advice confidence.

Table 4.22. Post-advice confidence

Reason type *N* participants = 66	*N* = Number of responses	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed	*sd*
No advice perceived to be useful	50	34.20	30.27	1.28	0.58
Solely numeric advice perceived to be useful	5	34.40	27.10	1.29	0.64
Reasons-based advice perceived to be useful	94	40.48	38.46	1.36	0.54

Table 4.22 shows that participants who found no advice to be useful, had the least post-advice confidence, yet there were no significant differences between groups ($F_{(2, 146)} = 0.34$, $p < 0.69$). No significant differences between groups were found when follow-up planned comparisons were performed on the data. Similarly, no differences were found between groups when the analysis was repeated for people who had changed judgment ($F_{(2, 57)} = 0.39$, $p < 0.68$). Further, there were no differences in post-advice confidence scores, between people who changed judgment, and people who did not (see Table 4.23), (t – 0.94, $df = 155$, $p < 0.35$).

Table 4.23 Judgment change: post-advice confidence scores

N participants = 43	*N* = number of responses	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed)	*sd*
Change judgment	61	35.43	39.51	1.26	0.58
Do not change judgment	96	38.65	32.35	1.35	0.55

Consideration was then given to whether participants who perceived reasons as most useful, and who changed judgment, differed in terms of their final confidence (see Table 4.24) – they did not ($F_{(3, 46)} = 0.91$, $p < 0.45$).

Table 4.24 Post-advice confidence: Participants who perceive reasons-based advice be useful and change judgment

N participants = 36	*N* = 50 number of responses	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed)	*sd*
Analogy	3	11.33	8.08	0.97	0.35
Parallel case	10	39.90	29.78	1.12	0.73
Authority	25	41.12	51.06	1.38	0.47
Motivation	12	36.28	32.76	1.31	0.56

Given the results of the preceding analysis of participants' post-advice confidence scores between groups, reported in Table 4.22, and the finding that there were no significant differences between groups when the post-advice confidence scores of participants who changed judgment, and perceived reasons to be useful, were examined (see Table 4.24); it is possible to tentatively infer that participants' confidence in the veracity of their post-advice estimate was bolstered to the same degree, no matter which type of advice they perceived to be most useful. This finding remains tentative however, due to the low number of participants in three out of four cells in Table 4.24, and the subsequent detrimental effects upon statistical power.

If participants' are *less* confident prior to accessing advice, are they *more* likely to revise their judgment post-advice?

Next, participants' pre-advice confidence scores were examined. Recall that one of the issues that is being explored here, is whether people who report narrow confidence intervals (i.e. are highly confident in the veracity of their unaided estimate), are less predisposed to change judgment, than participants who report wider confidence intervals in their unaided pre-advice estimate. To explore this issue, the pre-advice confidence levels of participants who do not subsequently change judgment, were compared with the pre-advice confidence levels of participants who did change judgment (see Table 4.25).

Table 4.25 Pre-advice confidence levels

N participants = 55	*N* = 103 number of responses	$\bar{x}$	*sd*	$\bar{x}$ ($Log_{(10)}$ transformed)	*sd*
Change judgment	55	72.25	68.17	1.62	0.53
Do not change judgment	48	55.38	48.19	1.56	0.43

Table 4.25 clearly shows that although participants who changed judgment had appear to have held less confidence initially, than participants who did not change judgment, this difference did not reach the level of significance (t = 0.55, df = 101, p < 0.59). However, the idea that participants (N = 71) who are *less* confident pre-advice are subsequently more disposed to revise their judgment in the light of advice, is supported when the correlation between judgment change, and pre-advice confidence was examined (Spearman's 'r' = - 0.19, p < 0.02, N = 163). Here the negative correlation coefficient suggests that the greater the numerical difference is between a participant's lower and upper bound confidence estimate, the more likely the same participant is to revise their judgment post-advice.

Does attending to advice, lead to changes in the degree of confidence held by participants?

The changes in confidence that our participant's reported were examined next. Recall that the confidence change measure here is calculated by subtracting a participant's initial confidence score from their final confidence score. Hence, negative values indicate *increasing* confidence. For the purposes of clarity, the sign of the mean confidence change score has been inverted, so that positive confidence change scores indicate *increasing* confidence. After extracting cases where participants did not change their level of confidence from their pre-advice estimates, to their post-advice estimates, these scores are tabulated in Table 4.26 (below).

Table 4.26 Changes in confidence

N participants = 51	*N* = 99 number of responses	$\bar{x}$	*sd*
No advice perceived to be useful	20	11.05	31.27
Solely numeric advice perceived to be useful	5	20.00	34.07
Reasons-based advice perceived to be useful	74	33.31	59.20

As these data did not meet the assumptions of normality, and $Log_{(10)}$ transformation of negative values is not meaningful, it was necessary to perform a Kruskal-Wallis test of significance upon the data in Table 4.26 to determine if there was a significant main effect of condition, in terms of the confidence change scores. The result of this test revealed a main effect of condition (no advice perceived as useful/reasons advice perceived as useful/numeric

advice perceived as useful) ($\chi^2 = 10.09$, $df = 2$, $p < 0.006$). Next a median test was performed on the data to determine which groups significantly differed from each other (see Table 4.27).

Table 4.27 Median scores: Frequency of changes in confidence scores

N participants = 51 Median = - 10	> median	<= median
No advice perceived to be useful	7	43
Solely numeric advice perceived to be useful	1	4
Reasons-based advice perceived to be useful	17	77

However, a subsequent test upon the group medians failed to detect significant differences between groups ($\chi^2 = 0.43$, $df = 2$, $p < 0.81$). Inspection of Table 4.27 only reveals that the majority of participants narrowed their post-advice confidence bounds in comparison to their pre-advice confidence bounds, but did so less so than the overall group median, on average (10 years).

4.7 Discussion

This study had set out to explore influences upon judgment revision (where both solely numeric and reasons-based advice constituted the advice that was available to participants) upon participants' deliberations, and subsequent estimation, of the dates of three historical events. The results indicate the importance that participants' attach to reasons (irrespective of the accuracy of the reasons) in their deliberations. In part, this preference may be accounted for by the fact that reasons are simply more elaborate than solely numeric advice, and hence more attractive to estimators. However, this may not be the only reason why participants did not regard solely numeric advice as useful, on average. Reinard (1988) speculates that the failure of statistical evidence to induce persuasion in judges is because such evidence fails to 'create vivid images in the minds of receivers' (p. 24). Statistical evidence in this context, potentially allows judges to make inferences about the nature of the population that the statistical evidence is drawn from. Given that the few existing studies (Harte, 1973; Ryland, 1973), which examine the persuasiveness of solely numeric, or statistical information, in comparison to the persuasiveness of testimonial assertions, show no differences - I considered why my participants' clearly sought out advice that offered reasons in addition to numerical information. Intuitively, advice that only consists of a simple four digit number is unlikely to 'fire the imagination' of participants – and it has been argued that effective argumentation relies, at least in part, in its power to create vivid images in the minds of an audience (Perelman and Olbrecht-Tyteca, 1969). Further, solely numeric information is only likely to be influential in the context of the three historical quiz questions utilized in this study, when people are able to incorporate such information within their own implicit theories of causation (Ross and Fletcher, 1985). The questionnaire devised in this study presented solely numeric advice, which is perhaps even less likely to create vivid imagery in the minds of participants than statistical information, as beyond an individual's world knowledge, few inferences about some target population can be made on the basis of an isolated historical date. Moreover, there

is a body of evidence to suggest, that on average, anecdotal reports are more persuasive than solely numeric advice, or statistics (Bar-Hillel, 1975; Kahneman and Tversky, 1972, 1973; Nisbett and Borgida, 1975). The findings of this study are consistent with those of these authors - in so far as participants in this study clearly found reasons more useful than solely numeric advice.

Moreover, participants in this study also tended to 'stick' with their initial advice preferences over subsequent estimation questions. Such inflexibility on the part of participants did not reflect accurate estimation however. Rather, participants may have been so constrained in their abilities to evaluate the veracity of advice substantiated by reasons, that investing cognitive resources in evaluating further advice options did not appear any more beneficial than persisting with the initial advice preference. If the assumption that participants were motivated to estimate accurately is correct, it follows that participants saw no reason to invest cognitive resources and adapt their estimation strategy to the demands of each individual historical quiz question; but neither were they guessing blindly – participants were at least motivated to estimate in an internally consistent manner (by persisting with an initial advice preference over later estimation questions). Self-consistency of advice preference, in this context, is best explained through a cognitive dissonance perspective. When faced with an estimation problem for which advice is available, participants scanned their previous decisions in order to make consistent future decisions about advice (Harmon-Jones, 1981), and hence reduce internally perceived cognitive dissonance.

A strong finding in this study is that in circumstances of incomplete domain knowledge, people seek cues from which an evaluation of (i) the accuracy of a participant's unaided estimate can be made, and (ii) the potential benefits of attending to advice. Of the advice substantiated by

reasons in this study, advice backed up by some authority clearly emerged as influential. Here, participant's sought out such advice, and inferred that the advice was useful on the basis of the credibility of the source. Recall that in the context of this study, numerical advice evidenced by some authority, was advice substantiated by the authority of three fictitious historians. This finding is consistent with the work of Yalch and Elmore-Yalch (1984). These authors presented their participants with messages that varied on the dimensions of quantity and expertise. The presentation of quantitative information in the Yalch and Elmore-Yalch study stimulated participants to rely upon source credibility cues as the basis of judgment. Here, advisors perceived by judges as 'expert' elicited greater persuasive effects, than advisors perceived as possessing *less* expertise. In contrast, participants in receipt of advice that did not include numeric information in the Yalch and Elmore-Yalch study, were uninfluenced by the perceived expertise of the advisor. Here, source credibility did not influence the persuasiveness of the message. Yalch and Elmore-Yalch (1984), contend that the quantitative component of reasons elicits the processing of peripheral cues such as source credibility. These findings imply that persuasion may be undermined where numerical advice is evidenced by some non-expert authority. In contrast, the advice substantiated by authority presented to participants in this study, explicitly highlighted the expertise of the source (see Appendix 1, Screen 4). Here, persuasion was enhanced, as participants' focused on the credibility of the source of the advice, and not on the accuracy of the numerical component of advice. In sum, people became disinterested in numbers and stopped paying attention to the content of the message (Anderson and Jolson, 1980). In our study, participants' rated advice evidenced by authority as most useful (even though it was not the most accurate), and people responding to this advice changed judgment to a greater extent (in the direction of the advice), than people who were not in receipt of reasons. In sum, our participant's responded to the credibility of the source of the advice, as a proxy for the veracity of the advice.

The findings discussed in the preceding section consistent with the idea of two routes to persuasion – central and peripheral processing of message arguments (Chaiken, 1980; Petty *et al.* 1981; Petty *et al.* 1983). Here, it is theorized that central processing of message arguments involves considerable cognitive resources, and results in detailed scrutiny of the merits of the message. In contrast, where the message is of little personal importance to the recipient, cognitive resources are sparingly allocated to the processing tasks, and judgment is based upon peripheral cues (such as the credibility of the source of the message).

Ultimately, the improvements in estimation accuracy reported by participants, can be explained by the idea that people conform to any available advice (to a degree), under conditions of uncertainty. This is because improvements in estimation accuracy were observed for the scenario where the advice lies between an unaided estimate, and the true answer. Simultaneously, estimation accuracy decreased where an unaided estimate lies between the advice and the true answer.

However, not all participants were swayed by the eight pieces of advice available to them in this study – 33.12% of our participant's rejected any advice and relied entirely on their own intuition. In these circumstances I observed that people who were highly confident in their own intuition were unlikely to change judgment post-advice. Similarly, people who were not confident in their own intuition were disposed to change judgment post-advice. These findings suggest that the reports of participant's confidence may also be a proxy for participants' pre-existing knowledge of the three historical quiz questions that they were tasked with estimating. However, there is no evidence that initial confidence is correlated with ultimate estimation accuracy. This implies that participants may have erroneously believed that their intuition was

accurate i.e. their confidence was not well calibrated with their estimation accuracy (Lichtenstein, Fischhoff and Phillips, 1982).

This study is exploratory in nature, and no specific predictions were made as to the estimation behaviour of participants post-advice. In this respect the study was severely limited, and future research should be based on experimentation where the influence of one set of independent variables upon various dependent measures can be closely controlled and measured. On reflection, it would have been appropriate to develop a more elegant counterbalancing scheme for the numeric component of advice, than the one utilized here – despite the fact that the counterbalancing scheme utilized here does not appear to have confounded the results of the study. A secondary, but highly important issue was the failure to successfully disentangle the influence of reason-based advice on judgment changes, from simple conformity to advice effects. In part, this is attributable to a conceptualization of what constitutes 'advice'. Here, the use of solely numeric advice, and reasons, was not entirely successful, as participants overwhelmingly discarded the solely numeric advice. Future studies should carefully consider and evaluate the potential benefits of using solely numeric advice. The findings of this study suggest that, solely numeric advice was not perceived as meaningful by our participants. The use of statistical advice however (and/or a description of the process from which the statistic is derived), allows participants to make inferences about the target population that a statistic is drawn from. Further, such elaboration may allow people to place advice within the context of their 'implicit theories of causation', and facilitate deeper processing of the message content of the advice – even though this remains at best an indirect route in facilitating enhanced estimation accuracy.

The limitations of this study serve also to highlight the underdeveloped status of the JAS paradigm. Recall, that JAS theorists argue that acknowledging role differentiation between judges and advisors, and placing judge/advisor interactions in a social context, enhances the paradigm's ecological validity in comparison to alternative nominal group decision-making techniques. However, these claims are difficult to substantiate, given there is currently no comprehensive theoretical framework of advice taking, and advice giving (Bonaccio, and Dalal, 2006). Moreover, it is perhaps not surprising that no comprehensive theoretical framework exists, given the heterogeneity of tasks, stimuli, and methods that characterize JAS studies. Hence, claims of the superior ecological validity of JAS research are difficult to sustain. Further, as the interaction between judge and advisor is impoverished and artificial in the extant JAS literature, it is difficult to isolate the cognitive mechanisms that underpin persuasion and judgment change. This is because of loose conceptualizations of what constitutes 'advice' (to date, advice has been largely operationalized in solely numeric terms). Further, the existing conceptualization of advice disguises the reasons that underpin deliberations that ultimately resolve in a numeric estimate of some quantity of interest. In sum, future studies should conceptualize and operationalize exactly what constitutes 'advice', and predict *a priori* its likely effects upon accuracy, judgment change and confidence. Further, future studies will also have to address the issue of the persuasive strength of any particular rationale that is included in an advice package. It is also clear that participants are able to use source credibility as a proxy for the veracity of advice, and simultaneously report confidence levels that may be based upon pre-existing knowledge of the quantity to be estimated. In such circumstances it will be necessary to take some measure of participants' general knowledge, in addition to a measure of the specific knowledge participants are able to bring to bear upon an estimation question.

Given the preceding discussion, and the clear need for further development of the JAS paradigm, the initial exploratory study outlined here, serves as the basis for future empirical work. One robust finding from the extant JAS literature (and a finding partially replicated in this study – 33.12% of participants did not find any advice useful in this study - see Table 4.2), is that people are not predisposed to revise their judgments in the light of advice. Yaniv and Kleinberger (2000), have described this phenomena as 'egocentric discounting' of advice. As advice is operationalized in solely numeric terms, it is argued that judges in the Yaniv and Kleinberger study, are aware of their own reasoning that resolves in a numerical estimate, but unaware of the reasoning behind any solely numeric estimate that any advisor(s) may offer. Hence, it is possible to envisage an empirical study that would extend understanding of the mechanisms of judgment change within the parameters of the JAS paradigm. Such a study will examine the influence of reasons in addition to numerical advice, upon egocentric discounting of advice, accuracy and confidence.

The present study also points to areas where the structured exchange of information may aid estimation performance. Participants in the present study clearly preferred reasons advice, in preference to solely numeric advice. Investigation of this finding could fruitfully be extended, by examining the circumstances in which further structuring information exchange, for both large and small highly uncertain quantities, may be beneficial, in terms of estimation performance. Such a study would both extend the parameters of the JAS paradigm, and build upon the existing findings of authors who have investigated the 'divide and conquer' principle, or 'algorithmic decomposition' for uncertain quantities.

Finally, the findings of the present study highlight the constraints upon human cognition, when people are tasked with differentiating between beneficial and poor advice, under conditions of

incomplete domain knowledge. This finding will be further tested in an empirical study, where both beneficial and poor numerical advice, is presented to participants under circumstances of incomplete domain knowledge.

Chapter 5: The Benefit of an Additional judgment.

5.1 Introduction

Seeking or using advice from a friend or colleague in a personal, work or social context seems very much in the realm of the commonplace. However, studies exploring the possible effects of advice upon decision-making are relatively few and largely concentrated in the Judge-Advisor Systems (JAS) stream of research (Sniezek, 1992; Sniezek and Buckley, 1995; Budescu and Rantilla, 2000; Budescu, Rantilla and Yu, 2003). Chapter 4 demonstrated that people tend to discount advice in favour of their own judgment; but where participants chose to revise their judgment post-advice, (where advice was constituted as both solely numeric, and reasons-based), participants tended to conform to advice, in an estimation task concerning the date of notable historical events. Further, participants overwhelmingly preferred reasons-based advice, when they chose to revise their judgment post-advice. The current chapter will focus upon the cognitive weighting strategies people employ, when deliberating upon solely numeric or reasons-based advice. It follows then, that if people persist in conforming to advice (irrespective of the quality of advice) over a range of tasks, they will place equal cognitive weight, upon either solely numeric advice, or reasons-based advice. Further, this study tests whether participants are able to adequately distinguish between potentially beneficial advice, and poor advice. Should people fail to distinguish between potentially beneficial advice, and poor advice, then this would lend further support to the idea that where people revise their judgment post-advice, it is likely that they are indiscriminately conforming to advice irrespective of its quality. A decision-making dyad consisting of an advisor and a judge is the simplest form that such a model takes, and is what is of concern here; the influence of an additional judgment, in terms of beneficial judgment revision.

Judgment Change and Egocentric discounting

Perhaps the most obvious way in which the benefit of an additional judgment can be demonstrated is where a judge changes his/her initial judgment following the receipt of advice - perhaps leading to a more accurate, or more beneficial final estimate. This implies for reductions in the mean error of judges' post-advice judgment, *judgment change* is necessary. Judgment change can be described as an instance where a judge gives an estimate that is different, post-advice, to the response made pre-advice. It follows that people who change judgment after receipt of advice are likely to change judgment in the direction of the available advice, on average. Further, if people are able to evaluate the quality of proffered advice, then such judgment change will also be in the direction of the correct answer. Whilst the receipt of advice may lead to improved estimation accuracy - on average - judgment change will be the process through which such improvements are realized. A further psychological mechanism, which may underlie a judge's degree of judgment change, is the weight that a judge puts upon advice – particularly in relation to his/her initial judgment.

Yaniv and Kleinberger (2000), and Yaniv (2004a; 2004b), have developed a measure delineating the Weight of Advice (WOA) to examine the degree to which judges revise their judgments post-advice. They argue that a final estimate is the product of a weighted combination of a judge's *initial estimate* and *advice*, and can be defined as $|f - i| / |a - i|$, where *i*, *f*, and *a*, represent initial, final and advice respectively. Further, it is argued that the weight of advice is well defined if the final estimate falls between the initial estimate and the advice. Expressed as a proportion, WOA reflects the weight that a participant places upon advice (it is inversely related to the extent that the advice is discounted). Hence, advice weighted as '0' indicates that a participant has completely discounted the advice and is fully convinced of the veracity of their own judgment, whereas a WOA of '1' indicates that a participant has

jettisoned their own initial judgment and entirely adopted the advice available to them in determining a final estimate. A WOA of 0.5 would indicate that both the advice and a participant's initial judgment were weighted equally. Yaniv *et al.* argue that people are likely to egocentrically discount advice in favour of their own judgment. This conceptualization has two possible explanations. The first, a cognitive explanation, argues that integrating advice into a global determination of some issue imposes a greater cognitive load than assessing the quality of the advice proffered (Harvey, Harries and Fischer, 2000). Hence, advice may be recognized as useful but may be under-utilized. A second explanation explores the idea that advice is egocentrically discounted relative to a judge's initial judgment (Yaniv and Kleinberger (2000); Yaniv 2004a, 2004b). This second task specific explanation, notes that in a situation where a judge is only able to access numerical advice from an advisor, the reasoning behind such advice will be unavailable. Hence, a judge's own initial judgment will dominate any numerical advice.

The findings from the JAS avenue of research do not, as yet, differentiate between the preceding two accounts, but simply support the notion that whilst advice is influential – and there is judgment change in the judge's estimates post-advice - judges grant their own judgments greater weight in the decision or forecasting process (Harvey *et al.* 2000; Yaniv and Kleinberger, 2000). Why should this be so? One explanation is that one's own judgment implicitly contains the reasons for a stated numerical judgment whereas, to date, advice in the JAS paradigm has been simply numerical estimates. Such numerical advice from advisors does not contain the reasons underlying the advice. Furthermore, numerical advice consistent with a judge's own judgment or forecast is likely to be given greater cognitive weight than discrepant advice (Harvey *et al.* 2000). Studies that have examined advice quality also highlight the sensitivity that judges exhibit towards, what is presumed to be, accurate advice.

Advisors that have a history of accuracy (or have more information at their disposal than judges making a final determination), are likely to be given more credence than advice proffered by, what is perceived, as a less accurate advisor (Budescu *et al.* 2003). However, reason-based advice has not been utilized in the extant JAS studies, and it is unknown whether such advice will be given more weight – relative to pure numerical advice, in the judge's post-advice judgment. The issue of what constitutes reason-based qualitative advice (in addition to purely numerical information) compared to other forms of qualitative advice is of significance here; it is contended that reason-based qualitative advice may convey not only information that a judge may be unaware of, but also a logic or rationale of how such information supports a numerical estimate received as advice (see Appendix II). In short, if reasons are attached to numerical advice, the effects of egocentric discounting should be mitigated, as judges would have access to the advisors' reasoning. If egocentric discounting is lessened then it will support the task specific account. Conversely, if there were an obtained persistence of egocentric discounting - despite the provisions of reasons in addition to numerical advice – it would lend support to the cognitive account of sub-optimal advice use. Utilizing the WOA measure allows some insight into how participants regard the advice available to them in relation to their own judgment given numerical advice with additional supporting reasons.

> H_1: Participants with purely numerical advice available to them will egocentrically discount that advice in favour of their own judgment. Participants with reason-based advice and numerical information available to them will egocentrically discount the advice to a lesser extent (than participants receiving purely numerical advice).

Advice and accuracy

Advice in the context of JAS has been almost exclusively numerical in nature (means/medians), and largely aimed towards the accuracy of forecasts extrapolated from time series data (Sniezek, 1989, 1990; Sniezek and Buckley, 1995; Harvey, 1995; Harvey *et al.*

2000). Where non-numerical advice has been utilized it has been of a superficial nature – qualitative 'advice' consisting of comments such as 'this is a guess' or 'I'm definite about this' (Sniezek and Van Swol, 2001). Hence, the potential beneficial effects of improved estimation facilitated by the introduction of reasons in addition to numerical advice is unexplored in the JAS literature. This is unfortunate, since one of the most obvious motives individuals may have for using advice is the desire to improve the quality of the decision (Harvey and Fischer, 1997). Intuitively, advice that is an elaboration of numerical estimates will contain important qualitative information (e.g. "I think that man will land on Mars in 2020. This is because my father works for NASA and I have seen their plans"). In the present study, I focus on an evaluation of the benefit of an additional judgment through advice available to judges from a single anonymous advisor. The advice is either (i) numerical advice or (ii) reasons-based advice. Further, the provision of qualitative reasons in addition to purely numerical advice may allow judges to discriminate between useful and spurious numerical advice, resulting in greater accuracy. One way to examine these assertions is to look at the error that an individual participant makes on the experimental task. Such errors can be averaged across participants and conditions (Mean Absolute Error). Subtracting the final MAE from the initial MAE is indicative of the degree accuracy that participants achieve. Alternatively, should participants revise their judgment and merely conform to *any* available advice, then the potential estimation performance improvements of attending to beneficial advice, may be negated by the decreases in estimation accuracy brought about by attending to poor advice. In this scenario, there will be little difference between the post-advice accuracy of participants in receipt of solely numeric advice, and participants in receipt of reasons-based advice.

This discussion leads to the hypothesis that –

> H_2: There will be significant differences (in MAE scores) between the condition where participants have numerical advice available to them, and the condition where participants have reasons in addition to numerical advice available to them. Specifically, participants who receive reasons in addition to numerical advice will improve their estimate (i.e. reduce their error) to a greater degree.

5.2 Method

Participants

Participants in the study (N = 256) were undergraduate and postgraduate students at Durham Business School from both the Durham and Queen's Campus (Stockton). Participation in the experiment was voluntary and no payments were made. Participants were students who the author recruited by visiting a lecture theatre (with the permission of staff colleagues) where students were gathered to attend their studies. All participants took approximately the same amount of time to complete the questionnaire.

Stimuli and materials

Following the work of Yaniv and Kleinberger (2000), and Yaniv (2004a; 2004b), I developed a 14-item questionnaire (see Appendix II) that tasked participants to ascertain the correct date - or a best estimate of the correct date - of historical events that occurred in the last 300 years. In addition, I derived two conditions under which advice would be available to participants: first, numerical advice (accurate/inaccurate), second, numerical advice and supporting additional reasons (very weak/weak/strong) (see Appendix II).

The problem of defining what is the essence of a compelling message or argument is a debate

that is longstanding and ongoing in social psychology. This is unfortunate for the purposes of the present study, as one of the objectives of the research was to examine the differential effects of reasons feedback/advice (i.e. reasons at different degrees of persuasiveness) upon accuracy, and judgment change. One way to approach this issue might be to adopt the perspective of the Elaborative Likelihood Model theory (ELM) advocated by Petty and Cacioppo (1986). Here, the experimenter (or pilot group, if practical) generates convincing and specious arguments about the question(s) of interest. These reasons are then independently rated on a 5-point Likert scale of *persuasiveness* (1 = unconvincing; 5 = compelling). In ELM studies of *message persuasiveness* these ratings would then be passed to a second rating group who would write down how favourable or unfavourable these ratings are to the advocated reasons (thought listing). i.e. very weak/weak/strong. As reasons are of interest and not persuasive messages I only retain the first step of Petty and Cacioppo's methodology; subsequently the reasons generated by the experimenter were categorized as 'very weak', 'weak', and 'strong' by two independent coders. These categories serve as methodological proxies for the argumentative quality of the generated reasons.

Procedure

Participants were randomly allocated to one of three experimental conditions (i) a control condition (where no advice was available to participants), (ii) a numerical advice condition (where only numerical advice is available to participants) (iii) a reasons-based advice condition (where reasons in addition to numerical advice is available to participants). Numerical advice was further delineated as accurate or inaccurate (participants received either accurate or inaccurate advice– not both - for each of the 14 items). This permutation of advice was held constant in the reasons-based and numerical advice condition, so that accurate numerical advice available to participants in the accurate numerical advice condition was identical to the accurate numerical advice in the accurate numerical advice and supporting reasons condition.

Exactly the same permutation of advice was maintained for inaccurate numerical advice, and reasons-based advice with inaccurate numerical information. The correct date for a specific historical event constituted accurate numerical advice. Inaccurate numerical advice varied the degree of inaccuracy from question to question, but was approximately balanced in terms of the sign of inaccuracy around the correct value (+ years or – years) (see Appendix II). Participants were given no information regarding the identity of the advisor beyond the description, 'advisor *C*', or 'advisor *P*'. For each of the fourteen questionnaire items in the two advice conditions the alphabetic designation of the advisor was varied (creating the illusion of fourteen different advisors).

In the reasons-based advice condition, the quality of reason-based advice was manipulated so that one third of participants received very weak supporting reasons; one third received weak supporting reasons, and one third received strong supporting reasons as advice, after making an initial estimate, and prior to making a final estimate.

In the control condition, participants estimated the date of particular historical events, and were then asked to reconsider their estimates after a short period of reflection before moving to the next question. Across conditions, all participants took approximately the same amount of time to complete the task.

The questionnaire items were counterbalanced across participants and conditions in order to offset any systematic presentation order effects of the questionnaire items. The experimental task was divided into two phases; initially participants were tasked with estimating the correct date (or best estimate) for a questionnaire item; secondly, participants were tasked with re-

considering their estimate in the light of available advice. Each phase of the questionnaire was printed on a separate sheet of the questionnaire booklet; so that advice was not available to participants in phase one. Participants were instructed to attempt each of the 14 questions sequentially, so that on completion of the first experimental phase participant's proceeded to the second phase, where a final estimate was made in the light of the available advice. An outline of the general procedure is shown below –

REASONS ADVICE CONDITION

Phase I (series of 14 questions)

In what year was the Suez Canal first opened for use?

Your best estimate _____

Please give full details of the reasons that have lead you to this estimate.

Phase II (same 14 questions)

In what year was the Suez Canal first opened for use?

Your previous best estimate was 1905

The best estimate of advisor *K* was 1830

Advisor *K* bases the estimate upon the notion that the Suez Canal was first opened for use during the 19th century colonial era to facilitate European trade between the Mediterranean and the Middle East.

Your final best estimate _____

5.3 Results

Prior to analyzing the data, response rates were considered to determine if missing data might influence the outcome of statistical analysis. Inspection of Table 5.1 shows that whilst response rates were not uniformly 100%, the responses rate reached 91.48% on average. This seemed high enough to proceed to analyze the data.

Table 5.1 Number of responses

Participants *N* = 255	Unaided estimate	Pre-advice Lower Bound	Pre-advice Upper Bound	Post-advice estimate	Post-advice Lower Bound	Post-advice Upper Bound
Responses *N*	3359/3570 (94.01%)	3266/3570 (91.48%)	3277/3570 (91.79%)	3368/3570 (94.34%)	3164/3570 (88.63%)	3164/3570 (88.63%)

Judgment Change and Egocentric discounting

The purpose of the results section which follows is to report data that addresses H_1 and H_2, and to ascertain whether the observed behavior of participants is consistent with the idea that where people revise their judgment post-advice, they do so by conforming to any available advice. First, a number of manipulation checks are reported that confirm that the experimental manipulations had an effect upon dependent variables. Second, the analysis examines the amount of judgment change reported by participants (i.e. the number of judgment revisions), before considering the cognitive weight participants placed upon advice (the WOA measure). Finally, the appropriateness of judgment revision is considered, in order to ascertain whether participants were able to distinguish between potentially beneficial advice and spurious advice.

Does attending to advice influence judgment change?

That the manipulations had some effect beyond that that might be expected simply by people 'thinking again' is demonstrated by the finding that there was no significant decrease in the mean absolute error that participants in the no advice condition made between their initial estimate ($\bar{x}$ = 44.51, *sd* = 16.36 years), and their final estimate ($\bar{x}$ = 43.82, *sd* = 16.28), indicating that 'thinking again' is insufficient to improve accuracy in the experimental task (t = 1.29, df = 29; p = 0.21). Further, by comparing the average mean differences between participant's initial and final estimates in the experimental conditions, with the average mean difference between participant's initial and final estimates in the no advice (control) condition, I was able to establish that the manipulations had had an effect upon the accuracy of participants' estimates ($F_{(8, 246)}$ = 4.62, p = 0.0005).

Table 5.2 Bonferroni comparison of error reductions (years) between control and experimental conditions

Advice	*N*	Initial estimate $\bar{x}$	*sd*	Final estimate $\bar{x}$	*sd*	Error reduction $\bar{x}$	*sd*	Mean difference	Std. Err	*p*
Control (no advice)	30	44.51	16.36	43.82	16.28	0.68	2.90			
Accurate numerical	24	28.91	16.00	10.84	7.76	18.07	15.09	-17.39	3.73	0.0005*
Inaccurate numerical	26	35.66	13.99	20.05	7.75	15.61	14.66	-14.93	3.65	0.002*
Accurate numerical and v.weak reasons	27	27.22	11.28	16.93	9.18	10.29	10.31	-9.61	3.61	0.30
Inaccurate numerical and v.weak reasons	32	31.61	15.45	22.35	12.62	9.26	11.60	-8.58	3.46	0.50
Accurate numerical and weak reasons	33	37.03	25.04	20.20	18.57	16.83	16.40	-16.15	3.43	0.0005*
Inaccurate numerical and weak reasons	25	34.04	16.85	20.23	8.96	13.82	13.29	-13.13	3.69	0.02*
Accurate numerical and strong reasons	24	33.95	17.21	13.92	12.27	20.03	16.35	-19.35	3.73	0.0005*
Inaccurate numerical and strong reasons	34	35.26	18.51	21.99	12.87	13.28	16.33	-12.59	3.41	0.01*

From the table above, it can be seen that apart from numerical advice with very weak supporting reasons, post-hoc Bonferroni tests reveal that all the experimental conditions significantly differ from the control condition (in terms of the reduction in error between the initial and final estimates), which supports the idea that the manipulations were successful. A further way of examining the reduction in error that participants made from the initial estimate (e1) to the final estimate (e2) is to look at the average error question by question. Collapsing the data across participants revealed a significantly reduced mean error on all of the 14 questions, post-advice – indicating the lack of a 'ceiling effect' due to the specific characteristics of the experimental task. Table 5.2 also shows that participants in receipt of inaccurate numerical advice were initially even *more* inaccurate than the advice on average. It should be no surprise therefore, that judgment revision in the direction of advice would benefit these participants. However, of interest to the present discussion is whether participants in receipt of accurate numerical advice were able to reduce their error post-advice, to a greater extent than participants in receipt of inaccurate numerical advice – irrespective of accompanying reasons. To ascertain whether this was so, a one-way ANOVA was performed on the error reduction scores of participants in receipt of advice. If people were sensitive to the asymmetries of the quality advice, then it might be expected that significant differences would be detected between conditions in terms of the extent to which participants were able to reduce their pre-advice error, post-advice. However, no significant differences between conditions in terms of error reduction scores were found $F_{(7,\ 217)} = 381.45$, $p = 0.08$). Hence, it is parsimonious to suppose that participants' estimates of historical events became more accurate (the average error decreased) from the initial estimate to the final estimate.

Another way that these data might be considered is by examining the average error participants made pre-advice with the average error that advisors committed in each experimental condition

(see Table 5.3). Examining the data in this way, reveals that participant's pre-advice estimates were more inaccurate than advice on average – hence judgment revision in the direction of advice *should* improve estimation accuracy, on average.

Table 5.3 Mean Pre-advice, and advisor error (years)

Participants (N = 255)	N = number of valid responses	Pre-advice error		Advisor error	
		$\bar{x}$	*sd*	$\bar{x}$	*sd*
Control	30	45	16		
Accurate statistical	24	29	16	0	0
Inaccurate statistical	26	36	14	14	1
Accurate statistical & v. weak reasons	27	27	11	0	0
Inaccurate statistical & v. weak reasons	32	32	15	14	2
Accurate statistical & weak reasons	33	37	25	0	0
Inaccurate statistical & weak reasons	25	34	17	15	0
Accurate statistical & strong reasons	24	34	17	0	0
Inaccurate statistical & strong reasons	34	35	19	15	1

Next, the *amount* of judgment change reported by participants was examined. Initially, I looked at the *amount* of judgment change in our analysis. Participants were able to change judgment on each of the 14 questions. Hence, the number of potential judgment changes per participant could range from '0' (an individual participant did not change judgment from their original estimate) to '14' (an individual participant changed judgment on all of the questionnaire items). The mean amount of judgment change is defined by the number of

judgment changes per participant divided by the number of valid responses for each participant. This score was subsequently averaged across conditions and the results are shown below (as some participants did not respond to all 14 of the questionnaire items these scores are presented as a score between 0 and 1). The table below shows that judges in receipt of advice change judgment approximately 50% of the time, whilst judges without advice change judgment approximately 20% of the time. A one-way between-subjects ANOVA confirmed that there are significant differences in the mean amount of judgment change between conditions ($F_{(2, 252)} = 11.84$, $p = 0.0001$). Employing the Bonferrroni post-hoc test, significant differences were found between the no advice (control) condition and (i) the numerical advice condition ($p < 0.0001$) and (ii) the reasons-based advice condition ($p < 0.0001$). However there was no significant difference in the mean amount of judgment change between participants in receipt of purely numerical advice, and participants in receipt of reasons-based advice ($p = 1$).

Table 5.4 Mean Amount of judgment change

	N	Mean $\bar{x}$	*sd*
Control	29	0.22	0.22
Numerical advice	50	0.49	0.27
Reasons-based advice	175	0.48	0.29

If participants in the reasons-based advice condition were to discount advice to a *lesser* extent than participants in the purely numerical advice condition, then it may follow that participants in the reasons-based advice condition are likely to have changed judgment *more* often than participants in receipt of purely numerical advice. As such, these findings do not support

hypothesis H_1 that participants in receipt of purely numerical advice egocentrically discount that advice to a greater degree than participants in receipt of reasons-based advice. Table 5.4 shows that participants in the purely numerical advice condition do not change judgment, any more or less, than participants in the reasons-based advice condition, on average. Whilst this finding does not lend support to H_1, it is consistent with the idea that participants who revise their judgment post-advice, indiscriminately conform to any available advice.

I next examined the *extent* of judgment change and egocentric discounting utilizing the WOA measure. Yaniv (2004a) argues, in the task specific account, that one of the underlying conditions for the occurrence of egocentric discounting is the inability of a judge to access the reasons behind an advisor's numerical estimate. Hence, in a situation where reasons are attached to numerical estimates egocentric discounting should dissipate. WOA is considered well defined if the values fall between 0 and 1 (i.e. if a participant's final estimate fell somewhere between his/her initial judgment and the advice). This was the case for 95% of the participants in advice conditions in this study; the remaining 5% of participants made estimates that were greater than the advice (e.g. if the initial estimate was 1900 the advice was 1940, but the final estimate was 1960). The overall mean WOA was 0.46 (*sd* = 0.31); by comparing the overall WOA to 0.5 (equal weighting of advice and a participant's own judgment) I found that participants discounted advice and preferred their own judgment (t = -1.92, df = 224, p = 0.03, 1 tailed). The extent of egocentric discounting reported here is far *less* than that of Yaniv and Kleinberger (2000), (WOA = 0.29), and Yaniv (2004), (WOA = 0.27). The difference in the extent of egocentric discounting between that found in the current study, and that found in previous research, is an issue taken up in the discussion section. Returning to the current study, there were no significant differences in the mean WOA of participants receiving reasons in addition to numerical advice ($\bar{x}$ = 0.46, *sd* = 0.32), and participants receiving numerical

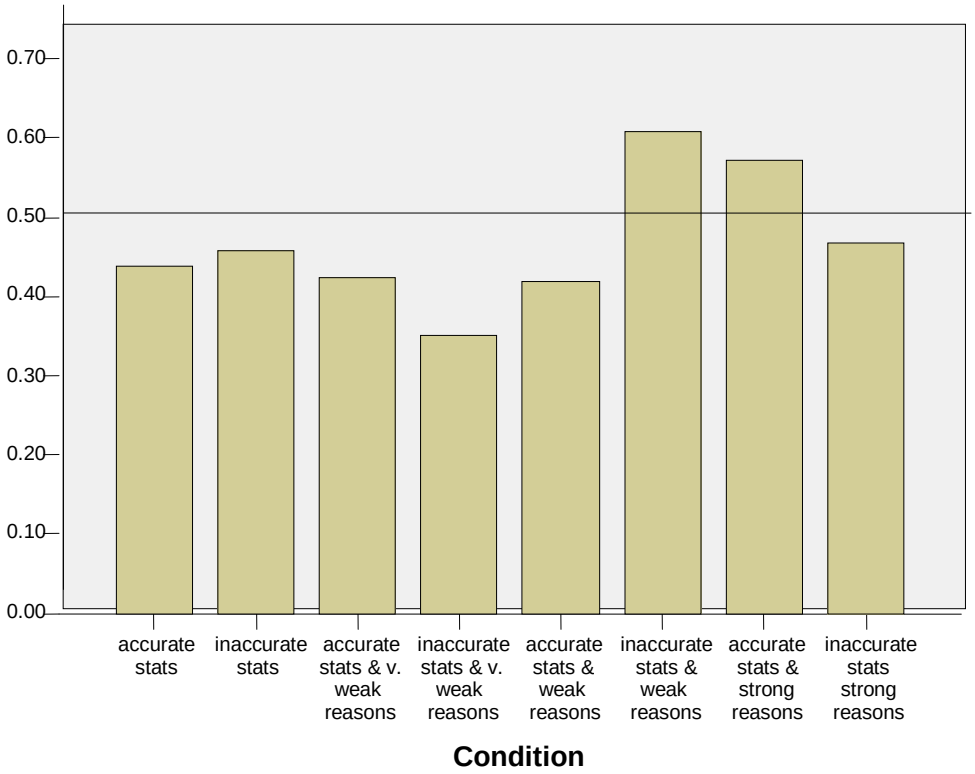
advice only ($\bar{x}$ = 0.45, *sd* = 0.28), (*t* = -0.334, *df* = 223, *p* = 0.37, 1 tailed). Comparing the mean WOA of participants in receipt of purely numerical advice, with 0.5 (equal weighting of advice and a participant's own judgment), revealed that participants did not discount the available advice in favour of their own judgment (*t* = -1.33, *df* = 49, *p* = 0.095, 1 tailed). Similarly, participants in receipt of reasons-based advice also did not choose to ignore the advice available to them (*t* = -1.494, *df* = 174, *p* = 0.07, 1 tailed). The table below compares the mean WOAs of the advice conditions against the scenario where people give equal weight to both their own, and an advisor's judgment (0.5 weighting).

Table 5.5 Mean WOA scores compared to 0.5 (equal weighting of own judgment and advice).

Advice	*N*	Mean $\bar{x}$	*sd*	*t*	*df*	*p* (1 tailed)
Numerical	50	0.45	0.28	-1.33	49	0.09
Numerical and additional reasons	175	0.46	0.32	-1.49	174	0.07
Accurate numerical	24	0.44	0.30	-1.03	23	0.16
Inaccurate numerical	26	0.46	0.27	-0.82	25	0.29
Accurate numerical and very weak reasons	27	0.42	0.31	-1.29	26	0.10
Inaccurate numerical and very weak reasons	32	0.35	0.25	-3.39	31	0.001*
Accurate numerical and weak reasons	33	0.42	0.29	-1.61	32	0.059
Inaccurate numerical and weak reasons	25	0.61	0.35	1.56	24	0.07
Accurate numerical and strong reasons	24	0.57	0.31	1.15	23	0.13
Inaccurate numerical and strong reasons	34	0.47	0.35	-0.55	33	0.29

The preceding analysis suggests little support for H_1 in that participants in the numerical advice condition and participants in the reasons-based advice condition had nearly identical mean WOAs. Whilst participants did significantly discount advice in favour of their own judgment on average, the introduction of additional reasons to numerical advice did not diminish egocentric discounting anymore than participants in receipt of purely numerical advice, contrary to the position outlined by Yaniv *et al.* (2000, 2004a, 2004b). Recall that Yaniv (2004) reported that his participants discounted purely numerical advice by a WOA of 0.27. Indeed, plotting WOA by the experimental conditions (Fig. 5.1) reveals that only participants in the numerical advice with very weak supporting reasons condition, significantly discounted the advice (compared to 0.5 weighting of a participant's own judgment and the advice).

Figure 5.1 Mean WOA by condition

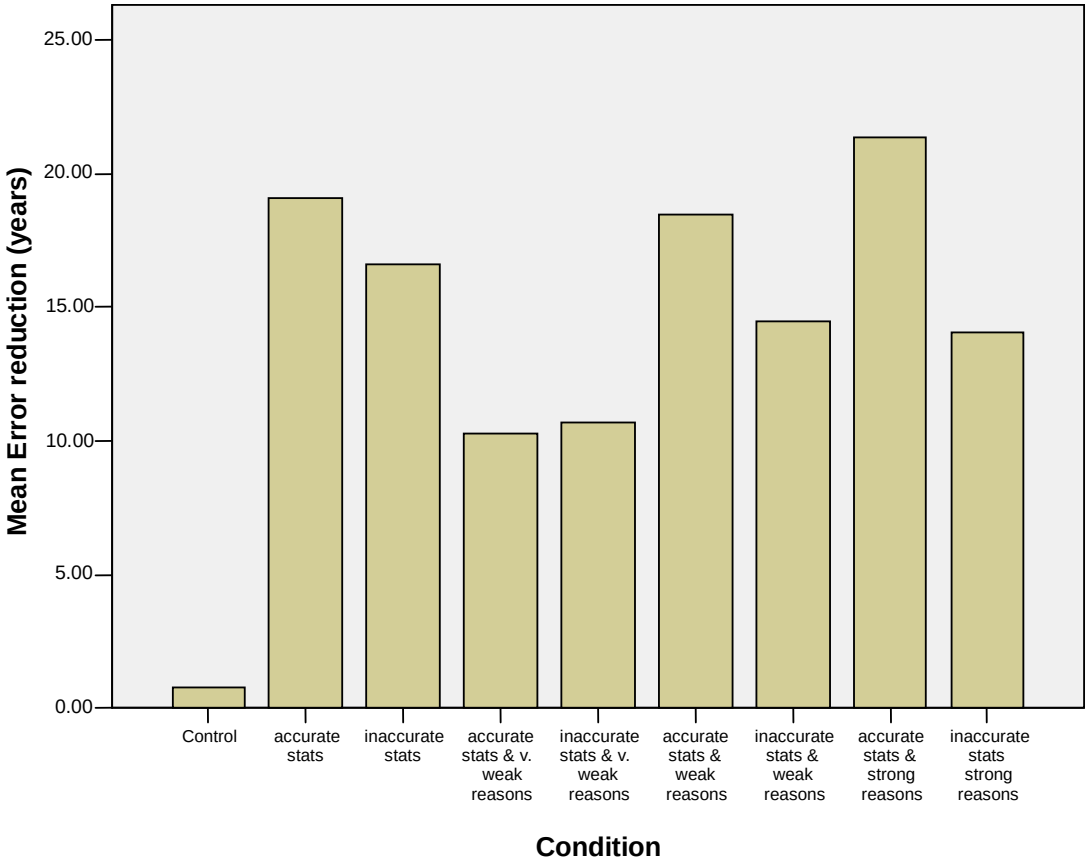


Whilst these results do not provide strong support for H_1 they are consistent with the idea that participants place equal cognitive weight upon either solely numeric or reasons-based advice. Hence, irrespective of the type of advice available to participants, where an initial judgment is revised, the advice is equally influential in participant's deliberations' – which is supportive of the conformity to advice account.

Advice and accuracy

I next examined the *appropriateness* of judgment change. On average, those participants who changed judgment did so in the direction of the correct answer. The mean absolute error was calculated for participants who changed their judgment (N = 235) at their initial estimate ($\bar{x}$ = 35.46 years, sd = 17.87 years), and their final estimate ($\bar{x}$ = 20.59 years, sd = 14.64 years). Similar results were obtained by examining those conditions where participants were in receipt of advice (N = 210); estimation error was reduced from ($\bar{x}$ = 34.45 years, sd = 17.77 years), at the initial estimation phase, to ($\bar{x}$ = 17.92 years, sd = 11.82 years) for the final estimate. A one-way within-subjects repeated measures ANOVA showed that there was a significant beneficial effect of judgment change ($F_{(1, 202)}$ = 266.21, p = 0.0005). This finding supports H_2 in so far as of those participants who changed their judgment, advice (irrespective of its quality) improved the accuracy of participants' estimation. However, there is little support for H_2 when the appropriateness of judgment change is examined by condition. There is no significant difference in beneficial judgment change between participants in the reasons-based advice condition (N = 162, $\bar{x}$ = 15.93, sd = 14.91 years), and participants in receipt of purely numerical advice (N = 48, $\bar{x}$ = 18.53, sd = 14.58), (t = 1.07, df = 208, p = 0.29). Figure 5.2 below shows that my participant's accuracy improved over iterations.

Figure 5.2. Error reduction (years) post-advice



Further support for the idea that participants were constrained in their abilities to distinguish between potentially beneficial advice, and poor advice (and where judgments were revised, tended to conform ultimately to any advice) is found when the appropriateness of judgment revision is examined. Taking average error as a dependent variable, it appears that the provision of reasons by a single anonymous advisor in addition to numerical advice, offers little in the way of greater diagnosticity in terms of accurate estimation. Whilst reasons in addition to numerical information do not appear to have facilitated enhanced accuracy in estimation, it is clear that participants were unable to discern the difference between accurate and inaccurate numerical advice. If participants were able to make such a distinction, no difference in the average error post-advice of the two groups would be expected. Instead it appears that participants receiving inaccurate numerical advice made nearly double the mean

average error ($\bar{x}$ = 20.05 years, *sd* = 7.75), than participants who received accurate numerical advice (mean error of $\bar{x}$ = 10.84 years, *sd* = 7.76) - a statistically significant difference (*t* = -4.19, *df* = 48, *p* = 0.0005). Similarly, participants in receipt of numerical advice (accurate/inaccurate), accompanied by supporting reasons were unable to distinguish between beneficial or spurious advice. Participants in receipt of accurate reasons-based advice were more accurate (*N* = 108, $\bar{x}$ = 15.78 years, *sd* = 13.49) on average, than participants in receipt of inaccurate reasons-based advice (*N* = 117, $\bar{x}$ = 21.27 years, *sd* = 10.94). This indicates that, on average, participants were unable to identify spurious advice; participants in receipt of inaccurate numerical advice and reasons-based advice, were less accurate than participants in receipt of accurate numerical and additional reasons-based advice (*t* = -3.37, *df* = 223, *p* = 0.001).

Logically, for advice to be beneficial overall, then the 'pull' of accurate advice (i.e. judgment change in the direction of the true answer), must be greater than the pull of inaccurate advice in the direction away from the true answer. As numerical accurate advice is the same as the true answer in the current study, the only scenario of interest that delineates between the pull of advice and the pull of the true answer, is where an individual's initial estimate falls between the true answer and the advice (i.e. initial estimate is 1880, the true answer is 1900, but the advice is 1860). I extracted those cases from the data that were consistent with this scenario, and looked at whether participants moved from their initial estimate towards the true answer, or from their initial answer towards the advice (see Table 5.6). Support for the cognitive account of egocentric discounting might be forthcoming should participants not become more inaccurate post-advice despite the availability of inaccurate advice. This is because this account is based upon the idea that the quality, or accuracy of advice, is of secondary importance to the potential benefits available to participants through deeper cognitive

processing of a particular problem and its relevant attributes (Brehmer and Hagafors, 1986). However, if participants merely conform to any available advice, then attending to poor advice must result in decreases in estimation accuracy post-advice, according to this account. Table 5.6 shows the number of cases extracted from the dataset where a participant's estimate (N = 117) fell between the true (but unknown) answer and the advice. The number of responses participants made in response to each question (that fulfilled the criteria that a participant's estimate fell between the true answer and advice) was tabulated in Table 5.6.

Table 5.6 Pre-advice error and post-advice error: Increases in error indicate movement towards advice.

Participants (N = 117)	N = number of valid responses	$\bar{x}$	*sd*	*t*	*df*	*p*
Question 1 – pre-advice error	32	7.84	5.66	-3.455	31	.002*
Question 1 - post-advice error	32	11.19	6.727			
Question 2 – pre-advice error	52	1.71	2.43	- 2.41	51	0.02*
Question 2 – post-advice error	52	2.46	3.32			
Question 3 – pre-advice error	33	3.18	2.34	- 1.77	32	0.09
Question 3 – post-advice error	33	4.91	5.80			
Question 4 – pre-advice error	28	4.96	2.10	- 2.21	27	0.04*
Question 4 – post-advice error	28	5.71				
Question 5 – pre-advice error	29	15.72	7.53	- 3.54	28	0.001*
Question 5 – post-advice error	29	19.41	8.25			
Question 6 – pre-advice error	43	24.05	12.31	- 3.83	42	0.0005*
Question 6 – post-advice error	43	30.05	12.11			
Question 7 – pre-advice error	21	3.67	2.42	- 1.45	20	0.16
Question 7 – post-advice error	21	4.14	2.33			
Question 8 – pre-advice error	35	2.49	1.34	- 2.00	34	0.05
Question 8 – post-advice error	35	2.80	1.28			
Question 9 – pre-advice error	21	5.62	6.05	- 2.11	20	0.05*
Question 9 – post-advice error	21	7.90	6.95			
Question 10 – pre-advice error	24	2.58	1.77	- 1.89	23	0.07
Question 10 – post-advice error	24	3.67	3.12			
Question 11 – pre-advice error	19	10.63	4.06	- 1.69	18	0.11
Question 11 – post-advice error	19	11.79	4.12			
Question 12 – pre-advice error	24	6.25	2.61	- 3.23	23	0.004*
Question 12 – post-advice error	24	7.92	2.69			
Question 13 – pre-advice error	21	5.43	5.64	- 3.20	20	0.005*
Question 13 – post-advice error	21	9.19	6.88			
Question 14 – pre-advice error	17	27.94	8.79	1.75	16	0.10
Question 14 – post-advice error	17	24.53	6.09			

From the preceding table, it is clear that of the 14 questions answered by my respondents, 13 questions show increases in mean error. Of the 13 questions showing increases in error, eight such increases appear significant. The participants chose to follow inaccurate advice that took them further away from the true answer than their initial estimate, in 8 out of 14 questions (57%). Further, of the 117 participants whose initial estimate fell between the inaccurate advice and the true answer, 65 of them revised their judgment in the direction of the advice to the detriment of their final accuracy. Put another way of the possible 1638 observed changes in error (117 participants x 14 questions), 910 observations resulted in decreases in accuracy (55.56%).

Table 5.7 Mean absolute error (MAE) for participants whose initial estimate fell between inaccurate advice and the true answer.

Where participants pre-advice estimates fell between advice and the true answer, participants in the conditions below followed advice on average – to the detriment of post-advice accuracy (i.e. post-advice error increased)							
Condition	Error pre-advice	*sd*	Error post-advice	*sd*	*t*	*df*	*p*
Inaccurate numerical advice (*N* = 12)	4.63	3.77	9.29	4.90	-5.25	11	0.0005*
Inaccurate numerical advice and very weak reasons (*N* = 18)	8.47	6.02	18.38	11.30	-4.66	17	0.0005*
Inaccurate numerical advice and weak reasons (*N* = 14)	9.21	7.56	21.06	10.28	-6.68	13	0.0005*
Inaccurate numerical advice and strong reasons (*N* = 21)	9.14	9.35	16.48	9.18	-6.75	20	0.0005*

The above table shows that irrespective of the reasons attached to inaccurate numerical advice, participants made significantly greater error post-advice compared to pre-advice. A one-way between subjects ANOVA revealed that there were differences between the conditions in terms

MAE post advice ($F_{(3, 61)} = 3.614$, $p = 0.018$). Employing the post-hoc Bonferroni method revealed that MAE of inaccurate numerical advice was significantly different from inaccurate very weak reasons-based advice ($p = 0.015$), and inaccurate weak reasons-based advice ($p = 0.015$) – post-advice. I also collapsed our data across conditions to examine if there were any differences between purely numerical advice and reasons-based advice. This was non-significant ($F_{(1, 115)} = 0.832$, $p = 0.48$). These findings indicate that the cognitive activation explanation of egocentric discounting can be dismissed, as when participants are allowed to revise their judgment in the light of advice, inaccurate advice can lead to decreases in estimation accuracy. However, this finding is consistent with the conformity to advice account of the data.

5.4 Discussion

At the outset of this study, it was contended that should participants indiscriminately conform to advice, post-advice, then they would simultaneously place equal cognitive weighting upon solely numeric advice, or reasons-based advice. Further, I posited two competing explanations of egocentric discounting – a task specific account, and a cognitive activation account. The data reported here does not fully support either of these two accounts, rather it points towards a third explanation – indiscriminate conformity to advice (to a degree), on the basis of 'anchor and adjust' (Tversky and Kahneman, 1974). The task specific account of egocentric discounting was tested by my initial hypothesis, that egocentric discounting would dissipate, where reasons-based advice was provided to judges, in addition numerical information. I found that in terms of the weight our participants put upon advice (WOA), that there was no significant difference between participants in receipt of reasons-based advice, and participants in receipt of purely numerical advice. This undermines the case for a task specific account of egocentric discounting, as it was argued that egocentric discounting occurs because people are

unable to access an advisor's reasoning - where advice is purely numerical. Hence, judges are more likely to discount advice, and rely upon their own judgment in making an estimate. Participants had access to the advisor's reasoning, yet did not weight advice differently to participants in receipt of purely numerical advice.

Such behaviour would not be inconsistent with the cognitive activation account of egocentric discounting - any advice is recognized as potentially beneficial, but under-utilized due to the cognitive load implied by integrating the advice into a revised estimate. Further support for this explanation is found in the accuracy data, where participants became more accurate post-advice irrespective of the quality of the advice available to them. However, for the cognitive activation account to be accepted, accuracy would have to improve, even where a participant's initial estimate falls between the true answer and the available advice. On examining cases where a participant's initial estimate fell between the true answer and advice, estimation accuracy decreased significantly in eight out of the fourteen questionnaire items. This would suggest that participants revise their estimates by conforming to advice by 'anchoring' upon the available advice and cognitively pulling themselves some way towards it from their initial position - to the detriment of estimation accuracy. The apparent contradiction between the finding that accuracy improved irrespective of the quality of advice, and the finding that where a participant's estimate fell between the true answer and the advice, accuracy decreased for 8 out of 14 questionnaire items, can be reconciled because of the number of possible permutations of *estimate, advice* and *true answer* (see Table 5.8). This table is identical to the one presented in Chapter 4 (Table 4.17), but is repeated here for the purposes of clarity.

Table 5.8 Permutations of advice in relation to initial estimate

Permutation	*Increase in* → *numerical value* →			judgment revision due to conformity effect
A	Estimate	True answer	Advice	towards 't'
B	True answer	Advice	Estimate	towards 't'
C	Advice	True answer	Estimate	towards 't'
D	Advice	Estimate	True answer	opposite to 't'
E	True answer	Estimate	Advice	opposite to 't'
F	Estimate	Advice	True answer	towards 't'

The above illustration shows that in only 2 out of 6 permutations (permutation D and E) is advice an alternative pole of attraction to the true answer. Hence, it is not surprising that the pull of advice is masked by the overall pull of the true answer and advice.

This account is somewhat tentative given the few studies in the extant JAS literature that have systematically manipulated the quality of advice available to judges. Of these studies, none as far as I am aware, has specifically looked at the effect that accurate and inaccurate advice has on accuracy, and judgment change. However, where the quality of advice has been manipulated as part of other dependent measures, there is some consistency the results reported

here. Harvey and Fischer (1997) manipulated advice quality (advice equal to the correct value, advice less than the correct value, advice more than the correct value) that their judges received, and found that people are unable to assess advice quality successfully – judges were no more influenced by correct than by incorrect advice. However, Yaniv and Kleinberger (2000), suggest that the quality of advice is influential mainly as a stimulus to further cognitive processing/retrieval. Here, it is argued that inaccurate advice can be perceived as inaccurate by a cognitive 'plausibility check' i.e. a person may not know that the American Declaration of Independence took place in 1776, but they may sense that advice of '1450' is not accurate. Further, accurate advice could potentially stimulate memory retrieval, and facilitate the recovery of memories that were inaccessible to an individual otherwise. This latter argument is consistent with the notion of a cognitive account of advice use.

This study tested the cognitive account of advice utilization through analysis of a measure of judgment change. Recall that numerical advice alone could have induced judgment change for two possible reasons. First, a judge is likely to make attributions about the advice beyond the 'numeracy' of the advice; second, in comparison to the condition involving reasons, numerical advice is 'cognition lite' and relatively easily assimilated into a combined estimate. In contrast, reasons-based advice is likely to have been perceived as cognitively demanding, and subsequently under-utilized as far as forming a final estimate is concerned. It follows, therefore, that participants may have been less inclined to undertake the cognitive effort to assimilate reason-based advice. For this account to be parsimonious, differences in the amount of judgment change between reason-based advice conditions and purely numerical advice conditions should be apparent. Instead, there is little to distinguish the two experimental advice conditions in terms of the amount of judgment change (both conditions reveal significantly more judgment change than participants who did not receive advice), suggesting

that the quality of advice appears to have little effect upon the *amount* of judgment change. I also examined the extent of judgment change – overall, participants beneficially changed their judgments (on average), irrespective of the advice that they had available to them. However, this finding masked the scenario where a participant's initial estimate fell between the true historical date for a particular event and the available inaccurate advice. Here, participants revised their judgments to the detriment of estimation accuracy in every permutation of numerical and reasons based advice, and did so for 8 out of 14 questionnaire items. This finding undermines the cognitive account of egocentric discounting, as there were no differences in the amount of judgment change between reasons-based and purely numerical advice; nor were there differences in the extent of judgment change between reasons-based and numerical advice and purely numerical advice. Further, for the cognitive account to be parsimonious any advice should have resulted in accuracy improvements – 57% of our participants revised their estimates inaccurately when inaccurate advice was available to them. Whilst both the task specific account, and the cognitive activation account do not adequately capture the observed behaviour of participants, such behaviour is not inconsistent with what might be expected under the conformity to advice explanation.

The findings reported here did not entirely replicate those of Yaniv *et al.* (2000; 2004a; 2004b), despite similarities in methodology and procedure. Utilizing the WOA measure, Yaniv and Kleinberger (2000), and Yaniv (2004) report egocentric discounting where participants are observed to significantly discount advice ($\bar{x}$ = 0.27 and 0.29). Whilst participants in this study significantly discounted advice overall, their WOA score was far closer ($\bar{x}$ = 0.46) to 0.5 than the WOA observed for participants in Yaniv and Kleinberger (2000), and Yaniv (2004). However, this should not be taken as evidence for the mitigating effects of reasons-based advice upon the discounting phenomena. This is because there was no

significant difference in WOA scores between participants receiving numerical advice, and participants in receipt of reasons-based advice. The difference in WOA scores between the findings reported here, and that of previous research is potentially explainable by several factors. Firstly, this study involved 255 participants, whereas Yaniv and Kleinberger (2000) utilized the responses of 25 participants in study 1 (examining the effects of egocentric discounting), and 80 participants in study 2 (examining the sensitivity of participants to advice quality). Similarly, Yaniv, (2004), utilized the responses from 30 participants in Study 1 (examining advice weighting as a function of knowledge). It is possible that my findings may be more generalizable than those of previous research, due to a larger *N*. Second, differences between the findings reported here, and those of previous research may have been influenced by how the experimental stimuli were presented to participants. Both Yaniv and Kleinberger, and Yaniv, had their participants receive the experimental stimuli via computer terminals, whilst participants in the present study filled in questionnaires in the traditional 'paper and pencil' format. I offer no theoretical articulation of why this might be so, but merely highlight this factor as a possible source of explanation in future studies. Third, participants in previous research have received financial payoffs for participation and accurate responses. This may have induced egocentrism in participants as participants would be likely to attribute accuracy to their own internal cognitions, and be more likely to discount advice. Participants here, received no payoffs for participation or accuracy, and hence may have been less egocentric in outlook and more sensitive to potentially beneficial advice. Further, previous research has stressed the ecological validity of randomly drawn advice from a pool of estimates provided by pre-testing; the advice in our study was directly manipulated by the experimenter. No feedback in this study was available to participants as to the accuracy of their estimates (unlike Yaniv and Kleinberger, Study 1 and 2), but here, accurate advice was the same date as the true answer (similar to Yaniv and Kleinberger; dissimilar to Yaniv). Given the inconsistencies in procedure between Yaniv and Kleinberger (2000), and Yaniv (2004), it appears unlikely that a

confounding priming effect could partially explain the difference between the extent of egocentric discounting in our study and that found in previous research. A final explanatory factor could potentially involve cross-cultural differences – the current research was based upon a sample of students drawn from the University of Durham (UK), while Yaniv and Kleinberger, (2000), and Yaniv, (2004), based their research upon a sample of Israeli students.

One limitation of the present study is that only supportive reasons-based advice was available to participants. It is possible that the cognitive activation account of egocentric discounting could be demonstrated in a scenario where participants' self-generate counterfactual reasons, for why their estimates may be inaccurate. Such a manipulation could reveal that counterfactual reasoning may lead to enhanced estimation accuracy. Supportive evidence for such an explanation can be found in the work of Koriat, Lichtenstein and Fischoff (1980). In Study 1, Koriat *et al.* tasked their participants with a forced dichotomous choice, listing reasons both for and against the choice, and rating the probability that the choice was correct. Results suggested that the appropriateness of confidence judgments was improved. The procedure was simplified in Study 2 so that participants either (i) listed one reason in support for the choice (ii) listed one reason why the choice was incorrect (iii) listed both a reason for and a reason against the choice being correct. Here, it appeared that only the listing of contradictory reasons improved the appropriateness of confidence judgments.

Whilst the appropriateness of confidence judgments was not the central thrust of this study, the depth of cognitive activity and its beneficial effect upon estimation is. Hence, future research about how people utilize advice should include cognitive measures such as depth of processing

to discover the underlying mechanisms that determine how advice is assimilated into a combined estimate (Bonaccio and Dalal, 2005).

The implications of these findings for judges may be that people making estimates of a quantitative nature are likely to anchor upon a reference point and adjust any revised estimate in the light of this anchor – it is this mechanism which may underpin the conformity to advice explanation of the observed behaviour of participants. It is likely that such adjustments are 'sticky' (i.e. constrained by egocentric cognitions). This finding has added significance when the available advice to a judge is poor and bounds both the truth and the judge's initial estimate; in such a scenario judges are susceptible to adjusting their revised estimate in the direction of greater inaccuracy. Either way, supportive reasons for quantitative estimates offer little in the way of enhanced diagnosticity. However, it is arguable whether these findings generalize to situations where more than one advisor contributes advice to a judge. Here, the influence of advisor confidence, redundant reasons, majority, and appropriate minority judgment is likely to be significantly influential. Moreover, the effects of multiple reasons upon egocentric discounting could potentially be different from the decision dyads utilized here. A further limitation of this study was that the reasons-based advice was generated by the experimenter and subsequently rated for persuasiveness by two independent coders. Future studies should examine the differences between participant generated rationales (supportive/counter-attitudinal) attached to quantitative estimates, and rationales (supportive/counter-attitudinal) generated external to participants.

Ultimately, the provision of intrinsic information of the quality of numerical advice did not benefit the cognitive weighting policies adopted by participants. Instead, participants are

observed to ‘anchor’ upon their initial intuitions, and revise their judgments in the direction of available advice (to a degree). Such behaviour is entirely consistent with social conformity to advice.

Chapter 6: Knowledge of advice generation

6.1 Introduction

The findings reported in the previous chapters suggest that people prefer reasons-based advice over solely numeric advice, are constrained in their abilities to distinguish between 'poor' and 'good' advice under conditions of uncertainty, and where they do revise their judgment post-advice, tend to conform to advice. People 'anchor' upon an initial intuition and only shift their judgment in the 'direction' of advice, to a limited degree. The current study examines the question of whether knowledge of the process of advice generation, provides participants with a means, by which a cognitive evaluation can be made of available advice, and participants' pre-advice intuitive estimates. Here, knowledge of the process of advice generation is operationalized in two ways. First, numerical advice supported by a testimonial assertion that speaks to the veracity of the advice is made available to participants from an anonymous advisor. In a second experimental condition, an anonymous advisor provides participants with a procedural algorithm by which participants self-generate their own advice through inputting sub-component estimates, and multiplicatively combining these sub-component estimates to form advice (output-of-algorithm). Subsequently, participants are in a position to evaluate advice, and choose whether to incorporate such advice into their deliberations. Should this evaluation result in participants appraising advice as valid and coherent, beneficial judgment revision appears to be a likely outcome. However, participants may also be unable to make such a determination, but choose to conform to advice, irrespective of its actual quality.

Judgment change and Advice

Of interest then, is the degree to which reason-based advice, influences judgment change, in a JAS-type advice-giving situation, or whether conformity to advice extends to situations where information of the process of advice generation is available to judges, in addition to advice itself. It is contended here that one useful deconstruction of reasons-based advice involves distinguishing between 'process' reasoning, and 'non-process' reasoning. In this framework, Non-Process reasons-based advice involves broad claims to knowledge, expertise, and some numerical estimate of a target quantity (i.e. where a testimonial assertion constitutes advice). In making a 'holistic' estimate of some target quantity, an estimator in receipt of Non-Process advice, simultaneously considers all the relevant attributes of the estimation problem – including the weight to place upon Non-Process advice. In contrast, a definition of Process reasons-based advice, involves a specific form of reasoning – an 'algorithmic decomposition' of some problem or issue. Here, the attributes of an estimation problem are broken down into various constituent parts, allowing consideration of the attributes of the problem individually, or as sub-groups of attributes. The decomposition of complex or uncertain estimation problems, aids decision making and estimation, because holistic judgments deteriorate over time due to the limits on human information processing capacity. The systematic decomposition of complex or uncertain estimation problems, however, relaxes the information processing demands upon an estimator, potentially allowing a more rigorous consideration of the attributes of an estimation problem (Fischer, 1977; Kleinmuntz, 1990).

In essence, algorithmic decomposition provides a specific procedural means of solving difficult problems, where information is uncertain or not available. As informational combination is mechanical, (i.e. informational sub-components can be combined either multiplicatively or additively), in such an approach, it is argued that reliability is improved

(MacGregor and Lichtenstein, 1991). In comparisons with unaided judgment, algorithmic decompositions of experimental almanac-type problems show improvements in estimation performance (Armstrong, Denniston and Gordon, 1975; MacGregor, Lichtenstein and Slovic, 1988; MacGregor and Lichtenstein, 1991; MacGregor, 2001). A typical example is provided by MacGregor *et al.* (1988, 1991), whose participants were asked, *'How many pieces of mail were handled by the US postal service last year?'* In order to assist participants in formulating an estimate in response to this question, MacGregor *et al.* provided their participants with the following algorithm –

A. What is the average number of post offices per state?
B. What is the number of states?
C. Multiply (A) times (B) to get the total number of post offices.
D. How many pieces of mail per day are handled by the average post office?
E. Multiply (C) times (D) to get the total pieces of mail per day for all post offices.
F. How many days are there in a year?
G. Multiply (E) times (F) to get the number of pieces of mail handled in a year by the postal service.

By decomposing the original question of interest ('*How many pieces of mail were handled by the US postal service last year?*') into sub-components ('A', 'B', 'D', and 'F' in the preceding example); assigning numerical values to these sub-components, and subsequently combining these numerical values ('C' and 'G' in the preceding example), it is argued that an individual is able to achieve a better estimate of some troublesome quantity, than that achievable, by holistic unaided estimation.

The preceding findings are consistent with the 'divide and conquer' principle (Henrion *et al.* 1993; Kleinmuntz *et al.* 1996; Morera and Budescu, 1998), where it is argued that it is less demanding (and less error prone) to produce estimates to the constituent parts of a problem - and subsequently re-combine these responses to form an overall estimate - than to produce an overall 'holistic' evaluation. The 'divide and conquer' principle is widely used to enhance estimation, or forecasting, in various contexts, from medicine to accounting. Early studies in clinical judgment showed that a linear model of clinical judgment outperformed holistic judgment (Meehl, 1957; Goldberg, 1968, 1970); whilst auditors' assessments of conditional probabilities were improved by algorithmic decomposition, in comparison to list-type aids, in the context of accounting (Bonner, Libby and Nelson, 1996). Similarly, studies (such as JAS) that task participants to make point estimations of target quantities in almanac-type tasks (where the answer is known to the experimenter, but not to the participant), show estimation improvements for participants utilizing the decomposition approach, above that achieved by holistic evaluation (Armstrong, Denniston and Gordon, 1975; MacGregor, Lichtenstein and Slovic, 1988; MacGregor and Lichtenstein, 1991; MacGregor, 2001).

The results from the preceding studies suggest that people make less accurate assessments through intuition, than they do using simple linear models (Dawes, 1979; Dawes, Faust and Meehl, 1989). However, what is less clear is whether people are likely to revise their initial unaided judgment, when subsequently presented with either Process advice (an algorithmic decomposition of some problem), or Non-Process advice (where a testimonial assertion constitutes advice). This is because many of the aforementioned studies do not measure judgment revision post-advice. Moreover, if people are disposed to change their initial unaided judgment in the light of advice, it is unclear to what extent either Process, or Non-Process advice influences judgment change. Should people indiscriminately conform to

advice, then either advice type is likely to be influential where people choose to revise their unaided intuitive estimate, post-advice. Further, the studies discussed here, do not successfully define the cognitive factors that facilitate judgment change (Dalal and Bonaccio 2006). The potential influence of Process advice is considered in the next section.

It could be argued that Process advice is more likely to facilitate judgment change, than other types of advice (such as a testimonial assertion), merely on the grounds that greater cognitive effort is involved in working through an algorithmic decomposition of some problem; in comparison to an holistic evaluation of a testimonial assertion. However, algorithmic decompositions, such as the 'US mail' problem, are based upon the assumption that the estimation errors of each sub-component are uncorrelated, and that these errors cancel each other out in the multiplicative re-composition process. Where errors on the sub-component estimates of a particular algorithmic decomposition do not cancel each other out during multiplicative re-composition, the resulting output of the algorithm is likely to be highly inaccurate. In these circumstances, people may judge themselves to be incapable of inputting credible inputs to the algorithm, and subsequently are *less* likely to revise an unaided estimate in the light of the output of the algorithm (MacGregor, 2001). Hence, cognitive effort, in isolation, may not be a primary determinant in facilitating judgment change.

A further possible determinant of judgment revision concerns the relationship between an individual's unaided pre-advice estimate of some troublesome quantity, the output of Process (or Non-Process advice), and the true answer to the estimation problem. Where participants have the opportunity to revise their judgment in the light of advice, and choose to do so, they are likely to revise their unaided pre-advice estimate, so that the absolute numerical difference

between their unaided pre-advice estimate, and the available advice, is greater than the absolute difference between their post-advice estimate, and the available advice. Moreover, such judgment revision is likely to be beneficial, on average, because the structured exchange of purely numerical information in JAS research, does not disallow the possibility that judgment revisions, are an artifact of the structured exchange. This might be so because any advice that falls between a participants' unaided estimate, and the true answer (or where the true value falls between a participant's initial estimate and the advice), is likely to be beneficial (permutations A, B C and F in the diagram below). Only an unaided estimate that falls between the true answer and subsequent advice has the potential to degrade estimation (permutations D and E). The diagrammatic representation below shows that in 2 out of 3 of the possible permutations discussed above, beneficial revisions of judgment can be accounted for by a conformity effect (i.e. an artifact of the structured exchange of information).

Fig. 6.1. Likely judgment revisions in the possible permutations of true answer and advice, in relation to an intuitive pre-advice estimate

Permutation	> *numerical* *value* →			judgment revision due to conformity effect
A	Estimate	True answer	Advice	towards 't'
B	True answer	Advice	Estimate	towards 't'
C	Advice	True answer	Estimate	towards 't'
D	Advice	Estimate	True answer	opposite to 't'
E	True answer	Estimate	Advice	opposite to 't'
F	Estimate	Advice	True answer	towards 't'

An important caveat to the assertion that beneficial judgment change may be attributable to the structure of the estimation task is that such an effect is also specific, to certain types of JAS-type estimation situations. Typically, such tasks are concerned with frequencies (e.g. estimate the total annual tonnage of lumber exported from New York in 1964), rather than judgment (e.g. 'given two estimates of property *X*, which do you think is closest to the true value?').

In sum, where people are provided with an algorithmic decomposition of an estimation problem, they are able to judge its logical coherence against their own intuition, and come to a view of the utility of the provided algorithm. If the logic of an algorithmic decomposition is accepted by a judge, and the output of the algorithmic decomposition is dissimilar to an individual's unaided holistic estimate, then it is likely that a judge will adopt the output of the algorithm, and subsequently revise their unaided pre-advice estimate of the quantity in question (i.e. appropriate judgment change is induced by the triumph of a superior rationale). People may however, effortlessly and flawlessly, recall accurate information that leads to an optimal estimate, or retrieve sufficient information to self-construct a viable cognitive algorithm that 'solves' the problem sufficient for the purposes of the environmental context, without the need for any external intervention – in such cases the output of an algorithmic decomposition is likely to be ignored. Alternatively, people may recognize that they are unable to make meaningful inputs into a specific algorithmic decomposition, but choose to incorporate the output of the algorithm in preference to relying upon their intuition. Having discussed the possible influence of Process advice upon judgment revision, I now turn to the possible influence of Non-Process advice upon judgment revision.

In contrast to Process advice, Non-Process advice offers little opportunity for judges to evaluate the internal logic of testimony. In my framework, this is because people were not provided with any numerical information about any of the potential attributes identified in the Non-Process advice (beyond the numerical output of Non-Process advice), nor were the advisors described as expert. Nevertheless, Non-Process advice is potentially beneficial to judges, in so far as such advice may highlight attributes of the estimation problem that judges were hitherto unaware of. In such a scenario, Non-Process advice could potentially facilitate beneficial judgment revision.

The purpose of this investigation then, was to determine the influence of knowledge of the process of advice generation upon judgment revision. Specifically, to ascertain whether participants changed judgment more often, and/or to a greater extent when in receipt of either Process or Non-Process advice, or where people revise an judgment post-advice, they do so by indiscriminately conforming to advice (to a degree).

6.2 Method

Participants

Participants in this study (N = 142), were drawn from members of staff, and students of the University of Durham, UK. Some participants were recruited by the author circulating an email invitation to staff and students of Durham Business School (N = 60). Here, participants accessed the questionnaire by 'clicking' on a hyperlink in the invitation email, that opened the questionnaire on an individual's computer. On completion, the data from the online questionnaire was saved automatically in a database on a Durham Business School server. However the majority of participants were students who the author recruited by visiting a

lecture theatre (with the permission of staff colleagues) where students were gathered to attend their studies. Here, students completed the questionnaire in a 'pencil and paper' format - all participants took approximately the same amount of time to complete the questionnaire. The sample was made up of 95 males and 45 females (two participants failed to provide details of their gender). 133 participants provided information as to their age, which allowed calculation of the mean ($\bar{x}$ = 20.38 years, sd = 5.43), and median (= 19 years), of the sample, where the range was from 18 years to 54 years.

Stimuli and Materials

Participants were tasked with providing estimates to four questions concerning uncertain quantities (see Appendix III). Some participants made their responses via a computer terminal connected to the internet (N = 60), and others made their responses in a traditional 'pencil and paper' format (N = 82). The presentation order of the four estimation questions was randomized to mitigate any possible learning effects.

Procedure

The experiment was conducted in two phases. Initially, all participants made an unaided estimate of an uncertain quantity, and were tasked with providing an upper and lower bound, so that they were 90% certain that the interval bracketed the true answer. At no time were participants made aware of the true answer. In the second phase of the experiment, participants were randomly allocated to one of three experimental conditions; a control group (no advice), a Process advice group, and a Non-Process advice group. Participants in receipt of Process advice were provided with a step-by-step algorithm that divided the estimation task

into sub-components. Participants were informed that, for each question, an algorithm had been provided by a different anonymous advisor; identifiable only by a first name (i.e. 'Kate', 'Joe', 'Tina' or 'David'). Participants were instructed to input sub-estimates into the system where indicated, and ultimately mechanically re-combine these sub-estimates to form an output of the algorithm. The output of the algorithm constituted the advice that was available to participants in the Process advice group. Participants were free to utilize or reject this advice in determining their final best estimate of the target quantity. Participants in the Non-Process advice group received advice that contained numerical information evidenced by a testimonial assertion, whose source was identified by a first name (e.g. 'Joe'). Participants in the Non-Process advice condition could choose to attend to this advice, or rely upon their intuitive pre-advice estimate. After each estimation question, all participants were given the opportunity to revise the intuitive pre-advice estimate they had provided in Phase I of the experiment, in the light of the advice available to them (or to 'think again' where people had not received advice). All participants were then asked to make a final best estimate of the uncertain quantity in the estimation task, and provide an upper and lower bound so that they were 90% certain that this interval contained the true answer. Finally, participants evaluated the utility of the advice made available to them on a 5-point Likert-type scale (anchored at one end by 'No use at all' and at the other by 'absolutely invaluable'). In all experimental conditions the order of presentation of our estimation question was randomized to mitigate learning effects. Participants who completed the questionnaire on-line received payment (£4) for taking part in the experiment; participants who completed the 'pencil and paper' format questionnaire did not receive payment. There is no theoretical rationale for this distinction, as payment was not linked to performance in the study - it was an incentive for participation. On the grounds of practicality and the availability of resources, the author was happy to accept data from participants who did not require payment.

6.3 Results

Before considering the data in terms of analysis, it was necessary to first inspect the rate of response to the four estimation tasks that participants were asked to complete. Table 6.1 (below) shows the number of responses collapsed across questions, indicating satisfactory levels of responses (even though it appears that participants fatigue may have been a factor post-advice as the percentage of participants providing a response falls somewhat).

Table 6.1 Number of responses

Participants *N* = 142	Unaided estimate	Pre-advice Lower Bound	Pre-advice Upper Bound	Post-advice estimate	Post-advice Lower Bound	Post-advice Upper Bound
Responses *N*	549/568 (96.65%)	510/568 (89.79%)	510/568 (89.79%)	434/568 (76.41%)	445/568 (78.34%)	387/568 (68.13%)

Given that there appears to be sufficient data to warrant analysis, the primary question of interest, in terms of the aims of this study, was to determine the influence of knowledge of the process of advice generation (Process, and Non-Process advice) upon judgment change, and to test the consistency of these results with the conformity to advice explanation, of judgment change. Secondly, a regression modelling approach was adopted in an attempt to identify some of the cognitive mechanisms that facilitate judgment change. Prior to proceeding with the analysis, it was necessary to check that the responses of participants who completed estimation tasks 'on-line', could be treated as the same population, as participants who completed questionnaires by hand. Hence, the absolute difference between a participant's pre-advice estimate, and their subsequent post-advice estimates was calculated for each question. Subsequently, the scores for participants who responded in the traditional 'paper and pencil' mode, were compared to the responses of participants who made their responses to the estimation questions online. As the dependent variable of absolute numerical difference

between pre-advice estimate, and post-advice estimate did not meet the assumption of normality the non-parametric Mann-Whitney U test was performed to test for significant differences in the absolute difference scores of people responding on paper, and online (see Tables 6.2-6.4). Whilst there was a significant difference between the absolute difference scores of people responding on paper, when compared to people responding online, for participants in the control condition (Question 2 only); and similarly for participants in the Process advice condition (Question 3 only), these differences do not appear to constitute a systematic pattern of variation in absolute difference scores attributable to the mode of response.

Table 6.2 Controls – absolute numerical difference between initial estimate and final estimate.

Estimation question	*N*	Mann-Whitney U	Participants responding on paper N_1	Participants responding online N_2	*p*
1	51	256.50	32	19	0.07
2	51	237.50	32	19	0.03*
3	51	256.50	32	19	0.07
4	51	285.00	32	19	0.27

*$p < 0.05$

The above table shows that where participants did not receive advice, in one out of four estimation problems, a significant difference in mean absolute difference scores between people responding online, and people responding in a 'pencil and paper' format was reported (Question 2). However, the removal of two outlying values resulted in any differences between people responding online, and people responding by 'pencil and paper', being non-significant ($U = 237.5$, $N_1 = 30$, $N_2 = 19$, $p < 0.06$).

Table 6.3 Process advice – absolute numerical difference between initial estimate and final estimate.

Estimation question	*N*	Mann-Whitney U	Participants responding on paper N_1	Participants responding online N_2	*p*
1	23	32.00	18	5	0.32
2	24	28.50	18	6	0.08
3	29	20.00	19	10	0.005*
4	29	74.50	20	9	0.46

*$p < 0.05$

Table 6.3 shows that where participants received Process advice, in one out of four estimation problems, a significant difference in mean absolute difference scores between people responding online, and people responding in a 'pencil and paper' format was reported (Question 3).

Table 6.4 Non Process advice – absolute numerical difference between initial estimate and final estimate.

Estimation question	*N*	Mann-Whitney U	Participants responding on paper N_1	Participants responding online N_2	*p*
1	29	57.00	21	8	0.19
2	35	80.0	28	7	0.45
3	35	64.50	28	7	0.16
4	26	39.00	19	7	0.11

Table 6.4 shows that there were no significant differences in absolute difference scores between people responding online and people responding by 'paper and pencil'. On average, the preceding results do not suggest systematic variation in mean absolute difference scores

attributable to the mode of response. Hence, for the purposes of the following analysis the data from people responding online and people responding by 'paper and pencil' are treated as the same population. Initially, the data from participants was aggregated and tabulated (see Table 6.5 and 6.6).

Table 6.5 Process advice - mean scores (1,000,000s), pre-advice estimates, advice, and post-advice estimates

Question	*N*	Pre-advice estimate $\bar{x}$	*sd*	Advice $\bar{x}$	*sd*	Post-advice estimate $\bar{x}$	*sd*	True
Q1. Lottery	23	17.58	25.11	15,478.66	49,928.98	488.79	1371.52	38.29
Q2. HE	24	3.80	10.14	543,945.69	249.93	5.42	12.56	0.41
Q3. Road deaths	29	0.27	0.73	40,061,290	211,942,069	1.13	2.86	0.003
Q4. Passengers	29	2.78	6.65	5048.91	27108.77	7.34	17.16	1.00

Table 6.5 shows that, on average, where the value of Process advice was numerically greater than a participant's pre-advice estimate, the same participant's post-advice estimate was greater than the numerical value of his/her pre-advice estimate. However, if the 'inflation' of participants' pre-advice, to post-advice, estimates, is attributable to Process advice, it is unfortunate that such judgment change was not beneficial in terms of accuracy. Participants in receipt of Process advice, who chose to revise their judgment post-advice, appear to be reducing only a proportion of the absolute numerical difference between their pre-advice estimate and Process advice, post-advice, which is consistent with the conformity to advice account of the data.

A similar pattern of results was observed when Non-Process advice was considered. Where Non-Process advice was numerically greater than a participant's pre-advice estimate, the same participant's post-advice estimate was greater than his/her pre-advice estimate, on average. Also, where Non-Process advice was numerically less than a participant's pre-advice estimate, on average, the same participant's post-advice estimate was less than his/her pre-advice estimate, on average.

Table 6.6 Non-Process advice - mean scores (1,000,000s), pre-advice estimates, advice, and post-advice estimates

Question	*N*	Pre-advice estimate $\bar{x}$	*sd*	Advice $\bar{x}$	*sd*	Post-advice estimate $\bar{x}$	*sd*	True
Q1. Lottery	29	12.54	19.77	20.00	0.00	15.68	8.86	38.29
Q2. HE	35	11.10	17.01	50.00	0.00	22.24	21.51	0.41
Q3. Road deaths	35	0.07	0.14	0.01	0.00	0.03	0.06	0.003
Q4. Passengers	26	2.71	5.12	0.50	0.00	0.84	0.94	1.00

Are participants in this study merely conforming (to a degree) to any available advice?

Ultimately, what these preliminary observations suggest, is that participants who choose to revise their judgment post-advice, are indiscriminately conforming to advice (irrespective of condition); to the extent that, they are revising their judgment by proportionately reducing the absolute numerical difference between their pre-advice estimates and the advice, post advice. Prior to pursuing this line of argument further, it was necessary to determine if there were differences between conditions, in terms of both the number of judgment changes, and also the amount of judgment change.

Does attending to advice influence judgment change?

The number of judgment revisions for each individual question was calculated, and it is clear from Table 6.7 that a greater number of people did not revise their judgment post-advice, in comparison to those people who did revise their unaided holistic estimate.

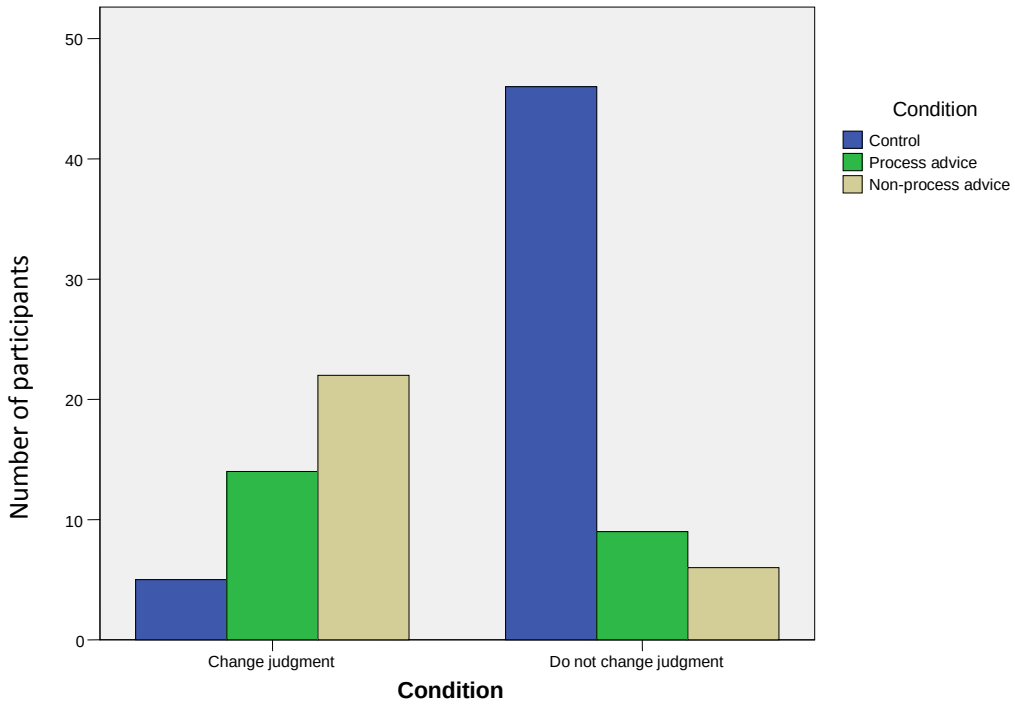
Table 6.7 Number of judgment revisions

Condition	Question	N	Change judgment
Control	1	51	5 (9.8%)
	2	51	7 (13.7%)
	3	51	5 (9.8%)
	4	51	2 (3.9%)
Non-Process advice	1	29	23 (79.3%)
	2	23	14 (60.9%)
	3	23	14 (60.9%)
	4	15	12 (80%)
Process advice	1	23	14 (60.9%)
	2	20	9 (45%)
	3	20	6 (30%)
	4	20	14 (70%)

The table above clearly indicates that people in receipt of advice tended to change judgment more often than people who did not receive advice. This data was next examined question by question. Fig 6.2 shows the number of judgment changes for each condition, for participants responding to Question 1 of the questionnaire. Here, it is clear that both advice conditions tended to result in more judgment changes, than participants in the control condition. However, there was no significant difference between the number of judgment changes that participants reported when in receipt of Process advice, than when in receipt of Non-Process

advice (χ^2 = 1.91, *df* = 1, *p* < 0.17). This finding further supports conformity to advice, as irrespective of advice condition, people tended to revise their judgment as often as each other, on average.

Fig 6.2. Q1 Judgment change



A similar pattern of results obtained for Question 2 (see Fig. 6.3), where people in receipt of advice, revised their unaided pre-advice estimate, more often than people who did not receive any advice. Here, the difference in the number of judgment changes between participants in receipt of Process advice, and participants in receipt of Non-Process advice did not reach significance (χ^2 = 0.45, *df* = 1, *p* < 0.50). The same pattern of results was observed for participants who responded to Question 3 (see Fig. 6.4). Once more there were no significant differences in the number of judgment changes between participants in the two advice

conditions ($\chi^2 = 3.67$, $df = 1$, $p < 0.06$).

Fig. 6.3 Q2 Judgment change

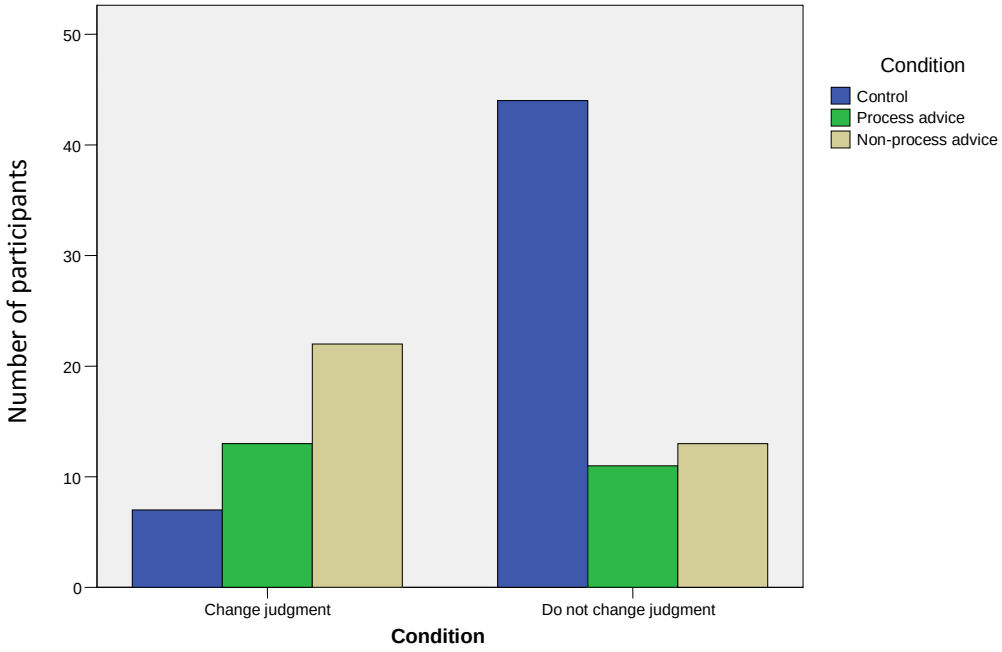
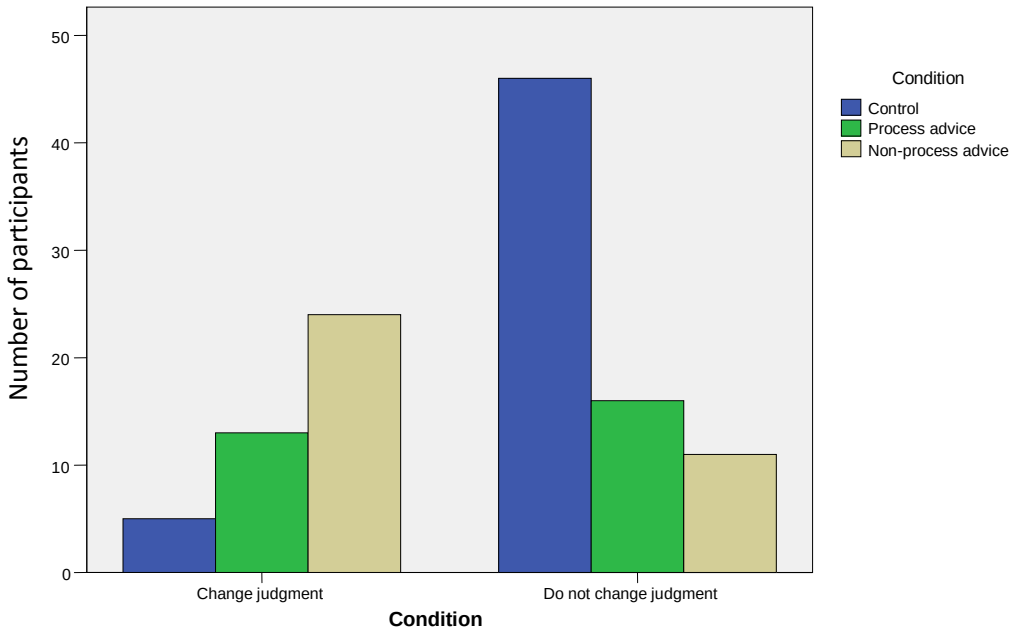
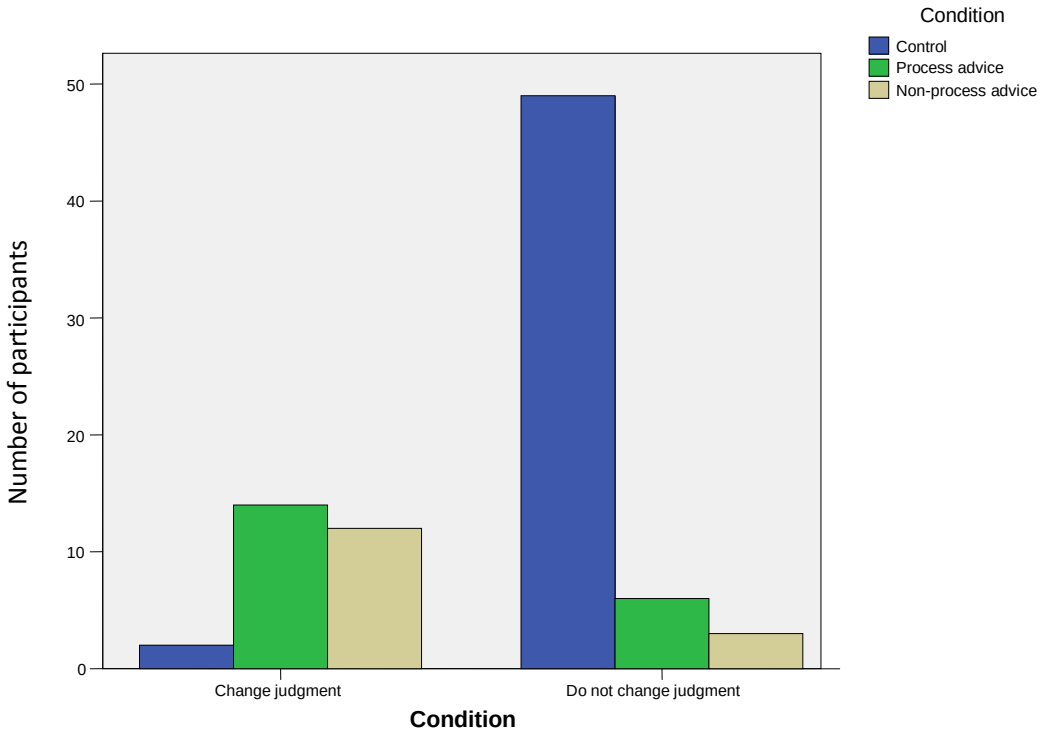


Fig 6.4. Q3 Judgment change



The same pattern of results that had been observed on the first three questions of the questionnaire was once more observed on the fourth question (see Fig. 6.5).

Fig. 6.5. Q4 Judgment change



Participants responding to Question 4 clearly changed judgment more often, when in receipt of advice, than when not in receipt of advice. However, there was no significant difference observed between the number of judgment changes reported by people in receipt of Process advice, and the number of judgment changes reported by people in receipt of Non-Process advice ($\chi^2 = 0.45$, $df = 1$, $p < 0.50$). The preceding findings strongly suggest that people are more disposed to revise their judgment when in receipt of some advice, and that where people revise their initial unaided estimate post-advice, knowledge of the process of advice generation does not appear to directly influence the number of judgment revisions. This finding is consistent with the conformity to advice account of these data.

Can participants discriminate between good and poor advice?

The preceding findings are not particularly illuminating – even though participants in Question 3 and in receipt of Process advice changed judgment *less* often than not, no significant difference between conditions was observed. Consideration was next given to the *extent* of judgment revision post-advice. Here, the dependent variable of amount of judgment change was not normally distributed, and it was necessary to Log transform the data to meet the assumptions of normality. Subsequently an independent samples t-test was performed on the data of participants who revised their judgment post-advice, to ascertain whether there was any significant difference in the absolute difference scores between conditions. Table 6.8 (below) shows that there was no difference in the mean absolute amount of judgment change between the advice conditions ($t = 1.29$, $df = 157$, $p < 0.20$).

Table 6.8 Mean absolute amount of judgment change

	Observations N	$\bar{x}$	sd	Log $\bar{x}$	sd
Process Advice	66	155,556,698.94	816,953,860.66	14.10	2.96
Non Process-Advice	93	8,132,201.40	12,975,150.35	13.45	3.23

Table 6.8 is further evidence that where participants revise their judgment post-advice, they appear to be indiscriminately conforming to advice, on average. This issue was next examined on a question-by-question basis (see Table 6.9).

Table 6.9 Mean absolute amount of judgment change

Question	Advice Condition	N	$\bar{x}$	*sd*	Log $\bar{x}$	*sd*	*t*	*df*	*p*
1	Process	15	666,040,366.67	1,654,302,311.29	16.71	2.80			
	Non-Process	23	13,238,260.87	15,772,936.73	15.72	1.47	1.25	19.11	0.23
2	Process	14	5,523,429	9981952.25	13.73	2.21			
	Non-Process	23	17,112,606.52	4644241.74	15.77	1.90	-2.99	35	0.005*
3	Process	15	1,660,622	3,250,645.87	11.57	2.83			
	Non-Process	24	65,532.5	122,508.79	9.22	2.15	2.94	37	0.006*
4	Process	22	7,904,513.64	18,881,090.67	14.28	1.99			
	Non-Process	23	2,463,130.43	4,961,518.69	13.27	1.60	1.88	43	0.07

*$p < 0.05$

From Table 6.9 it can be concluded there were significant differences in the mean absolute amount of judgment change between participants in receipt of Process advice, and participants in receipt of Non-Process advice, in Questions 2 and 3. Participants in receipt of Non-process advice, and responding to Question 2, reported greater mean absolute difference scores on average, than participants responding to Question 2, and in receipt of Process advice (t = -2.99, df = 35, $p < 0.005$). Conversely, participants in receipt of Process advice, and responding to Question 3, reported greater mean absolute difference scores on average, than participants in receipt of Non-process advice and responding to Question 3 (t = 2.94, df = 37, $p < 0.006$). There were no significant differences between participants in receipt of Non-Process advice, and Process advice, in terms of mean absolute difference scores, for participants responding to Question 1 and 4. These findings strongly suggest that participants in receipt of Process advice, changed judgment no more often, and to no greater extent, than participants in receipt of Non-Process advice. This may suggest that participants are, at least in part, indiscriminately conforming to advice, merely because the advice is available. The following section discusses this possibility.

Recall that judgment revision is possibly an artifact of the structured exchange of numerical information in JAS-type tasks, in so far as, if the absolute numerical difference between a participants' pre-advice estimate, and the true answer is *greater*, than the absolute numerical difference between an advisors estimate and the true answer; participants choosing to attend to advice will revise their judgment beneficially. In contrast, where the absolute numerical difference between a participant's pre-advice estimate and the true answer is *less*, than the absolute numerical difference between an advisor's estimate and the true answer; participants choosing to attend to advice will not revise their judgment beneficially. Hence, indiscriminately following the advice offered by advisors is not always an optimal strategy for

accurate estimation (in the four estimation tasks participants completed for the purposes of this study). To test the possibility that participants indiscriminately attend to advice, a measure of judgment change relative to advice was formulated (¦pre-advice estimate – advice¦ - ¦post-advice estimate – advice¦). Should the values of the observations for judgment change relative to advice, be positive, on average, then it is possible to infer that participants have revised their judgments so that, the absolute numerical difference between their post-advice estimate, and the advice, is less than the absolute numerical difference between their pre-advice estimate and advice. Similarly, should the value of the measure (judgment change relative to advice) be negative, this indicates that a participant has rejected the available advice. Table 6.10 shows that where participants in receipt of Process advice, did revise their pre-advice estimate, they did so by reducing the absolute numerical difference between their post-advice estimate and the advice (i.e. the mean value for each question is positively signed).

Table 6.10 Mean judgment change relative to advice - for participants in receipt of Process advice

Question	N	$\bar{x}$	*sd*
1	14	551,124,642.86	1,227,778,216.52
2	13	2,593,692	8,574,989.25
3	13	1,891,064	3,448,676.26
4	22	171,218.18	18,126,510.37

Similarly, participants in receipt of Non-Process advice, who revised their judgment post-advice, did so by reducing the absolute difference between their pre-advice estimate and advice; in comparison to the absolute numerical difference between their post-advice estimate and advice (see Table 6.11).

Table 6.11 Mean judgment change relative to advice - for participants in receipt of Non-Process advice

Question	N	$\bar{x}$	*sd*
1	23	10,151,304.35	15,701,680.37
2	22	15,526,818.18	12,238,752.42
3	24	65,449.17	122,541.07
4	22	2,397,818.18	5,136,045.58

In sum, the mean scores for judgment change relative to advice are positively signed, indicating that irrespective of condition, participants who revised their judgment post-advice did so by reducing the absolute numerical difference between their estimate and the advice. These findings are consistent with the idea that where participants are influenced by advice, they revise their judgments by moving towards the numerical value of the advice, irrespective of the quality of advice (Harvey and Fischer, 1997). This finding was further supported when participant's ratings of the utility of the advice were considered.

Do participant's who value advice, follow it?

Advice in this study was differentiated between Process and Non-Process advice, on the basis that Process advice offered a coherent algorithm for solving seemingly intractable estimation problems. In contrast, Non-Process advice was constituted as a testimonial assertion. In these circumstances, it might be expected – if participants were motivated to estimate accurately – that people would value Process advice highly. The results of a Median test (reported in Table

6.12), show that there were no differences between treatment conditions, in terms of the value participants attributed to advice.

Table 6.12 Median test of participants' perceived utility of advice ratings

Question	Process advice *N*	Median	> Median	<= Median	Non-Process advice *N*	> Median	<= Median	*df*	χ^2	*p*
1	22	4	2	20	29	1	28	1	0.72	0.40
2	23	3	8	15	35	15	20	1	0.38	0.54
3	29	3	10	19	35	17	18	1	1.29	0.26
4	29	3	13	16	26	14	12	1	0.45	0.50

Given the present analysis of judgment change, where it appears that participants who choose to revise their judgment, do so by indiscriminately following advice; it is perhaps unsurprising that people appear to value different types of advice equally.

Consideration was next given to whether people who revised their judgment post-advice rated the utility of the available advice higher than people who did not change judgment. Superficially, it may seem attractive to pool participant's utility ratings for the purposes of analysis. However, this analysis is not particularly meaningful given the individual characteristics of each question that are pertinent to the observed results (e.g. there are very large differences between the absolute magnitude of the true answer and advice between questions). Given this caveat, the analysis of perceived utility of advice

ratings was performed question by question (see Table 6.13). In the Process advice condition, where people revised their judgment post-advice in response to Question 3; participants attributed higher utility ratings of advice, than ratings attributed by people who did not choose to revise their judgment post-advice ($U = 37.5$, $N_1 = 13$, $N_2 = 16$, $p < 0.002$). There were no significant differences (between people who revised their judgment post-advice, and people who did not revise their judgment), in participants' utility ratings of advice on Questions 1, 2 and 4 for people in receipt of Process advice. Where people received Non-process advice, it was clear that people who revised their judgment post-advice, valued the advice, to a greater extent than people who did not change judgment for two out of four of the estimation problems.

Table 6.13 Participants in receipt of advice – perceived utility ratings of advice

Advice	Question	Change Judgment *N*	Median	Do not Change Judgment *N*	Median	Mann-Whitney *U*	*p*
Process	1	14	3	8	3	52.5	0.80
	2	13	3	10	2	47.5	0.26
	3	13	4	16	2	37.5	0.002*
	4	22	3	7	3.5	57.5	0.30
Non-Process	1	23	4	6	3	37	0.05*
	2	22	4	13	3	92	0.07
	3	24	4	11	2	74	0.03*
	4	22	4	4	3.5	37.5	0.62

$p < 0.05$

These findings are consistent with the idea that the perceived utility of advice is an important determinant in judgment change. However, participants in receipt of Process advice responding to Questions 1, 2 and 4, did not report any differences in utility ratings of advice irrespective of whether people revised their judgment or not, post-advice. This may have been because of individual differences in task relevant knowledge of the items to be estimated, pre-advice. In order to investigate this possibility, significance tests for differences between questions, in terms of the width of pre-advice confidence intervals, were carried out.

If participants are *less* confident prior to accessing advice, are they *more* likely to revise their judgment post-advice?

Recall that 'narrow' confidence intervals, around a pre-advice estimate, may be indicative of greater pre-existing knowledge of the quantity to be estimated, than wide confidence intervals. The data from all four estimation questions was pooled in order to check for significant differences in the width of participants' pre-advice confidence intervals. As the dependent variable of initial pre-advice confidence did not meet the assumption of normality, it was necessary to perform a non-parametric Kruskal-Wallis test to ascertain whether there was a significant difference between conditions. This showed that, on average, there was no difference between conditions, in terms of participant's pre-advice initial confidence levels ($\chi^2 = 4.32$, $df = 2$, $p < 0.12$). This analysis was repeated for each estimation question. Participants responding to Question 1 did not report any significant differences between conditions, on average, in terms of pre-advice initial confidence ($\chi^2 = 2.35$, $df = 2$, $p < 0.31$). Although people responding to Question 1, and responding online reported higher levels of confidence, on average, than people responding on paper ($U = 746$, $N_1 = 63$, $N_2 = 32$, $p < 0.04$). Participants responding to Question 2 reported significant differences in pre-advice initial confidence scores, on average ($\chi^2 = 14.15$, $df = 2$, $p < 0.001$). A follow-up Median test revealed that this difference was between controls and participants in receipt of Non-Process advice ($\chi^2 = 15.30$, $df = 2$, $p < 0.0005$). People responding to Question 2, and responding online reported higher levels of confidence, on average, than people responding on paper ($U = 539$, $N_1 = 65$, $N_2 = 30$, $p < 0.0005$). Where participants responded to Question 3, there were no significant differences between conditions in pre-advice initial confidence score on average ($\chi^2 = 1.27$, $df = 2$, $p < 0.53$). Consistent with the preceding findings, participants responding to Question 3 reported higher levels of confidence when responding online ($U = 797$, $N_1 = 65$, $N_2 = 35$, $p < 0.01$). Participants who responded to Question 4 reported no significant differences in levels of pre-advice confidence between conditions on average ($\chi^2 = 1.43$, $df = 2$, $p < 0.49$).

Nor did participants responding to Question 4 report any significant difference between those responding online, and those responding on paper, in terms of pre-advice initial confidence scores ($U = 882$, $N_1 = 62$, $N_2 = 35$, $p < 0.13$).

In sum, there is little evidence to suggest systematic differences between conditions, or questions, in terms of participant's pre-advice initial confidence scores. However, in three out of four of the estimation questions participants responded to online, participants reported significantly greater pre-advice initial confidence, than participants responding on paper. Next, the relationship between pre-advice initial confidence and subsequent judgment change was examined. In the aggregate there was no significant relationship between pre-advice initial confidence and subsequent judgment change (see Table 6.14).

Table 6.14. Correlation between pre-advice initial confidence and subsequent judgment change

Question	N	Spearman's '*r*'	*p*
1	95	0.15	0.14
2	101	- 0.37	0.0005*
3	106	0.07	0.48
4	106	0.01	0.90

$*p < 0.05$

Table 6.14 shows that there was only a significant relationship between pre-advice initial confidence, and subsequent judgment change, for participants responding to Question 2. Here, the correlation is negative, indicating that as pre-advice initial confidence become 'wider' (i.e. decreased), so subsequent judgment change became more likely. However, there is no strong

evidence in the remaining three estimation questions to suggest that pre-advice initial confidence is significantly correlated to subsequent judgment change.

Post-advice, there were no significant differences between conditions in terms of participant's post-advice final confidence scores, on average ($\chi^2 = 4.71$, $df = 2$, $p < 0.10$). Further, it is perhaps unsurprising that participants who made their responses online, reported that they were more confident in their estimates post-advice than participants responding on paper ($U = 13203.5$, $N_1 = 256$, $N_2 = 132$, $p < 0.0005$).

Recall that most participants did not revise their intuitive pre-advice estimate, post-advice. The data analyzed so far, seemed to suggest only that participants who changed judgment, did so under the indiscriminate influence of advice. In order to determine whether this was the 'whole story', participants' data was re-analyzed by developing theoretical models and testing these models through regression analysis. Such an analysis is likely to be uninformative where the data from all questions is pooled, due to the differing characteristics of each question (the magnitude of the true answer and the advice are notably different for each question). It is preferable therefore, to perform regressions for individual questions. The following section will outline the regression approach and discuss variables that will be entered into the regressions.

The main questions multiple regression answers

Multiple regression is an appropriate method of data analysis in the context of this study, as the results of this analysis provides some answers to several questions of interest. First, multiple regression attempts to explain the variation in a single dependent variable due to *n* independent variables (regressors). The degree to which the regressors, (taken together), explain the

variation in the dependent variable is assessed by the value of R^2_{adj}. It is customary to regard an R^2_{adj} of above 75% as very good; 50-75% as good; 25-50% as fair; and below 25% as poor. Multiple regression is also able to determine if the regressors, (taken together), are significantly associated with the dependent variable - assessed by the *F* statistic in ANOVA. A third advantage of utilizing multiple regression in the current analysis, is that the relationship between a single regressor, and the dependent variable can be ascertained while holding all other regressors constant – by examining regression coefficients. The conventional way to compare regression coefficients calculated in differing units is through *z* score transformation. This transformation results in standardized regression coefficients (betas) that are directly comparable. Finally, it is possible to determine whether the relationship of the regressors with the dependent variable, are statistically significant with all other regressors taken into account. The preceding discussion sets out the appropriateness of the regression approach, I now turn to the specific justification for the approach in the context of JAS studies.

JAS has not widely utilized the benefits of a multiple regression approach in the analysis of experimental data. Those studies that have adopted the approach, acknowledge that it is possible to consider variables that potentially explain the percentage variance of some criterion variable, simultaneously (Harvey *et al.* 2000; Azen and Budescu, 2003; Budescu and Azen, 2004). In terms of developing models that can be tested through regression analysis, I follow the suggestion of Cronbach and Petty (1970), and Edwards (1995), who argue that endogenous variables should be utilized as criterion variables that control for other exogenous variables in regression analysis. Put another way, variables such as a, 'judge's initial estimate and the advisor's recommendation are exogenous, whereas the judge's final estimate is endogenous' (Dalal and Bonaccio, 2006, p42). Having established the theoretical justification for the regression approach, I now turn to two issues for which models were developed and tested

through multiple linear regression. First, I consider the issue of the potential predictors of estimation accuracy. This issue is of importance to the central question of this thesis, as where people revise their judgment, neither reasons-based advice, or solely numeric advice appears to be any more persuasive. Hence, it is necessary to consider if there are variables other than advice 'type' that predict estimation accuracy. Second, a logistic regression model is developed and tested that seeks to determine the variables that predict judgment change. Consideration is first given to the rationale for the variables selected to enter into a regression analysis that attempts to predict ultimate estimation accuracy.

In the first instance, it is of interest to ascertain whether participant's ultimate estimation accuracy could be successfully modelled. Clearly, a participant's initial estimation accuracy is largely dependent upon the same individual's capacity to muster their existing knowledge in order to solve the estimation problem at hand. Moreover, it is likely that any subsequent post-advice revision of judgment may be 'anchored' upon this initial intuition (Harvey and Fischer, 1997; Lim and O'Connor, 1995). Hence, ultimate estimation accuracy, will be, at least in part, moderated by an individual's initial unaided intuition. Of interest then, in terms of multiple regression, is to determine (holding other variables constant) the amount of variance in ultimate estimation accuracy accounted for by initial pre-advice estimation accuracy. This will be examined by entering the variable of unaided pre-advice estimation accuracy into the regression analysis. Further, the extent to which available advice either illuminates elements of the estimation problem not originally considered, or merely adds cognitive 'noise', is another factor relevant to ultimate estimation accuracy. Recall that advice in this study is operationalized as Process or Non-Process. This operationalization of advice contains important differences between conditions. Process advice is generated by an individual participant inputting sub-component estimates into an algorithm, and subsequently combining

these sub-components multiplicatively to generate advice. Hence, advice is self-generated by participants, and is different for each participant. In contrast, Non-process advice is constituted as a testimonial assertion from an anonymous advisor, and so is constant for participants in receipt of Non-process advice (in reality advice was generated by the experimenter). The practicalities of this distinction for multiple regression analysis, is that the variable of advice accuracy can only be input into a regression for participants in receipt of Process advice (Non Process advice is invariant and hence is not able to account for variance in ultimate estimation accuracy). Given this limitation, the variable of advice accuracy was input to the regression analysis, as the amount of variance in ultimate estimation accuracy accounted for by the accuracy of advice (while holding other variables constant), is clearly important in ascertaining whether participants act upon advice when formulating a final estimate.

Similarly, post-advice confidence in the veracity of the output of an individual's deliberations may also have a relationship with ultimate estimation accuracy. This variable is of particular interest, as the relationship between ultimate estimation accuracy, and appropriate post-advice confidence is unclear (Rowe, Wright and McColl, 2005). Koriat, Lichtenstein, and Fischhoff, (1980), for example reported that final confidence appears to be related to improvements in ultimate estimation accuracy. However, there is considerable variability in the strength of this relationship due to task demands and testing conditions (Koehler, 1994). Typically, people are overconfident in the veracity of their own judgments, so it is of interest to determine the variance in ultimate estimation accuracy attributable to post-advice confidence (holding other variables constant). Having established the rationale for the selection of variables to enter into a regression analysis that attempts to predict ultimate estimation accuracy, I now turn to the issue of a logistic regression that attempts to predict judgment change.

Logistic regression is a statistical technique designed to acknowledge circumstances where a dependent variable is dichotomous. Logistic regression can be used to predict a dichotomous variable from a set of predictor variables. With a dichotomous categorical dependent variable, logistic regression is an appropriate analytical technique, if the predictor variables are a mix of continuous and categorical variables and/or if they are not normally distributed (logistic regression makes no assumptions about the distributions of the predictor variables). For a logistic regression, the predicted dependent variable is a function of the probability that a particular subject will be in one of two categories (i.e. disease present/disease not present or yes/no etc.). In the context of this study, it is appropriate to use this technique as one dichotomous dependent variable of interest, is judgment revision (i.e. whether participants change judgment, or not).

One variable that is a likely predictor of judgment revision is an individual participant's pre-advice level of confidence in their own intuitive estimate. However, the measure of pre-advice initial confidence may simultaneously capture elements of an individual's pre-existing knowledge of the immediate estimation problem, as well as how confident the same individual is of that domain relevant knowledge. Further, it is possible that elements of a participant's global sense of self-efficacy may also be captured in this measure, which implies that caution should be exercised in interpreting any logistic regression that utilizes this particular measure.

However, despite these reservations, it is reasonable to believe, that at least in part, judgment revision may be mediated by the degree of confidence a participant may hold in an intuitive pre-advice estimate. This is because participant's who report 'wide' confidence intervals (indicative of low levels of pre-advice confidence), maybe more disposed to regard advice as

beneficial, than participants who report 'narrow' confidence intervals (indicative of high pre-advice confidence) (Yaniv and Foster, 1995, 1997). Participants who report high levels of pre-advice confidence, have previously been shown (in Chapter 4) to be less disposed to revise their judgment post-advice, than their less confident counterparts – irrespective of advice quality. Given the differential effects of pre-advice confidence upon judgment revision, the variable of pre-advice confidence was entered into the logistic regression in order to determine the amount of variance in judgment change it accounted for, holding other variables constant.

A second variable of interest that is likely to be influential upon judgment change is pre-advice estimation accuracy. This variable was entered into the logistic regression analysis for the same reasons as it was selected for entry into the multiple regression analysis. Although here, the selection of pre-advice initial estimate accuracy is entered into the logistic regression in order to ascertain the amount of variance in judgment revision attributable to this variable whilst holding other variables constant.

The accuracy of advice is also likely to be influential upon an individual's decision whether to revise their unaided initial intuition, or not (if the same individual is able to successfully evaluate the quality of the proffered advice). This issue is closely related to the measure of perceived utility of advice. Here, participants were asked to rate the quality of the advice available to them. Should participants successfully evaluate the quality of advice in comparison to their own intuitions, then positively evaluated advice is likely to loom large in their deliberations. Alternatively, participants could evaluate advice positively, but be disinclined to incorporate such advice in their deliberations. The accuracy of advice and the perceived utility of advice were entered into the logistic regression, in order to ascertain the

variance in judgment revision attributable to each (holding other variables constant). Prior to considering the results of the multiple regression analysis, and the logistic regression analysis; the data (i.e. residuals) was examined for evidence of multicollinearity and hetereoscedascity, both for Process and Non-Process advice, on a question by question basis.

It is important to consider the issue of multicollinearity when performing regression analysis, to check for undesirable strong correlations between independent variables. Strong correlations between independent variables are undesirable because small changes in data values may lead to large changes in the estimates of regression coefficients. The conventional method for investigating the issue of multicollinearity in regression analysis is to examine residual values. Should these scores *not* be independent, then only limited confidence in the estimated regression coefficients is possible. Eigenvalues for each regression were calculated, and examined for values that approached zero. Eigenvalues that are close to zero indicate the possibility that the independent variables entered into the regression are highly intercorrelated. Similarly, condition indices were calculated for each regression. Condition indices are calculated by taking the square root of the ratio of the largest eigenvalue, to each successive eigenvalue. Condition indices greater than 15 indicate a possible problem of intercorrelation between independent variables, and condition indices greater than 30 indicate a serious problem of intercorrelation between independent variables. On examination of these data, no eigenvalues approached zero, nor did any condition index approach, or exceed 15. Hence, it is possible to infer that no violation of the independence of residuals was found.

Each regression for each question and condition was also checked for evidence of hetereoscedascity. Hetereoscedascity is defined as differences in the variance of the error term

between observations. The presence of hetereoscedascity does not bias the estimates of the coefficients in regression analysis, but it does bias estimates of their standard errors. One way of determining the influence of heteroscedascity in regression analysis is by close visual inspection of residual plots, and where hetereoscedascity is suspected, formally testing for it by performing White's test. Here, a regression of the squares of the residuals is run on the variables suspected of causing the hetereoscedascity, their squares, and cross products. The purpose of the test is to determine whether the absolute value of the residuals can be explained by the original independent variables. This should not occur, as the residual values are supposedly random and non-predictable. Significant hetereoscedascity can be confirmed by comparing the product of R^2 multiplied by the number of participants N, with the appropriate value and degrees of freedom in a χ^2 distribution. Significant hetereoscedascity is detectable when the product of $N*R^2$ exceeds the appropriate test statistic and degrees of freedom in a χ^2 distribution. No significant hetereoscedascity was detected in the when the data was examined at the level of condition, and then at the level of each estimation question.

Table 6.15. Process advice: Multiple regression analysis – ultimate estimation accuracy (*$p < 0.05$)

Question	Experimental condition	Variable	B	SE	Beta	*t*	*p*
1	Process advice	constant	9898051067.96	9591609765.89		1.03	0.32
		LN Final CI	78674011.01	246305934.71	0.07	0.32	0.75
		LN Advice accuracy	173033090.58	100722426.94	0.36	1.72	0.10
		LN E1 accuracy	-743322670.78	542189130.25	-0.29	-1.37	0.19
2	Process advice	constant	-33382591.80	11230869.07		-2.97	0.01*
		LN Final CI	4421207.57	1993804.17	0.32	2.20	0.04*
		LN Advice accuracy	1466487.44	401558.27	0.61	3.65	0.002*
		LN E1 accuracy	871534.72	972875.93	0.15	-0.90	0.38

Question	Experimental condition	Variable	B	SE	Beta	*t*	*p*
3	Process advice	constant	-2252793.96	1292759.76		-1.74	0.10
		LN Final CI	367526.23	96234.74	0.67	3.82	0.001*
		LN Advice accuracy	18211.21	54106.49	0.06	0.34	0.74
		LN E1 accuracy	-98994.00	131293.00	-0.13	-0.75	0.46
4	Process advice	constant	-29621474.52	50768517.63		-0.58	0.57
		LN Final CI	1356821.40	1717896.56	0.19	0.79	0.44
		LN Advice accuracy	440539.48	1285416.87	0.07	0.34	0.74
		LN E1 accuracy	920590.92	3756088.74	0.06	0.24	0.81

Table 6.16. Process advice: Logistic regression – predictors of judgment change.

Question	Condition	Variable	B	SE	Wald	df	*p*	Exp (B)
1	Process advice	LN Pre-advice CI	0.10	0.21	0.23	1	0.63	1.10
		LN E1 accuracy	-0.07	0.89	0.01	1	0.94	0.94
		LN Advice accuracy	0.14	0.23	0.35	1	0.55	1.15
		Advice utility	0.13	0.52	0.06	1	0.81	1.13
		Constant	-3.90	17.31	0.05	1	0.82	0.02
2	Process advice	LN Pre-advice CI	0.006	0.23	0.001	1	0.98	1.01
		LN E1 accuracy	-0.19	0.35	0.32	1	0.57	0.82
		LN Advice accuracy	0.13	0.12	1.07	1	0.30	1.13
		Advice utility	-0.19	0.44	0.19	1	0.67	0.83
		Constant	0.81	3.28	0.06	1	0.81	2.24

Question	Condition	Variable	B	SE	Wald	df	*p*	Exp (B)
3	Process advice	LN Pre-advice CI	0.19	0.30	0.42	1	0.52	1.21
		LN E1 accuracy	0.07	0.37	0.03	1	0.86	1.07
		LN Advice accuracy	0.006	0.11	0.003	1	0.96	1.01
		Advice utility	-1.54	0.67	5.26	1	0.02	0.22
		Constant	2.56	2.96	0.75	1	0.39	12.95
4	Process advice	LN Pre-advice CI	-0.22	0.25	0.79	1	0.37	0.80
		LN E1 accuracy	0.05	0.64	0.007	1	0.93	1.06
		LN Advice accuracy	0.43	0.27	2.54	1	0.11	1.54
		Advice utility	-0.52	0.54	0.91	1	0.34	0.60
		Constant	-4.19	9.41	0.20	1	0.66	0.02

Table 6.17. Non-Process: Multiple regression analysis – ultimate estimation accuracy (*$p < 0.05$)

Question	Experimental condition	Variable	B	SE	Beta	*t*	*p*
1	Non process advice	constant	-7377178.81	36274247.56		-0.20	0.84
		LN Final CI	7652771.79	1635219.97	0.69	4.68	0.0005*
		LN E1 accuracy	1492625.87	2129409.61	0.10	0.70	0.49
2	Non process advice	constant	-116054614.86	24471068.57		-4.74	0.0005*
		LN Final CI	3523710.65	1494781.12	0.30	2.36	0.03*
		LN E1 accuracy	5737532.36	1248859.28	0.58	4.59	0.0005*
3	Non process advice	constant	-182360.42	41042.89		-4.44	0.0005*
		LN Final CI	10664.32	4524.25	0.36	2.36	0.03*
		LN E1 accuracy	13385.07	4178.04	0.49	3.20	0.004*
4	Non process advice	constant	-823806.03	2070767.31		-0.40	0.69
		LN Final CI	185738.04	136893.42	0.34	1.36	0.19
		LN E1 accuracy	-59628.97	152555.38	-0.10	-0.39	0.70

Table 6.18. Non-Process advice: Logistic regression – predictors of judgment change

Question	Condition	Variable	B	SE	Wald	df	*p*	Exp (B)
1	Non Process advice	LN Pre-advice CI	0.72	0.58	1.55	1	0.21	2.05
		LN E1 accuracy	-0.42	1.03	0.17	1	0.68	0.66
		Advice utility	-1.12	0.80	1.97	1	0.16	0.33
		Constant	-1.37	22.21	0.004	1	0.95	0.25
2	Non Process advice	LN Pre-advice CI	-0.22	0.29	0.56	1	0.45	0.80
		LN E1 accuracy	0.36	0.28	1.68	1	0.20	1.43
		Advice utility	-0.66	0.34	3.67	1	0.06	0.52
		Constant	-0.59	3.35	0.03	1	0.86	0.55

Question	Condition	Variable	B	SE	Wald	df	*p*	Exp (B)
3	Non Process advice	LN Pre-advice CI	0.26	0.27	0.90	1	0.34	1.30
		LN E1 accuracy	-0.41	0.33	1.59	1	0.21	0.66
		Advice utility	-0.52	0.39	1.77	1	0.18	0.60
		Constant	2.04	1.94	1.10		0.30	7.66
4	Non Process advice	LN Pre-advice CI	0.24	0.73	0.11	1	0.74	1.27
		LN E1 accuracy	-2.63	1.72	2.34	1	0.13	0.07
		Advice utility	-1.02	0.98	1.09	1	0.30	0.36
		Constant	33.10	28.07	1.39	1	0.24	2.4e+014

The first step in the analysis was to seek to determine the amount of variance in the dependent variable of ultimate estimation accuracy, attributable to the predictor variables of pre-advice initial accuracy, post-advice confidence, and the accuracy of advice. As pooling the data would not be meaningful (due to the differing characteristics of the estimation questions), the analysis was performed for each estimation question separately. The model was a very poor fit for Question 1 ($R^2_{adj} = 0.10$), and there was no significant overall relationship ($F_{(3, 17)} = 1.82$, $p < 0.18$). None of the predictor variables in Question 2 met the assumptions of normality, and so the natural log transformed values were utilized in the regression analysis. The regression was a fair fit ($R^2_{adj} = 0.56$), and the overall relationship was significant ($F_{(3, 19)} = 10.19$, $p < 0.0005$). Inspection of the standardized beta weights, lead to the formulation of a regression equation that predicts a 1% change in ultimate estimation accuracy -

A 1% change in ultimate estimation accuracy = 0.32 LN Final confidence + 0.61 LN Advice accuracy + 0.15 LN E1 Accuracy

Hence, with other variables held constant, ultimate estimation accuracy increased by 0.32%, for every unit decrease in post-advice confidence (this is because 'wide' confidence intervals indicate *less* confidence). Similarly, for every unit decrease in LN Advice accuracy, the percentage change in ultimate estimation accuracy increased by 0.61%, holding other variables constant (this is because values greater than zero indicate increasing inaccuracy). Similarly, for every unit decrease in initial estimation accuracy, the percentage change in ultimate estimation accuracy increased by 0.15, holding other variables constant (this is because values greater than zero indicate increasing inaccuracy). Both LN Final confidence ($t = 2.20$, $df = 19$, $p < 0.04$), and LN Advice accuracy ($t = 3.65$, $df = 19$, $p < 0.002$) were significant predictors of ultimate estimation accuracy.

Similarly, none of the predictor variables in Question 3 met the assumptions of normality, and the natural log transformed values were utilized in the regression analysis. The model was a medium fit ($R^2_{adj} = 0.33$), and there was a significant overall relationship ($F_{(3, 22)} = 5.17$, $p < 0.007$). Examination of the standardized beta weights, lead to the formulation of a regression equation that predicts a 1% change in ultimate estimation accuracy –

1% change in estimation accuracy = 0.67 LN Final Confidence – 0.13 LN Initial Estimate Accuracy + 0.06 LN Advice Accuracy

Hence, with other variables held constant, ultimate estimation accuracy increased by 0.67%, for every unit decrease in LN Final Confidence (this is because values greater than zero indicate increasing inaccuracy). Similarly, for every unit decrease in LN Initial Estimate Accuracy the percentage change in ultimate estimation accuracy increased by 0.13%, holding other variables constant. This finding confirms that the more accurate participants' initial pre-advice responses to Question 3 were, the more accurate these same participants were post-advice. Ultimate estimation accuracy increased by 0.06%, for every unit decrease in LN Advice Accuracy (this is because values greater than zero indicate increasing inaccuracy). Only LN Final confidence was a significant predictor of ultimate estimation accuracy ($t = 3.82$, $df = 22$, $p < 0.001$). However, the regression for the data in Question 4 fitted the model very poorly ($R^2_{adj} = 0.07$) and there was no overall relationship ($F_{(3, 22)} = 0.49$, $p < 0.70$).

Next, consideration was given to the determinants of judgment change for participants in receipt of Process advice at the level of each estimation question (see Table 6.16). Here, a logistic regression was performed on the data with judgment change as the DV and initial participant confidence, initial estimation accuracy, accuracy of advice, and perceived utility of advice as sub-scales of the LSI-R, as predictor variables. A total of 20 cases were analyzed

from Question 1, but the full model was not significantly reliable ($\chi^2 = 1.13$, $df = 4$, $p < 0.89$), although the model fitted the data adequately (Hosmer and Lemeshow $\chi^2 = 9.07$, $df = 8$, $p < 0.34$). This model accounted for between 5.5% and 7.6% of the variance in judgment change, with 92.3% of the participant's who changed judgment post-advice successfully predicted. However, only 14.3% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 65% of predictions were accurate. None of the predictor variables reliably predicted judgment change. Similarly, a total of 23 cases were analyzed for the data in Question 2 - the full model was not significantly reliable ($\chi^2 = 2.47$, $df = 4$, $p < 0.65$). However, the model fitted the data adequately (Hosmer and Lemeshow $\chi^2 = 9.81$, $df = 8$, $p < 0.28$). This model accounted for between 10.2% and 13.6% of the variance in judgment change, with 84.6% of the participant's who changed judgment post-advice successfully predicted. However, only 50% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 69.6% of predictions were accurate. None of the predictor variables reliably predicted judgment change. A total of 25 cases were analyzed from Question 4, but the full model was not significantly reliable for the data ($\chi^2 = 5.77$, $df = 4$, $p < 0.22$), although the model fitted the data adequately (Hosmer and Lemeshow $\chi^2 = 7.37$, $df = 6$, $p < 0.29$). This model accounted for between 20.6% and 30.9% of the variance in judgment change, with 94.7% of the participant's who changed judgment post-advice successfully predicted. However, 33.3% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 80% of predictions were accurate. None of the predictor variables reliably predicted judgment change.

The data from Question 3 was successfully modelled however - a total of 26 cases were analyzed and the full model was significantly reliable ($\chi^2 = 12.74$, $df = 4$, $p < 0.01$). The model also fitted the data adequately (Hosmer and Lemeshow $\chi^2 = 13.61$, $df = 7$, $p < 0.06$). This

model accounted for between 38.7% and 51.7% of the variance in judgment change, with 75% of the participant's who changed judgment post-advice successfully predicted. However, 78.6% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 76.9% of predictions were accurate. Table 6.16 gives coefficients, the Wald statistic and associated degrees of freedom and probability values for each of the predictor variables.

Inspection of Table 6.16 shows that only perceived utility of advice reliably predicts judgment change for the data in Question 3. The values of the coefficients reveal that a decrease in 1 unit of perceived utility of advice is associated with an increase in the odds of judgment change by a factor of 0.22. This appears counterintuitive in so far as it might be expected that participants who rate advice highly might be expected to be more disposed to revise their judgments post-advice, than participants who rate advice less highly. However, this result could be explained by the idea that one of the determinants of judgment change might be the similarity between a participant's pre-advice judgment and advice. If so, participants might rate advice that is similar to pre-advice judgments highly, regarding such advice as confirmation of participants' pre-advice judgments; implying that participants are *less* likely to revise their judgments post-advice. In contrast, advice that is rated less highly, may be divergent from a participant's pre-advice initial judgment, and subsequently motivate the same participant to re-consider their judgment post-advice. This idea was tested by calculating the correlation between judgment change and perceived utility of advice (R_s = - 0.37, $p < 0.03$, N = 35). This finding supports the idea that advice which is rated highly is less divergent from a participant's pre-advice initial judgment, than advice that is rated less highly. The correlation coefficient is negative, indicating that participants who did not change judgment rated advice more highly than people who did change judgment post-advice.

I next considered the data from participants in receipt of Non-Process advice, here a multiple linear regression was performed where ultimate estimation accuracy was the dependent measure and pre-advice initial accuracy and post-advice confidence, were predictor variables. The regression was a fair fit for the data in Question 1 (R^2_{adj} = 0.47), and the overall relationship was significant ($F_{(2, 24)}$ = 12.61, $p < 0.0005$). Inspection of the standardized beta weights, lead to the formulation of a regression equation that predicts a 1% change in ultimate estimation accuracy -

1% change in ultimate estimation accuracy = 0.69 LN post-advice confidence + 0.10 LN initial accuracy

Hence, with other variables held constant, ultimate estimation accuracy increased by 0.69%, for every unit decrease in post-advice confidence (this is because 'wide' confidence intervals indicate *less* confidence). Similarly, for every unit decrease in initial estimation accuracy, the percentage change in ultimate estimation accuracy increased by 0.10%, holding other variables constant (this is because values greater than zero indicate increasing inaccuracy). Only LN post-advice confidence (t = 4.68, df = 24, $p < 0.0005$), was a significant predictor of ultimate estimation accuracy. Consideration was next given to participants in receipt of Non-Process advice and responding to Question 2. The model was a fair fit (R^2_{adj} = 0.51), and the overall relationship was significant. ($F_{(2, 32)}$ = 18.59, $p < 0.0005$). Examination of the standardized beta weights, lead to the formulation of a regression equation that predicts a 1% change in ultimate estimation accuracy -

1% change in ultimate estimation accuracy = 0.30 LN pre-advice initial accuracy + 0.58 LN Final Confidence

Hence, holding other variables constant ultimate estimation accuracy increased by 0.30%, for every unit decrease in LN pre-advice initial accuracy (this is because values greater than zero indicate increasing inaccuracy). Similarly, for every unit decrease in LN Final Confidence the percentage change in ultimate estimation accuracy increased by 0.58%, holding other variables constant (this is because 'wide' confidence intervals indicate *less* confidence). Both LN Final confidence ($t = 2.36$, $df = 32$, $p < 0.03$), and LN pre-advice initial accuracy ($t = 4.59$, $df = 32$, $p < 0.0005$) were significant predictors of ultimate estimation accuracy.

The data from participants in receipt of Non-Process advice, and responding to Question 3 was examined next. The model was a fair fit ($R^2_{\text{Adj}} = 0.48$), and the overall relationship was significant ($F_{(2, 25)} = 13.52$, $p < 0.0005$). Examination of the standardized beta weights, lead to the formulation of a regression equation that predicts a 1% change in ultimate estimation accuracy –

1% change in ultimate estimation accuracy = 0.36 LN Final Confidence + 0.49 LN Initial estimate accuracy

Hence, for every unit decrease in LN Initial Estimate Accuracy the percentage change in ultimate estimation accuracy increased by 0.49%, holding other variables constant. Similarly, with other variables held constant, ultimate estimation accuracy increased by 0.36%, for every unit decrease in LN Final Confidence. This finding confirms that the more accurate participants' initial pre-advice responses to Question 3 were, the more accurate these same participants were post-advice. Both LN Final confidence ($t = 2.36$, $df = 25$, $p < 0.03$) and LN Initial estimate accuracy ($t = 3.20$, $df = 25$, $p < 0.004$) were significant predictors of ultimate estimation accuracy. Finally, the data from participants in receipt of Non-Process advice, and

responding to Question 4 was examined. The model was a very poor fit ($R^2_{adj} = 0.07$), and there was no significant overall relationship ($F_{(2, 18)} = 0.95$, $p < 0.40$).

In order to ascertain the determinants of judgment change for participants in receipt of Non-Process advice, a logistic regression was performed with judgment change as the DV and initial participant confidence, initial estimation accuracy, and perceived utility of advice as sub-scales of the LSI-R, as predictor variables. A total of 27 cases were analyzed from Question 1, but the full model was not significantly reliable ($\chi^2 = 6.17$, $df = 3$, $p < 0.10$), despite fitting the data adequately (Hosmer and Lemeshow $\chi^2 = 1.50$, $df = 7$, $p < 0.98$). This model accounted for between 20.4% and 33.1% of the variance in judgment change, with 96% of the participant's who changed judgment post-advice successfully predicted. However, only 20% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 81.5% of predictions were accurate. None of the predictor variables reliably predicted judgment change. A total of 35 cases were analyzed in Question 2, but the full model was not significantly reliable ($\chi^2 = 5.97$, $df = 3$, $p < 0.11$). Further, the model did not adequately fit the data (Hosmer and Lemeshow $\chi^2 = 12.64$, $df = 7$, $p < 0.08$). This model accounted for between 15.7% and 21.4% of the variance in judgment change, with 86.4% of the participant's who changed judgment post-advice successfully predicted. However, only 53.8% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 74.3% of predictions were accurate. None of the predictor variables entered into the regression proved to be significant. Similarly, the data from Question 3 was not modelled successfully - a total of 33 cases were analyzed ($\chi^2 = 5.60$, $df = 3$, $p < 0.13$). However, the model fitted the data adequately (Hosmer and Lemeshow $\chi^2 = 15.21$, $df = 8$, $p < 0.06$). This model accounted for between 15.6% and 22.1% of the variance in judgment change, with 100% of the participant's who changed judgment post-advice successfully predicted. However, 40% of predictions for the group of participants who did not change

judgment post-advice were accurate. Overall, 81.8% of predictions were accurate. None of the predictor variables significantly predicted judgment change. An identical pattern of results emerged from the logistic regression performed on the data from Question 4 (a total of 23 cases were analysed) the model was not reliable ($\chi^2 = 4.24$, $df = 3$, $p < 0.24$). The model fitted the data adequately however (Hosmer and Lemeshow $\chi^2 = 5.46$, $df = 8$, $p < 0.71$), and this model accounted for between 16.8% and 37.8% of the variance in judgment change, with 95.2% of the participant's who changed judgment post-advice successfully predicted. However, 50% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 91.3% of predictions were accurate. None of the predictor variables reliably predicted judgment change. Table 6.19 summarizes the significant predictors that emerged from the preceding regression analyses.

Table 6.19 Summary of significant predictors from regression analyses

Advice	Criterion variable	Question	Predictor	*t*	*df*	*p*
Process	Ultimate estimation accuracy	2	LN Final confidence	2.20	19	0.04
		2	LN Advice accuracy	3.65	19	0.002
		3	LN Final confidence	3.82	22	0.001
Non-Process	Ultimate estimation accuracy	1	LN post-advice confidence	4.68	24	0.0005
		2	LN Final confidence	2.36	32	0.03
		2	LN pre-advice initial accuracy	4.59	32	0.0005
		3	LN Final confidence	2.36	25	0.03
		3	LN Initial estimate accuracy	3.20	25	0.004

Advice	Criterion variable	Question	Predictor	B	SE	Wald	df	*p*	Exp (B)
Process	Judgment change	3	Advice utility	-1.54	0.67	5.26	1	0.02	0.22

In sum, the analyses of data provided by participants in receipt of Process advice showed that levels of confidence post-advice were influential in predicting improvements in ultimate estimation accuracy for Question 2 and Question 3. However, final confidence explained more variance in ultimate estimation accuracy only in Question 3. The analyses of data from participants responding to Question 2 showed that the accuracy of Process advice was more important than final confidence in Question 2. This may indicate the mediating effect of perceived differences between the two estimation tasks (recall that only participants responding to Question 3 reported differences in the perceived utility of advice scores between participants who revised their judgment post-advice, and those who did not). Further, there

were no differences in pre-advice initial confidence scores between those who subsequently revised their judgment post-advice, and those who did not, for participants in both Question 2 and Question 3. This may suggest that the operationalization of measures of confidence in this study are 'noisy', and may have captured elements of global self-efficacy, and task specific competence, in addition to confidence in the veracity of participants ultimate estimates.

Similarly, the analyses of data from participants in receipt of Non-Process advice showed the influence of post-advice confidence in accounting for the variation in ultimate estimation accuracy. This variable was most influential upon ultimate estimation accuracy in Questions 1 and 2. However, the data in Question 3 showed that although final confidence was a significant predictor of ultimate estimation accuracy, initial estimation accuracy was a more important predictor of ultimate accuracy. These findings may not be contradictory, if one accepts that there are perceived differences of task characteristics between questions, and individual differences between participants. Overall, final confidence appears to be related to improvements in ultimate estimation accuracy. This finding is not inconsistent with the work of Koriat, Lichtenstein, and Fischhoff (1980), who found a similar relationship between self-assessed confidence and accuracy; but as has been found here, there is considerable variability in the strength of this relationship due to task demands and testing conditions (Koehler, 1994).

Little evidence has been found however, for the determinants of judgment change. Only analyses of the data of participants in receipt of Process advice and responding to Question 3, suggested that the perceived utility of advice was influential in post-advice judgment revision. Whilst in isolation this finding is consistent with a conformity to advice account (if an individual views advice favourably, and rates it positively, it is not unreasonable to predict that

the same individual may revise their judgment post-advice, if their own unaided intuition seems weak in comparison to the proffered advice), it is caveated by the differences in experimental tasks that have not been controlled for here.

6.4 Discussion

Within the constraints of the four estimation tasks completed by participants, the preceding analysis shows that Process advice is no more effective in facilitating judgment change, than Non-process advice, either in terms of the number of appropriate judgment revisions, or the extent of judgment revision. Further, the measure of judgment change relative to advice indicated that participants – when they did revise their judgment post-advice – did so by reducing the absolute difference between their initial unaided estimate and advice, post-advice. Both Process, and Non-Process advice was valued equally by participants, and there were no significant differences in pre-existing knowledge (as determined by the width of pre-advice confidence bounds), held by participants in each condition. These findings suggest that participants who revised their judgment post-advice were indiscriminately conforming to advice. Subsequently, participants' data was modelled through multiple linear regressions, in order to predict ultimate estimation accuracy for each estimation problem, in both the Process and Non-Process advice conditions. Further, the determinants of judgment change were modelled in a logistic regression for each condition, and each estimation question.

A preliminary inspection of participants' data in this study revealed that people tend to make significantly larger estimates of a target quantity, after working through an algorithmic decomposition, than they would otherwise make. MacGregor *et al.* (1988, 1991), argue that such an outcome reflects the positive correlation between the errors of the component parts of

a poorly motivated algorithm. However, without knowing the beneficial attributes of an algorithmic decomposition *a priori* (if this were the case it would be a simple matter to choose between holistic estimation and some form of decomposition), people may have found it difficult to assess the accuracy of advice (output-of-algorithm). Henrion *et al.* (1993) compared the accuracy of estimates of subjective probability distributions produced by (i) people generating their own decompositions and (ii) people working through algorithmic decompositions provided by an experimenter. They found no differences in the accuracy of the outputs of these two conditions. Instead, Henrion *et al.* found that algorithmic decomposition 'inflated' people's estimates, in comparison to a control condition. Hence, people who underestimated a target quantity initially, and persisted in underestimating (to a lesser degree) a target quantity after working through an algorithmic decomposition can be expected to improve their estimation accuracy. Conversely, people who overestimate a target quantity, initially, can expect to diminish their estimation accuracy after working through an algorithmic decomposition. Hence, beneficial judgment revision can be accounted for by the nature of the structural exchange of numerical information in JAS-type tasks. Similar findings are presented here, and some possible cognitive factors are discussed below.

The data from participants in receipt of advice in this study was modelled to ascertain (i) the determinants of ultimate estimation accuracy (ii) the determinants of judgment change. As Process advice was generated by participants (and was different for each participant and each question), whilst Non-Process advice was constituted by a testimonial assertion by a single anonymous advisor, each advice type was analyzed separately. The data from participants in receipt of Process advice, and responding to Question 2 (an estimate of Higher Education enrolments in the UK 2004/5), and Question 3 (an estimate of the number of fatalities resulting from a road traffic accident in the UK in 2004) could be successfully modelled. Here, a

participant's post-advice confidence, in conjunction with the accuracy of the available advice, determined the same participant's ultimate accuracy, when estimating the number of Higher Education enrolments in the UK 2004/5. The ultimate accuracy of participants estimating the number of fatalities resulting from a road traffic accident in the UK, in 2004, was determined by the same participants' post-advice level of confidence. Significant models for ultimate estimation accuracy were reported for participants in receipt of Non-Process advice, and responding to Question 1 (an estimate of the number of Lottery tickets sold in the first week of June 2006 in the UK), Question 2 (HE enrolments in the UK 2004/5), and Question 3 (an estimate of the number of fatalities resulting from traffic accidents). For Questions 1 a participant's post-advice confidence determines the same participant's ultimate accuracy. The data from participants responding to Question 2 and 3 indicates that, the accuracy of participant's pre-advice unaided estimate, in conjunction with the same participants' post-advice confidence, determines ultimate accuracy. The determinants of judgment change were only successfully modelled through logistic regression for participants in receipt of Process advice and responding to Question 3 (an estimate of the number of fatalities resulting from traffic accidents). Here, participants' perceived utility of advice was a significant predictor of judgment change. However, the correlation between perceived utility of advice and judgment change indicated that participant's attributed higher utility ratings to advice that was *less* divergent from participants' pre-advice initial judgments.

Clearly, the success (or otherwise) of this model building exercise is sensitive to the characteristics of each different estimation task. Further, the mode of response (online/paper) influences participants' self-reported levels of confidence. In the aggregate, there were no reported significant differences, in the widths of pre-advice initial levels of confidence. Here, it is inferred that this reflects broadly similar levels of pre-existing task relevant knowledge,

amongst participants randomly allocated between conditions. However, in three out of four estimation questions participants responding online reported higher levels of pre-advice initial confidence than participants responding on paper. One explanation for this apparent discrepancy might be that people responding online may have felt greater self-efficacy in their deliberations due to their IT skills. Hence, participants may have held broadly similar degrees of task relevant knowledge, whilst simultaneously experiencing enhanced self-efficacy when responding online.

The findings reported in this study show that, in conditions where people make estimates of highly uncertain quantities, they are unable to either spontaneously evaluate advice of a holistic nature, or, indeed, use an algorithmic decomposition to mechanically combine sub-component estimates into a useful estimate of the quantity in question. This may imply that in such circumstances people are more likely to utilize peripheral environmental cues concerning the source of advice, rather than invest cognitive resources, in what potentially maybe a fruitless effort, to evaluate the quality of advice on offer (in conditions of less uncertainty it is possible that people may be able to successfully scrutinize the quality of advice). Moreover, where people have little or no task relevant knowledge, and are unable to input more accurate sub-component estimates into an algorithmic decomposition than what may be achievable by a global holistic evaluation of an estimation problem, then algorithmic decomposition may not be of any greater utility than holistic estimation.

Further, people will, often as not, follow advice under conditions of uncertainty, irrespective of the quality of such advice. This is because where people have little, or no, task relevant knowledge it is difficult to evaluate the quality of advice. Also, evaluations of advice based upon ‘world knowledge’ are insufficient for the purposes of distinguishing between poor and

good advice. Lichtenstein (1990) makes a similar point in reporting the results of her second study, where she argues that algorithmic decomposition is of little utility where it is (i) poorly understood, and (ii) 'noisy' due to the poor quality information generated by participants, coupled with a lack of arithmetic skills - which are necessary to combine the various component steps of the algorithm. She goes on to argue that the conditions for the successful use of algorithmic decomposition are likely to involve the careful and deliberate design of algorithms, and various computational aids for participants.

The preceding discussion is difficult to reconcile with the idea that Process advice relaxes cognitive demands upon people and facilitates rigorous consideration of each attribute of complex estimation problems (Fischer, 1977; Kleinmuntz 1990; MacGregor, 1988, 1991, 2001). Indeed, the benefits of a 'divide and conquer' strategy (Henrion *et al.* 1993; Kleinmuntz *et al.* 1996; Morera and Budescu, 1998), in the four estimation tasks participants in this study responded to, appear to be offset by the difficulties participants encountered when attempting to make accurate sub-component estimates in the Process advice condition. Recall that if sub-component estimate errors are positively correlated, then the output of an algorithmic decomposition is likely to be highly inaccurate. In such circumstances a holistic estimate of some troublesome quantity may be no less effective, than algorithmic decomposition (and hence one strategy may appear no more valuable to participants than another). Such findings are consistent with the work of Lichtenstein and Weathers (1998), who found that, even when people were trained to create their own algorithmic decompositions (to estimate uncertain quantities), performance was not improved to any great degree. Moreover, these authors also found that on some occasions the use of algorithmic decomposition could inhibit accurate estimation (Study 2).

The discussion so far has highlighted some of the constraints upon the decompositional approach to estimation, in comparison to a holistic approach. However, the results reported in this study are limited in several important respects. Firstly, the formulation of Process advice in this study may possibly have been weak. The algorithms used were not known *a priori* to be effective in facilitating judgment change (beneficial or otherwise). The formulation of advice utilized by MacGregor *et al.* (1988, 1991), however *was* known to be effective in facilitating beneficial judgment revision *a priori*. However, for algorithmic decomposition to retain some general utility, it is unlikely that estimators will have a repertoire of accessible 'good' algorithms. Perhaps a more significant point is that Process advice in this study, was generated by individual participants, and hence was different for each person and each question. In contrast, Non-Process advice was not generated by participants, and hence was identical for each question, and for each participant. This implies that direct comparisons between the two advice conditions, is imperfect (for reasons discussed previously).

A further limitation of the study was that in the Process advice condition, all four algorithms produced advice values that were hugely discrepant from participants' initial estimates. This may have lead participants to believe that the advice was of little use, and hence the process that produced such unbelievable advice was equally worthless. Subsequently, participants may not have been motivated to address the procedure with sufficient seriousness. The preceding discussion implies that future studies should carefully consider the issues of internal validity when attempting to formulate meaningful algorithms in studies of this type. A further problem arises when reflecting upon the characteristics of the four estimation problems for which algorithms were constructed. Although there was no *a priori* reason to suppose that participants would find one estimation problem any more difficult than another, drawing upon tasks from multiple task domains may risk undermining my participant's motivation to

complete the tasks. Similarly, not fully controlling for task difficulty and utilizing obscure tasks from multiple domains leaves such experimental work open to criticisms on the grounds of ecological validity.

In sum, the preceding discussion has outlined circumstances in which algorithmic decomposition is no more effective in facilitating judgment change than a holistic evaluation of a testimonial assertion, and where conformity to advice is observed. Ultimately, the most effective determinant of judgment change appears to be the perceptions people hold, of the utility of advice, as an aid in solving an estimation problem. One possible direction for further research would be to test algorithms that are known *a priori* to be effective estimation aids, to ascertain whether such algorithms were also (i) effective in facilitating judgment change (ii) valued highly by respondents.

Chapter 7: Knowledge of advice generation with algorithms that are known to be effective.

7.1 Introduction

The preceding study investigated the possibility that intrinsic information about the quality of advice could influence a participant's decision of whether judgment change is appropriate and beneficial. Ultimately, in the context of the estimation tasks that participants completed, knowledge of the process by which advice is generated proved no more influential than 'holistic' advice (constituted as a testimonial assertion) upon judgment revision, instead conformity to advice was observed. This finding was surprising, given that MacGregor *et al.* (1988, 1991), argue the case for the potential effectiveness of algorithmic decomposition (in estimating uncertain quantities), in comparison to holistic estimation. However, these authors were solely interested in the potential benefits of algorithmic decomposition, in terms of the ultimate accuracy of the output-of-algorithm; whereas the issue of interest in the previous study was how knowledge of the process by which advice is generated could result in beneficial judgment revision. Moreover, the algorithms utilized in the previous study may have been weak (their effectiveness was not known *a priori*), in comparison to the algorithms utilized by MacGregor *et al.* Further, the previous study may not have fully acknowledged the constraints that MacGregor (2001) outlined as prerequisites for the appropriateness of the algorithmic approach in estimating uncertain quantities. Uncertainty in this context can be defined as a participant's lack of domain relevant knowledge (as opposed to some calculation of risk). Given these limitations, it seemed appropriate to replicate MacGregor *et al.* and analyze the data in terms of beneficial judgment change, instead of ultimate estimation accuracy. Such a formulation of the issue would pitch two 'good' (i.e. known to be effective (*a priori*) algorithms against each other in terms of beneficial judgment change. Should

conformity to advice obtain, then participants will reduce the absolute difference between their initial intuitive estimate and advice, post-advice – irrespective of experimental condition.

MacGregor *et al.* (1988) formulated a number of algorithmic decompositions – the 'US Mail' problem, and the 'Number of Forested Miles in the US state of Oregon' problem, being two of them. These authors have argued that participants responding to the 'Forested Miles' problem, do not show the same improvements in estimation accuracy as participants in the 'US mail' problem, as participants in the former condition, initially gave more accurate unaided estimates, than participants in receipt of the US mail problem. Hence, there was less opportunity for participants in the 'Forested Miles' to improve their estimation accuracy. However, the two algorithms were not equivalent in the number of sub-component steps available to participants – the US Mail algorithm comprised of eight sub-component steps, whilst the Forested Miles problem comprised five sub-component steps. This may suggest that differences in the complexity of knowledge of the process of advice generation may interact with the characteristics of the estimation task, leading participants to different levels of estimation accuracy (and judgment change). Moreover, there are further reasons for suspecting that the results reported by MacGregor *et al.* (1988, 1991), can be interpreted by an alternative account. These authors drew their participants from the University of Oregon, and hence, it is not impossible that their participants were able to bring 'local knowledge' to bear in forming an unaided estimated, of the Forested Miles in the state of Oregon. It follows then, that, algorithmic decomposition is unlikely to offer greater improvements in estimation accuracy, than unaided estimation in these circumstances.

Replication of MacGregor *et al*

To control for the possibility that participants' in MacGregor *et al.* may have been able to bring 'local knowledge' to bear upon the estimation tasks at issue, MacGregor *et al.* (1988, 1991), was replicated using participants drawn from students and staff at the University of Durham (UK). It is likely that UK participants would have no more pre-existing knowledge about the numbers of pieces of mail handled by the US postal service in 1987, than the area of the US state of Oregon that is forested, and hence, improvements in estimation accuracy (and judgment change) through knowledge of the process by which advice is generated, might be expected to be broadly similar. Moreover, the preceding study of algorithmic decomposition has shown that, at least in part, any improvements in estimation accuracy are artifactual (i.e. people tend to follow any plausible advice in conditions of uncertainty); and it is of interest to determine whether similar patterns of results obtain in a replication of MacGregor *et al.* Further, it is arguable that differences between the two algorithmic decompositions used in MacGregor *et al.* may account for the lack of improvement in final estimation accuracy for participants responding to the 'Forested Miles' problem. Specifically, that the number of component 'steps', or sub-estimates, were greater in the 'US mail' problem than the 'Forested Miles' problem, and hence participants in receipt of the 'Forested Miles' problem had less of an opportunity, to input accurate information, that may have lead to improvements in estimation accuracy. If this premise is accepted, then MacGregor *et al.* cannot account for the lack of improvement in estimation accuracy in the 'Forested Miles' problem in these circumstances. Further, respondents in MacGregor *et al.* consistently underestimated the true answer to the 'Forested Miles' problem, and the 'US mail' problem, hence post-algorithm estimate 'inflation' can potentially explain accuracy improvements without recourse to explanations based upon decomposition.

The current study is not an absolute replication of MacGregor *et al.* however. These authors did not adequately control for the possibility that their participants may have held differing perceptions of the utility of each of the two algorithms. Subsequently, it was impossible to determine if the accuracy improvements reported for participants in receipt of the US Mail problem, are entirely artifactual, or that these participants recognized, and valued, the utility of the provided algorithm. This was controlled for in the current study, by asking participants to rate the utility of the algorithm given to them, and hence it was possible to subsequently test, whether there were any differences between the ratings participants attributed to the US Mail algorithm, and the ratings participant's attributed to the 'Forested Miles' problem. In sum, the purpose of this partial replication is to test whether people conform to any available advice, are sensitive to minor differences in the characteristics of the estimation task, the degree of pre-existing knowledge available to an estimator, and the domain from which the estimation problem is drawn; but relatively insensitive to the quality of any advice that may be considered in forming a definitive estimate of some uncertain quantity.

7.2 Method

Participants

Participants were drawn from staff and students at Durham Business School, University of Durham (UK), (N = 105). Members of staff were recruited by a personal invitation by the author to participate in the study, however the majority of participants were students who the author recruited by visiting a lecture theatre (with the permission of staff colleagues) where students were gathered to attend their studies. All participants took approximately the same amount of time to complete the questionnaire. Participation was entirely voluntary, and a verbal instruction by the author that participants should write their age and gender on the first page of the questionnaire, was given to all participants before they started the questionnaire

and at the completion of the questionnaire prior to collection by the author. Verbally instructing participants to provide demographic data, instead of providing them with clearly marked text boxes in which to provide this information on the first page of the questionnaire, may have reduced the rate of response as 85 participants provided demographic data (20 participants did not, or chose not, to provide demographic data) - 45 males participated (age $\bar{x}$ = 24.33 years, *sd* = 3.05), as did 40 females (age $\bar{x}$ = 23.20 years, *sd* = 1.64).

Stimuli and Materials

Participants completed either the 'US Mail' estimation problem, or the 'Number of Forested Miles in Oregon problem' questionnaire booklets (not both) (see Appendix IV). In each questionnaire booklet for both estimation problems an identical opening paragraph of text informed participants that the purpose of the questionnaire was to 'determine how you go about estimating an uncertain quantity'. However, in reality the true purpose of the experiment was to ascertain the potentially beneficial effects of knowledge of the process of advice generation upon judgment revision and accuracy. Here, knowledge of the process of advice generation was operationalized as step-by-step algorithms of known effectiveness for solving estimation problems. On the first page of the questionnaire booklet, participants provided an estimate to one of the estimation problems. Subsequently, participants were tasked with providing a lower, and upper bound, within which they were 90% certain that their unaided estimate and the true answer (whatever that might be) fell. Before participants provided this information they were provided with a visual example of how they should enter the lower and upper bounds of the 90% confidence interval (see Appendix IV).

On the second page of the questionnaire booklet, participants either received a distractor task, the 18-item 'Need for Cognition (NFC) scale', (Petty, Kao, and Caccioppo, 1982), or advice.

The NFC scale is an instrument designed to differentiate between people on the grounds of focussed or peripheral thinking styles. Participants circle quantitative values (ranging from 1- extremely uncharacteristic of you; 5 - extremely characteristic of you) in response to questions such as 'Thinking is not my idea of fun' that characterise their own thinking style. The use of the NFC scale here is limited to merely ensure that participants not in receipt of advice spend approximately the same amount of time thinking about the estimation problem, as those participants in receipt of advice. Participants who received a distractor task in their questionnaire booklet were invited to reconsider their unaided estimate to the estimation problem on completion of the distractor task, since they had 'had an opportunity to think about the issue' (see Appendix IV). These participants could then choose to revise, or not revise their unaided estimate, and provide a 90% confidence interval within which their estimate and the true answer fell on the final page of the questionnaire booklet. Where people received advice, it was provided by an anonymous advisor 'Kate' or 'Joe', whose advice consisted of a step-by-step algorithm that if followed, resulted in a single numeric value. Participants then rated the utility of this advice on a Likert-type scale, before deciding whether to revise, or not revise, their own unaided judgment, and give an upper and lower bound, within which a participant was 90% certain that their revised estimate, and the true value fell.

Procedure

Participants were randomly assigned to either one of two experimental conditions, and tasked with completing either the US mail estimation problem, or the Forested Miles problem (not both). Where participants were assigned to complete the US mail estimation problem, one group was provided with the algorithmic decomposition (N = 24) presented by MacGregor *et al.* Another group of participants in receipt of the US mail problem were tasked with completing the estimation problem unaided (N = 30), before completing a distracter task - the

short form of the Need-for-Cognition scale (Petty, Kao and Cacioppo, 1982). No participant was permitted to be in both the 'algorithm' group and the 'distractor task' group. Subsequently, all participants were invited to reconsider their estimate of the number of pieces of mail that the US postal service handled in 1987. An equivalent procedure was followed for the number of Forested Miles in Oregon estimation problem, where one group of participants received the algorithmic decomposition (N = 25), and a second group completed the task unaided (N = 26), before completing the distracter task, and then given the opportunity to revise their estimate.

7.3 Results

Prior to beginning the analysis of the data provided participants, it is useful to consider the rate of responses to the estimation tasks participants were asked to complete. Inspection of Table 7.1 shows that there was a high and satisfactory rate of response.

Table 7.1 Number of responses

Participants N = 105	Unaided estimate	Pre-advice Lower Bound	Pre-advice Upper Bound	Post-advice estimate	Post-advice Lower Bound	Post-advice Upper Bound
Responses N	104/105 (99.05%)	101/105 (96.19%)	101/105 (96.19%)	105/105 (100%)	104/105 (99.05%)	102/105 (97.14%)

The distribution of data points for participants' unaided estimates, and post-algorithm estimates, proved to be skewed. Therefore, in order to meaningfully analyze the data, it was necessary to Log transform values where appropriate. In the first instance, participants' data was aggregated and tabulated (see Table 7.2). Table 7.2 indicates that the advice produced by both the 'US Mail' algorithm, and the 'Forested Miles' algorithm, was numerically greater than participants' pre-advice estimates, on average. Subsequently, participants' appear to have

revised their estimates so that the absolute numerical difference between their post-advice estimate, and advice, is smaller than the absolute numerical difference between their pre-advice estimate, and advice – which is consistent with the conformity to advice account of judgment change. The issue of conformity to advice was further explored by analyzing both the *amount* (number of judgment revisions), and the *extent* of judgment revision (absolute difference scores). Further, a measure of judgment change relative to advice was calculated for each participant in each of the experimental conditions. Positive scores, on average, would indicate that where people did revise their judgment, post-advice, such judgment revision was in the 'direction' of the available advice, irrespective of the quality of advice. The perceived utility of each of the algorithms presented to participants was a further measure of the conformity to advice account of these data. Should participants report that they valued each algorithm equally, on average, then it is possible to infer, that people were constrained in their abilities to assess advice quality, in these particular estimation tasks.

Further analyses were carried out to ascertain if similar advice to a participant's pre-advice initial estimate was influential in terms of judgment change. Ultimately, the analysis concludes with testing potential explanatory models through regression (based on the same rationale set out in Chapter 6, p217-22). The regression modelling approach is appropriate as it allows (i) the determination of the percentage variance in a criterion variable, based upon the simultaneous consideration of multiple predictor variables, and (ii) where a criterion variable such as judgment change (yes//no) is dichotomous, logistic regression is an appropriate procedure for the simultaneous consideration of multiple predictor variables.

First, the aggregated data for participants' unaided estimates, advice, and post-advice estimates was inspected (see Table 7.2 below).

Table 7.2 Participants in receipt of advice – pre-advice estimate, advice, post-advice estimate (1,000,000s)

Estimation problem	Pre-advice estimate $\bar{x}$	*sd*	Median	Advice $\bar{x}$	*sd*	Median	Post-advice estimate $\bar{x}$	*sd*	Median	True answer
US Mail (*N* = 24)	943.97	2394.33	50.00	9,634,253.00	46,540,000.00	5053.43	2,221.88	3702.54	257.50	89,000.00
Forested Miles (*N* = 25)	5.84	13.08	0.10	9,000,000.00	41,960,000.00	0.07	8,400,004.00	42,000,000.00	0.08	0.47506

These preliminary findings are consistent with the idea that people generally under- estimate the size of large quantities, but overestimate the size of small quantities (Lichtenstein *et al.* 1982). Here, the algorithmic decomposition of the US mail problem acted to significantly inflate the median value of participants' estimates, post-algorithm (z = - 2.43, N – Ties = 12, p < 0.015, two-tailed). As the median value of participants' unaided estimates was less than the median value of participants' post algorithm estimates, estimate 'inflation' can account for the improvement in accuracy reported by participants in receipt of the algorithmic decomposition of the US mail estimation problem. However, when the estimates participants made in response to the Forested Miles estimation problem were considered, no significant differences in median values were found, pre to post algorithm (z = 0.71, N – Ties = 9, p < 0.48). The issue of the similarity between pre-advice initial estimate, and advice as a factor in judgment change will be returned to later in the analysis.

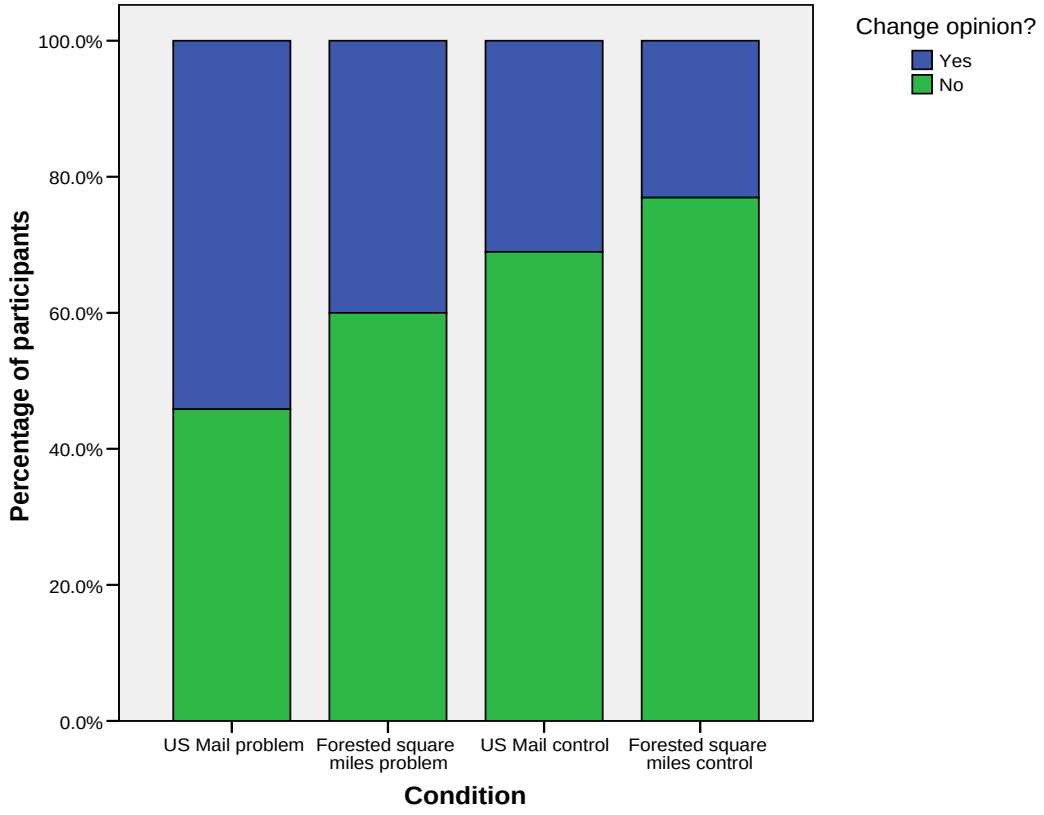
Consideration was next given to the number of judgment revisions made by participants. This analysis is relevant to the conformity to advice account as should people indiscriminately conform to advice, then it might be expected that no differences in the number of judgment revisions between experimental conditions would be observed. Initially, these data were aggregated and tabulated (see Table 7.3).

Table 7.3 Number of judgment changes

	Change judgment	Do not change judgment
US Mail problem	13 (54.2%)	11 (45.8%)
US Mail controls	10 (33.3%)	20 (66.7%)
Forested Miles problem	10 (40%)	15 (60%)
Forested Miles controls	6 (23.1%)	20 (76.9%)

Table 7.3 shows that people *not* in receipt of an algorithmic decomposition change judgment less often than people who worked through the algorithmic decompositions, on average. However, people in receipt of the US Mail algorithm, changed judgment (more often than not), on average; whilst people in receipt of the Forested Miles algorithm did not change judgment (more often than not), on average. These findings are illustrated in Fig. 7.1.

Fig 7.1 Proportion of judgment revisions in each condition



A significance test was performed in order to ascertain whether people in receipt of the US Mail algorithm changed judgment more often, than would be expected by chance, and this proved to be non-significant (χ^2 = 0.17, df = 1, p < 0.68). However, people not in receipt of the US Mail problem did revise their judgment post-advice significantly less often than people who did revise their judgment (χ^2 = 4.17, df = 1, p < 0.04). The same procedure was carried out, on the data from people in receipt of the Forested Miles algorithm, and again, the result

was non-significant – people in receipt of this algorithm were just as likely not to revise their judgment post-advice ($\chi^2 = 1.0$, $df = 1$, $p < 0.32$). However, where people did not receive the Forested Miles algorithm more people were predisposed not to revise their judgments post-advice ($\chi^2 = 7.54$, $df = 1$, $p < 0.06$). These findings suggest that the US Mail algorithm was no more effective in facilitating judgment change, than the Forested Miles algorithm. Further, it is possible to infer that participants in receipt of both the US Mail algorithm and the Forested Miles algorithm appear to be just as likely not to revise their unaided estimate as to revise their unaided estimate, after considering the output of the algorithm. In so far as these findings are consistent with a conformity to advice account, it is only possible to say that the provision of purportedly beneficial algorithms marginally tips the balance of the chance of a participant revising their judgment post-advice from probably not likely, on average, to as likely to as not, where algorithms are provided.

However, whilst the US Mail algorithm may not be any more effective in facilitating judgment change, than the Forested Miles algorithm, consideration was next given to whether the US Mail algorithm facilitated a greater *extent* of judgment change, than the Forested Miles algorithm. This might be so, as the US mail algorithm was more complex (i.e. contained more sub-component steps) than the Forested Miles algorithm. The absolute difference between a participant's unaided estimate, and the same participant's estimate post-advice, was calculated for each participant, and subsequently aggregated for each of the algorithm conditions. The raw scores did not satisfy the assumption of normality, and so were Log transformed, A significance test was performed to determine whether participants who had worked through the US Mail algorithm (and who changed judgment, $N = 13$) ($\bar{x} = 2.65$, $sd = 2.11$) changed judgment to a greater degree, than participant's who had worked through the Forested Miles algorithm (and who changed judgment, $N = 10$) ($\bar{x} = 3.12$ $sd = 4.59$). No significant

difference between conditions, in terms of the extent of judgment change participant's reported, post-advice, was found (t = - 1.0, df = 9.00, p < 0.34). Whilst participants in receipt of the Forested Miles problem, and who changed judgment post-advice, recorded a greater degree of judgment change, than participants in receipt of the US Mail problem, this difference did not reach significance. The greater complexity of the US Mail algorithm, did not result in a significantly greater degree of judgment change, than participants in receipt (of the less complex) Forested Miles algorithm.

Having established that participants in receipt of the US Mail algorithm changed judgment to no greater extent, and no more often, than people in receipt of the Forested Miles algorithm, the influence of advice (the numerical output-of-algorithm) was next considered. A measure of judgment change relative to advice was utilized, expressed as ¦pre-advice estimate – advice¦ - ¦post-advice estimate – advice¦. Should the product of this measure be positively signed, it is possible to infer that the absolute difference between a participant's post-advice estimate, and advice, is *less* than the absolute difference between a participant's pre-advice estimate, and advice. Such a finding indicates that where a participant revised their estimate post-advice, they did so by 'shifting' their judgment in the direction of the advice (see Table 7.4).

Table 7.4 Judgment change relative to advice (output-of-algorithm) (1,000,000s)

Estimation problem	Judgment change relative to advice $\bar{x}$	*sd*
No. items of US Mail in 1987 (N = 13)	235.92	446.11
No. of Forested Miles in Oregon (N = 10)	2,100,001	6640781.33

Table 7.4 shows that participants in receipt of algorithmic decompositions, revised their judgment, by 'shifting' towards advice (output-of-algorithm), (i.e. the product of the dependent measure in each condition is positively signed). Further, a significance test on the data in Table 7.5 revealed a significant difference between conditions (U = 32.0, N_1 = 10, N_2 = 13, $p <$ 0.04). Hence, it is possible to infer that participants 'shifted' their post-advice estimates, towards advice (output-of-algorithm), to a greater extent where people were in receipt of the Forested Miles algorithm, than when in receipt of the US Mail algorithm. Recall, that participants in receipt of the Forested Miles problem became *less* accurate post-advice, on average. In itself, this may not be surprising as the ratio between participant's pre-advice initial estimates and advice, for participants in receipt of the Forested Miles problem, was much greater than the ratio between participant's pre-advice initial estimates and advice for participants in receipt of the US Mail algorithm, on average. I now turn to measures of the perceived utility of each of the algorithms provided to participants.

The complexity of the US Mail algorithm, in comparison to the Forested Miles algorithm, is certainly reflected in the finding that people in receipt of the US Mail algorithm, attributed to it, greater utility ratings, than those reported by participants in receipt of the Forested Miles algorithm, irrespective of whether they revised their judgment post-advice, on average ($U = 194.5$, $N_1 = 24$, $N_2 = 25$, $p < 0.03$). This pattern of results also obtained for people who did not revise their judgment, post-advice. These participants also attributed higher utility ratings to the US Mail algorithm, than ratings attributed by participants in receipt of the Forested Miles algorithm ($U = 45.0$, $N_1 = 11$, $N_2 = 15$, $p < 0.05$). Participants who *did* revise their judgment post-advice, valued both algorithms to the same degree ($U = 52.5$, $N_1 = 13$, $N_2 = 10$, $p < 0.45$). These findings support the idea that participant's perceived the US Mail algorithm to be useful (i.e. it superficially appears systematic); but ultimately participants who revised their judgment post-advice, did not perceive this algorithm to hold any greater utility, than the ratings participants attributed to the Forested Miles algorithmic decomposition. In sum, neither algorithmic decomposition was any more effective than the other - the differences between conditions in terms of post-advice accuracy can be accounted for by 'estimate inflation'. The findings from this study show that the benefits of algorithmic decomposition, in terms of estimating uncertain quantities are constrained, and problematic. Those participant's who revised their pre-advice estimates, did so by reducing the difference between their pre-advice estimate and the available advice, on average, post-advice.

One remaining issue concerns the idea that the extent of judgment revision may depend upon the discrepancy between a participant's pre-advice unaided estimate, and the available advice, on average. This idea was supported when the absolute numerical difference between participants' pre-advice estimates and the advice (output-of-algorithm) was considered. The numerical value of the absolute difference between participants' pre-advice estimates and

advice was calculated, and aggregated for each condition. These raw scores were subsequently Log transformed to satisfy the assumption of normality. It was then possible to determine, if the difference between participants' pre-advice estimate, and advice, differed between conditions, for people who changed judgment, post-advice, on average (see Table 7.5). As the result of this test was significant (t = 2.80, df = 21, p < 0.01), the possibility that people in receipt of the algorithmic decompositions changed judgment, on the basis that the absolute numerical difference between a participant's pre-advice estimates, and the available advice, was different between advice conditions is supported. This relationship is not linear however, as it should be recalled that the advice available to participants in receipt of the US Mail algorithm was numerically greater than participants' pre-advice estimates by a factor of 0.01 million, on average. Contrastingly, advice available to participants in receipt of the Forested Miles algorithm was numerically greater than participants' pre-advice estimates by a factor of 1.54 million, on average. This may indicate that people were more disposed to revise their judgment to a greater degree when the available advice was more similar to the same participants' pre-advice intuitive estimate.

Table 7.5 Mean absolute Log transformed numerical differences between pre-advice estimate, and advice (output-of-algorithm) for participants who revised their judgment post-advice.

	N	$\bar{x}$	*sd*
US Mail	13	21.53	2.63
Forested Miles	10	15.51	7.18

This account is further supported when the pre-advice estimates of people who subsequently changed judgment, post-advice, are considered. Recall that MacGregor *et al.* argue that people in receipt of the Forested Miles algorithm, are able to make less erroneous pre-advice estimates, than people in receipt of the US Mail algorithm. These authors speculate that this is because (i) the true numerical answer to the US Mail problem (89,000,000,000) greatly

exceeds the true numerical answer to the Forested Miles problem (47,506) (ii) people are likely to be more familiar with the Forested Miles of Oregon task, than the US Mail problem. Support for these premises is found here. The idea that participants in the Forested Miles condition were able to provide more accurate estimates pre-advice, than people tasked with estimating the number of items handled by the US Mail in 1987, is supported (see Table 7.6). Prior to performing a significance test, the raw scores were Log transformed to meet the assumptions of normality. For people who subsequently changed judgment, there was a significant difference between conditions, in terms of pre-advice accuracy (t = 11.88, df = 9.002, $p < 0.0005$).

Table 7.6 Mean absolute Log transformed numerical differences between pre-advice estimate, and the true answer (pre-advice accuracy), for people who changed judgment post-advice.

	N	$\bar{x}$	*sd*
US Mail	13	25.20	0.34
Forested Miles	10	14.21	0.92

These findings support the view that the benefits of algorithmic decomposition, in terms of estimating uncertain quantities, are largely a product of post-advice 'estimate inflation', for people estimating the size of large quantities, (Lichtenstein *et al.* 1982; Henrion, 1983). The next stage of the analysis involved testing predictive models of ultimate estimation accuracy, and judgment change through multiple, and logistic regression (see Chapter 6, p217-22 for justification of approach and selection of predictor variables). Prior to reporting the results of this analysis checks for violations of assumptions of the model involving multicollinearity, and hetereoscedascity, were carried out and no violations were found.

A multiple linear regression was performed on the data from participants in receipt of the US Mail problem, where ultimate estimation accuracy was the dependent measure, and pre-advice

initial estimation accuracy, pre-advice initial confidence, and the accuracy of the advice (output-of-algorithm) were entered into the model as predictor variables (see Table 7.7). The model was a poor fit ($R^2_{Adj} = 0.12$), and the overall relationship was not significant ($F_{(4, 19)} = 1.79$, $p < 0.17$). Inspection of the standardized beta weights in Table 7.6 lead to the formulation of a regression equation that predicts a 1% change in ultimate estimation

accuracy -

1% change in ultimate estimation accuracy = 0.007 LN post-advice confidence + 0.32 LN advice accuracy + 0.33 LN initial accuracy.

This regression equation can be interpreted as for every 1 unit decrease in LN post-advice confidence, ultimate estimation improves by 0.007% (this is because increasingly 'wide' confidence intervals imply less confidence), holding other variables constant. Similarly, for every one unit decrease in LN advice accuracy, ultimate estimation accuracy increases by 0.32% (this is because values approaching zero indicate greater accuracy), holding other variables constant. For every one unit increase in LN initial accuracy, ultimate estimation accuracy is predicted to increase by 0.33%. However, none of the variables entered into the regression were significant predictors of ultimate estimation accuracy.

Table 7.7 Multiple regression – predictors of ultimate estimation accuracy

Experimental condition	Variable	B	SE	Beta	*t*	*p*
Forested Miles	constant	-3.51	3.66		-0.96	0.35
	LN Final CI	695,933,080,676	2,901,927,706,333	0.06	0.24	0.81
	LN Advice accuracy	4,408,371,024,658	1,655,784,211,602	0.68	2.66	0.01*
	LN E1 accuracy	-2,086,866,409,171	4,022,063,023,334	-0.14	-0.52	0.61
US Mail	constant	81,549,663,785	2,449,225,411		33.30	0.0005
	LN Final CI	345,434,954	23,54,074,879	0.007	0.03	0.97
	LN Advice accuracy	473,275,422	308,281,273	0.32	1.54	0.14
	LN E1 accuracy	458,275,180	290,233,125	0.33	1.60	0.13

Next, the data from participants in receipt of the US Mail problem was examined in order to ascertain the determinants of judgment change. A logistic regression was performed with judgment change as the DV and initial participant confidence, initial estimation accuracy, accuracy of advice, and perceived utility of advice as sub-scales of the LSI-R, as predictor variables (see Table 7.8). A total of 24 cases were analyzed, but the full model was not significantly reliable ($\chi^2 = 4.44$, $df = 4$, $p < 0.35$). However, the model adequately fitted the data (Hosmer and Lemeshow $\chi^2 = 4.53$, $df = 8$, $p < 0.81$). This model accounted for between 16.9% and 22.6% of the variance in judgment change, with 84.6% of the participant's who changed judgment post-advice successfully predicted. 63.6% of predictions for the group of

participants who did not change judgment post-advice were accurate. Overall, 75% of predictions were accurate. None of the proposed predictor variables proved to be significant.

Table 7.8 Logistic regression – determinants of judgment change

Experimental condition	Variable	B	SE	Wald	*df*	*p*	Exp(B)
Forested Miles	constant	23.66	10.63	4.96	1	0.03	18,932,217,891
	LN initial CI	-0.60	0.41	2.17	1	0.14	0.55
	LN Advice accuracy	0.55	0.30	3.48	1	0.06	1.73
	LN E1 accuracy	-1.62	0.85	3.65	1	0.06	0.20
	Advice rating	-1.63	1.02	2.54	1	0.11	0.20
US Mail	constant	0.43	2.06	0.04	1	0.84	1.54
	LN initial CI	0.73	1.51	0.23	1	0.63	2.07
	LN Advice accuracy	0.07	0.19	0.15	1	0.70	1.08
	LN E1 accuracy	0.16	0.18	0.80	1	0.37	1.18
	Advice rating	-0.80	0.48	2.77	1	0.10	0.45

The data from participants in receipt of the Forested Miles problem was examined next. A multiple linear regression was performed on the data from participants in receipt of the Forested Miles problem, where ultimate estimation accuracy was the dependent measure, and pre-advice initial estimation accuracy, pre-advice initial confidence, and the accuracy of the

advice (output-of-algorithm) were entered into the model as predictor variables (see Table 7.7). The model was a fair fit ($R^2_{Adj} = 0.29$), and the overall relationship was significant ($F_{(3, 20)} = 4.18$, $p < 0.02$). Inspection of the standardized beta weights in Table 7.6, lead to the formulation of a regression equation that predicts a 1% change in ultimate estimation accuracy –

1% change in ultimate estimation accuracy = 0.06 LN post-advice confidence + 0.68 LN advice accuracy – 0.14 LN initial accuracy.

This regression equation can be interpreted as for every 1 unit decrease in LN post-advice confidence, ultimate estimation improves by 0.06% (this is because increasingly 'wide' confidence intervals imply less confidence), holding other variables constant. Similarly, for every one unit decrease in LN advice accuracy, ultimate estimation accuracy increases by 0.68% (this is because values approaching zero indicate greater accuracy), holding other variables constant. For every one unit increase in LN initial accuracy, ultimate estimation accuracy is predicted to decrease by 0.14%. However, only LN Advice accuracy was a significant predictor of ultimate estimation accuracy ($t = 2.66$, $df = 20$, $p < 0.01$).

Next, the data from participants in receipt of the Forested Miles problem was examined in order to ascertain the determinants of judgment change. A logistic regression was performed with judgment change as the DV and initial participant confidence, initial estimation accuracy, accuracy of advice, and perceived utility of advice as sub-scales of the LSI-R, as predictor variables (see Table 7.8). A total of 24 cases were analyzed and the full model was significantly reliable ($\chi^2 = 19.05$, $df = 4$, $p < 0.001$). In addition, the model adequately fitted the data (Hosmer and Lemeshow $\chi^2 = 5.85$, $df = 8$, $p < 0.66$). This model accounted for between 54.8% and 74.7% of the variance in judgment change, with 88.9% of the participant's

who changed judgment post-advice successfully predicted. 93.3% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 91.7% of predictions were accurate. None of the proposed predictor variables proved to be significant.

7.4 Discussion

The present study has shown that knowledge of the process of advice generation is not a clearly beneficial influence on judgment change. The claimed effectiveness of algorithmic decomposition, in terms of estimating uncertain quantities, can be explained by means other than the algorithmic approach. It is contended here that the improvements in estimation accuracy reported by MacGregor *et al.* for participants in receipt of their algorithmic decomposition of the US Mail problem, can be accounted for by (i) the tendency of people to underestimate particularly large quantities pre-advice, and (ii) the idea that algorithmic decompositions have the potential to 'inflate' people's estimates post-advice (Lichtenstein *et al.* 1982; Henrion, 1983). Moreover, the argument that people in receipt of the Forested Miles problem were able to formulate less erroneous pre-advice estimates, than people in receipt of the US Mail problem is explainable by an alternative account. The idea that people in receipt of the Forested Miles problem (47, 506 [or 5 digits]), were unable to improve their estimation accuracy, post-advice, because the quantity to be estimated was substantially numerically less, than the quantity to be estimated in the US Mail problem (89,000,000,000 [or 11 digits]) is problematic. This is because participants in the current study overestimated the true value of the number of Forested miles in Oregon, pre-advice. However, the data analyzed here may point to alternative account, where participants in receipt of the Forested Miles problem evaluate the advice available from the output of the algorithm as 'similar' to their own intuitions, and subsequently are more likely to incorporate advice into their post-advice

deliberations irrespective of the numerical accuracy, or perceived utility of the advice – in this instance leading to greater inaccuracy.

The analysis presented here, suggests neither algorithm was any more effective than another, in terms of the number of judgment revisions, or the *amount* of judgment change post-advice. Instead, participants shifted a proportion of the absolute numerical value between their pre-advice estimate, and the advice (output-of-algorithm), post-advice, on average (i.e. people conformed to advice under conditions by revising their pre-advice estimate – a little). Further evidence for conformity to advice was found when the measure of judgment change relative to advice was considered. Here, participant's where they did revise their judgment post-advice tended to shift their estimate in the direction of the advice, post advice. That participants were insensitive to the asymmetries of advice quality was also found where perceived utility of advice was measured. Here, where participants revised their estimate post-advice, they valued the US mail algorithm, and the Forested miles algorithm equally, on average. However, a person's propensity to revise their judgment is influenced by how similar the numerical component of advice is to the same individual's initial intuition.

The cognitive mechanisms that determine ultimate estimation accuracy, and also judgment change were investigated through the use of multiple linear, and logistic, regression statistical procedures. No significant models were found for the determinants of ultimate estimation accuracy from the data of participants in receipt of the US Mail problem. Neither did the logistic regression procedure successfully model the determinants of judgment change for the same participants. However, participants in receipt of the Forested Miles problem reported a significant model that determined ultimate estimation accuracy. Here, participants were

influenced by what they perceived as the accuracy of advice (output-of-algorithm), even though this did ultimately lead to greater inaccuracy, post-advice.

Currently, the results presented here are limited to the two algorithms reported by MacGregor *et al.* and it is as yet unclear whether these findings generalize more widely to the algorithmic approach. A further limitation to the results reported here, are the relatively low numbers of participants in each condition who changed judgment post-advice. Recall that the majority of participants in this study *did not* change judgment post-advice, and the analysis presented here has focused upon participants who changed judgment post-advice. Further, where people did revise their judgment post-advice, they did so in a very conservative fashion. Studies that have measured judgment change by a participants who have received advice from 'advisors' other than themselves, report judgment change that is in the region of 20-30% of the absolute numerical difference between an unaided estimate and advice (Yaniv and Kleinberger, 2000; Yaniv 2004a; Yaniv 2004b; Harvey and Fischer, 1997). Hence, it remains unclear whether the limited evidence of judgment change reported in this study, is due to the difference between individual judgment, and social interaction with advisor(s), or due to the low number of participants taking part in this study.

In conclusion, what can be stated with some degree of confidence is that the claims of theorists, who advocate the algorithmic approach to estimating uncertain quantities, should be closely scrutinized. Where conformity to advice results in post-advice estimate inflation or deflation; an illusion of beneficial judgment change may disguise the weakness of the particular algorithmic decomposition being used. The preceding account has concentrated upon estimation problems where the quantities to be estimated are typically 'large' (i.e. seven

digits or more). Of further interest then, is whether the conformity to advice account extends to explain observed behaviour where the quantities to be estimated are constituted as relative frequencies (e.g. 1 in 5000, or 1 in 100,000).

Chapter 8: Knowledge of advice generation when estimating relative frequencies.

8.1 Introduction

Given that both the results of my own algorithmic decompositions of estimation problems reported in Chapter 6, and the partial replication of MacGregor *et al.* (1988), reported in Chapter 7, showed evidence of conformity to advice, and simultaneously failed to show clear evidence of the benefits of algorithmic decomposition for quantities that were highly uncertain (i.e. seven digits or more), I next considered if knowledge of the process of advice generation could possibly benefit estimators of small quantities. In their discussion of what constitutes 'extremity' (usually calculated as the number of digits that the true answer of some quantity to be estimated contains), Armstrong and MacGregor (1994) also state that, "extremity could also be defined in terms of small numbers. An example would be, 'What is the chance that a person in the US will die next year because of botulism? (The answer is 1/100,000,000)', (p496). The symmetry of the argument is clear – if algorithmic decomposition benefits enhanced estimation accuracy for extremely large quantities, it should also benefit extremely small quantities. Likewise, knowledge of the process of advice generation should enable participants to evaluate advice against participant's unaided intuition. Of interest here then, is the veracity of the claims of advocates of algorithmic decomposition, and the veracity of an alternative account of conformity to advice. This assertion was tested experimentally.

Algorithmic decomposition with small quantities

In their definition of extreme, but small, numerical quantities, MacGregor and Armstrong (1994), indicate that algorithmic decompositions of estimation (problems) of small quantities are often expressed in terms of *relative frequencies* (i.e. the 'chance' of some event occurring can be described as the ratio of a single instance of the target quantity, to the number of

possible instances of the target quantity). Hence, the 'botulism fatality' example cited by MacGregor and Armstrong, (1994) is expressed as 1/100,000,000. Of interest, is whether knowledge of the process by which advice is generated for target quantities expressed as relative frequencies, could facilitate judgment change and potentially deliver more accurate estimates, than unaided holistic estimation. Equally important is whether the conformity to advice account is sufficient to explain the observed behaviour of participants. Consistent with Armstrong and MacGregor's definition of extreme but small numerical target quantities, I developed two algorithmic decompositions of estimation problems concerning unlikely life events. Participants were tasked with estimating either the relative frequency of the chance that a person would be killed in a road traffic accident in the UK in 2005, or the relative frequency of the chance that a single maternity in the UK in 2005, would result in triplet births (no participants were asked to complete both estimation problems) (see Appendix V).

8.2 Method

Participants

Participants (N = 134) were all students attending the University of Durham (UK). Of our sample, 133 students provided demographic data. Hence our sample consisted of 51 males (age $\bar{x}$ = 24.37 years, sd = 2.69 years), and 82 females (age $\bar{x}$ = 24.07 years, sd = 2.30 years). Participants were recruited by the author (with the permission of staff colleagues) in a lecture theatre at the University of Durham, where students had gathered to attend their studies. Participation in the experiment was voluntary, and participants had the right to withdraw at any time – fortunately none chose to do so. All participants took approximately the same amount

of time to complete the questionnaire. No payments were made for participation in the experiment.

Stimuli and Materials

Participants were tasked with completing questionnaire booklets of either the 'road fatalities', or the 'triplet births' estimation problem (see Appendix V). These booklets were set out so that participants completed the estimation task unaided (page 1 of the questionnaire booklet), prior to receiving advice (page 2 of the questionnaire booklet), then participants had an opportunity to re-visit the estimation problem in the light of advice (page 3 of the questionnaire booklet). Some participants received questionnaire booklets that were identical to the road fatalities problem, or the triplet birth problem, but they did not contain advice. Instead, these participants were tasked with completing a distractor task - the 18-item Need for Cognition (NFC) Scale (Petty, Kao and Caccioppo, 1982). People in this condition provided a baseline 'think again' measure.

Participants in receipt of the road fatalities problem were provided with background information about road safety and the difficulties surrounding the compilation of accurate statistics (see below).

> Road safety has become an increasingly important issue both for the UK government, and the general public in recent years - given the increases in the numbers of motorists, year on year. Regrettably, fatal accidents occur throughout the year on Great Britain's road network involving both motorists and pedestrians. In the absence of accurate statistics, or where the cost of obtaining such information is prohibitively high, how would road safety managers assess the chance of a pedestrian being killed in a motor vehicle accident during the 12 months of 2006? This brief questionnaire is designed to elicit how *you* would go about such a problem.

Hence, participants were encouraged to believe that the experiment was concerned with road safety issues, but in fact the true purpose of the experiment was to determine whether knowledge of the process of advice generation (where the quantity to be estimated was a small quantity expressed as a relative frequency) is a sufficient cue for participants to engage in beneficial judgment revision. Knowledge of the process of advice generation is here operationalized as an eleven step algorithm, purportedly provided to participants by an anonymous advisor (named either 'Joe' or 'Kate') that could offer a solution to the estimation problem. In using the algorithm, participants had to estimate numerical inputs for the sub-components of the algorithm. If participants utilized the advice, the algorithm ultimately resolved in a single numerical value expressing the chance of a pedestrian becoming a fatality in a motor vehicle accident in the UK in 2006 (see Appendix V). Participants were then invited to evaluate the advisor's advice on a 5-point Likert-type scale anchored at one end by '1' (no use at all), and at the other by '5' (absolutely invaluable). Participants were then tasked with deciding whether to revise their unaided estimate in the light of the advisor's advice. Should a participant decide to revise his/her pre-advice estimate, a place to do so was clearly indicated on the questionnaire.

Similarly, participants in receipt of the triplet birth problem were provided with background information about the difficulties health managers face when trying to assess the chance that a maternity will result in a triplet birth (see below).

> Clinicians generally regard triplet maternities as more problematic and potentially dangerous for mothers, and babies, than single births. These maternities often require different medical resources than single births, so a salient question for health managers is how likely it is that a pregnancy will result in a triplet birth In the absence of accurate statistics, or where the cost of obtaining such information is prohibitively high, how are managers able to determine the chance of a triplet birth? This brief questionnaire is designed to elicit how *you* would go about such a problem.

Knowledge of the process of advice generation is here operationalized as an eight step algorithm, purportedly provided to participants by an anonymous advisor ('Joe' or 'Kate') that could offer a solution to the triplet birth estimation problem. Participants had to estimate numerical inputs for the sub-components of the algorithm. Should the algorithm be utilized by participants, it ultimately resolved in a single numerical value expressing the chance that a single maternity would result in a triplet birth in 2005 in the UK. Participants were then invited to evaluate the advisor's advice on a 5-point Likert-type scale anchored at one end by '1' (no use at all), and at the other by '5' (absolutely invaluable). Participants were then given an opportunity to revise their unaided estimate in the light of the advisor's advice. Should a participant decide to revise his/her pre-advice estimate, a place to do so was clearly marked on the questionnaire.

Procedure

Participants were randomly assigned to either the road fatalities estimation problem, or the triplet births estimation problem (see Appendix V), and asked to complete a questionnaire booklet. Some of my participants were provided with algorithmic decompositions of the road fatalities problem, and others were provided with the triplet birth problem in their questionnaire booklet, a third group of participants were tasked with completing the estimation task unaided, and then completing the short form of the Need-for-Cognition Scale (Cacioppo, Petty and Kao, 1982), that was contained in their questionnaire booklet (no participant completed more than one questionnaire booklet). The distractor task was utilized to ensure that all participants took approximately the same amount of time to complete the questionnaire. In order to convince participants that completing the distractor task was a meaningful exercise, an introductory line of text preceded the NFC scale – 'We now move on to ask you about your general reasoning style'. On completion of the distractor task these participants were asked to

re-consider the estimation problem since they had 'had time to think about the issue' (see Appendix V).

Where participants were in receipt of either the eleven step algorithmic decomposition of the road fatalities problem, or the eight step triplet birth problem, it was necessary for them to submit sub-component estimates to the algorithm, and complete simple mathematically operations in order to combine sub-component estimates into a single value that constituted the advice from 'Joe' or 'Kate'. Once a single output-of-algorithm value was available, participants were invited to evaluate the utility of the advice on 5-point Likert-type scale anchored at one end by '1' (no use at all), and at the other by '5' (absolutely invaluable). Finally, all participants were invited to reconsider their estimate, and could choose to provide a different estimate to the one they had given pre-advice. Lastly, participants were thanked for their participation.

8.3 Results

Prior to analysing the data it was necessary to determine that participants provided sufficient responses for the purpose of data analysis. Table 8.1 shows this to be the case.

Table 8.1 Number of responses

Participants N = 134	Unaided estimate	Post-advice estimate	Perceived utility of advice rating
Responses N	132	133	93

Initially, participants pre-advice, advice (output-of-algorithm), and post-advice estimates were aggregated and tabulated. Inspection of participant's pre-advice estimates, and post-advice estimates revealed that the data was not normally distributed. Natural Log transformation of pre-advice initial estimates, and post-advice estimates, allowed the assumptions of normality to be met and facilitated the use of parametric statistical techniques. Next, consideration was given to judgment change relative to advice – should the product of this measure be positive, it is possible to infer that participants, when they did revise their judgment post-advice, did so by reducing the absolute difference between their pre-advice intuitive estimate and the advice, post-advice. This analysis was followed by an examination of the perceived utility ratings that participants attributed to either the road fatalities problem, or the triplet maternities problem. It might be expected (if people indiscriminately conform to any available advice), that no significant differences in the perceived utility ratings of advice would be reported by participants. The aforementioned analyses were supported by consideration of both the amount of judgment change (number of judgment revisions), and the extent of judgment change (absolute difference between pre-advice initial estimate and post-advice estimate). Ultimately, the analysis concludes with testing potential explanatory models through regression. The regression modelling approach is appropriate as it allows (i) the determination of the percentage variance in a criterion variable, based upon the simultaneous consideration of multiple predictor variables, and (ii) where a criterion variable such as judgment change (yes/no) is dichotomous, logistic regression is an appropriate procedure for the simultaneous consideration of multiple predictor variables. First, the data of participants' unaided pre-advice estimates, advice (output-of-algorithm), and post-advice estimates were aggregated and tabulated (see Table 8.2 below).

Table 8.2 Participants in receipt of advice –pre-advice estimate, advice, post-advice estimate (1,000,000s)

Estimation problem	Pre-advice estimate $\bar{x}$	*sd*	Advice $\bar{x}$	*sd*	Post-advice estimate $\bar{x}$	*sd*	True answer
Road fatalities (*N* = 53)	0.040	0.069	0.079	0.206	0.042	0.076	0.00001
Triplet births (*N* = 40)	0.028	0.061	0.048	0.106	0.027	0.06	0.002

Table 8.2 reveals that participants greatly overestimated the relative frequency of the chance that a pedestrian would become a road fatality, and also overestimated the chance that a maternity would result in triplets. Moreover, the output-of-algorithm (advice), for both of the decompositions presented to participants, resulted in values numerically greater than participants' pre-advice estimates, on average. Participants in receipt of the road fatalities algorithmic decomposition do not appear to have revised their pre-advice estimates to any great extent post-advice, on average; similarly participants in receipt of the triplet maternities algorithm also do not appear to have revised their pre-advice estimates to any great extent post-advice, on average.

A significance test was performed on the transformed values reported in Table 8.2 to determine if participants' post-advice estimates differed from their pre-advice estimates. For participants in receipt of the road fatalities problem, there was no significant difference between pre-advice estimates, and post-advice estimates, on average (t = 1.92, df = 51, p < 0.06, two-tailed). Similarly, there were no significant differences between pre-advice estimates, and post-advice estimates for people in receipt of the triplet maternities (t = 0.83, df = 38, p < 0.41, two-tailed). In sum, where participants are in receipt of an algorithmic decomposition of some extremely small quantity, expressed as a relative frequency, the algorithm does not significantly either 'inflate', or 'deflate', post-advice estimates.

Given that people do not appear to be significantly revising judgment post-advice, a measure of judgment change relative to advice (expressed as ¦pre-advice estimate – advice¦ - ¦post-advice estimate – advice¦), was formulated to determine if, where people did revise their judgment, they did so by reducing the absolute numerical difference between their pre-advice estimate, and the advice, post-advice (see Table 8.3). Should the product of this measure be

positively signed, it is possible to infer that participants 'shifted' their judgment towards the advice (output-of-algorithm) post-advice, on average.

Table 8.3 Mean judgment change relative to advice

Estimation problem	Judgment change relative to advice $\bar{x}$	*sd*
Road fatalities (N = 10)	- 0.00267	0.04465
Triplet maternities problem (N = 8)	0.00591	0.06404

Table 8.3 shows that the product of the measure judgment change relative to advice is positively signed for participants in receipt of the triplet maternities problem, but negatively signed for participants in receipt of the road fatalities problem. This implies that people in receipt of the triplet maternities algorithm reduced the absolute numerical difference between their pre-advice initial estimate, and the advice (output-of-algorithm), post-advice. However, participants in receipt of the road fatalities algorithm *increased* the absolute numerical difference between their pre-advice initial estimate, and the advice (output-of-algorithm), post-advice. To determine whether this pattern of results was the result of chance, or a statistically significant effect of the quality of the two algorithms, a significance test was performed on the data in Table 8.3 (t = - 0.34, df = 16, p < 0.74). This finding indicates that although people appear to be following the advice (output-of-algorithm) when in receipt of the triplet maternities problem, and rejecting advice when in receipt of the road fatalities problem the difference between the two conditions is not statistically significant in terms of the measure of judgment change relative to advice. Further, a one-sample t test was performed on the mean judgment change relative to advice data from participants in receipt of the Road fatalities

problem. Here, it was found that the observed value was not significantly different from zero (t = 0.19, df = 9, p < 0.85). Similarly, a one-sample t test was performed on the mean judgment change relative to advice data from participants in receipt of the Triplet maternities problem. Here, it was found that the observed value was not significantly different from zero (t = 0.26, df = 7, p < 0.80).

Given that there was no real difference in the reduction (or increase) of the absolute numerical difference between participants' pre-advice estimate and advice, post-advice, on average; and that this proportionate shift is not dependent on the experimental condition that they were randomly assigned to, it is possible to infer that these participants are indiscriminate in their judgment revisions. Further support for this assertion would be forthcoming, if people did not distinguish between the two algorithmic decompositions, in terms of the value, or utility, that people placed upon them post-advice. The utility ratings of participants who changed their judgment post advice, were cross-tabulated, and no significant relationship was found to exist between experimental condition and utility ratings (χ^2 = 7.34, df = 3, p < 0.06). Hence, it is possible to infer that there was no difference in the utility ratings that participants (who changed judgment post-advice) attributed to the algorithmic decompositions of the estimation problems.

Consideration was next given to the number of judgment changes for each experimental condition (see Table 8.4).

Table 8.4 Number of judgment changes

	Change judgment	Do not change judgment
Road fatalities problem	10 (19.2%)	42 (80.8%)
Road fatalities control	3 (21.4%)	11 (78.6%)
Triplet maternities problem	7 (17.9%)	32 (82.1%)
Triplet maternities control	5 (18.5%)	22 (81.5%)

Table 8.4 shows that the majority of participants did not change judgment, irrespective of whether they received the algorithmic decomposition of the estimation problem. Further, there were only small percentage differences between the numbers of participants in each condition who did change judgment. A significance test was performed to determine if the number of judgment revisions in each treatment condition differed - this proved not to be the case ($\chi^2 = 0.02$, $df = 1$, $p < 0.88$).

Given that participants did not differ in the number of judgment revisions irrespective of treatment condition, the degree, or *amount* of judgment change was next considered. The mean amount of judgment change was calculated as the absolute difference between an individual participant's pre-advice initial estimate, and the same participant's post advice estimate, aggregated by condition. Table 8.5 shows the mean absolute difference between participants' pre-advice estimates, and post-advice estimates, for each condition.

Table 8.5 Amount of judgment change: Mean absolute difference between pre-advice estimates and post-advice estimates

	Absolute difference between pre-advice estimate and post-advice estimate $\bar{x}$	*sd*	Log absolute difference between pre-advice estimate and post-advice estimate $\bar{x}$	*sd*
Road fatalities problem $N = 10$	0.03090	0.03067	- 4.4482	2.1073
Road fatalities control $N = 3$	0.13340	0.15267	- 4.0080	3.9436
Triplet maternities problem $N = 8$	0.03094	0.05520	- 8.0866	4.8584
Triplet maternities control $N = 5$	0.00036	0.000564	- 9.7395	2.7365

As *N* is low in each of the cells in Table 8.5, performing a test of significance between conditions is not very meaningful due to low statistical power – particularly as the untransformed scores did not meet the assumptions of normality. However, inspection of the mean values in Table 8.5 suggests that those participants in receipt of algorithmic decompositions of the estimation problems appear to have revised their judgment to the same amount, on average. It is not possible to make meaningful comments about the mean values for people in receipt of the distractor task for the triplet maternities problem, and the road fatalities problem beyond noting that a single observation appears to have inflated the mean value.

Of further interest, were the potential determinants of ultimate estimation accuracy, and judgment change – analyzed here by regressing potential predictor variables upon a criterion variable, similar to the analytical procedures carried out in Chapter 6 (see p217-22). Second, a logistic regression procedure is carried out to determine what factors are influential in judgment change. A multiple linear regression was performed on the data from participants in receipt of the Road fatalities problem (N = 52), where ultimate estimation accuracy was the dependent measure, and natural log transformations of pre-advice initial estimation accuracy, and advice accuracy were predictor variables. The model was a fair fit (R^2_{Adj} = 0.40), and the overall relationship was significant ($F_{(2,\ 49)}$ = 17.89, p < 0.0005). Inspection of the beta weights reported in Table 8.6, leads to the formulation of a regression equation that predicts a 1% change in ultimate estimation accuracy. –

1% change in ultimate estimation accuracy = 0.63 LN E1 initial accuracy + 0.08 LN advice accuracy.

It is possible to interpret this regression equation as every 1 unit decrease in LN E1 initial accuracy, ultimate estimation accuracy improves by 0.63%, holding other variables constant. Likewise, for every 1 unit decrease in LN advice accuracy, ultimate estimation accuracy improves by 0.08% (recall that advice accuracy improves as values approach zero), holding other variables constant. However, only LN E1 Initial accuracy was a significant predictor of ultimate estimation accuracy (t = 5.71, df = 51, p < 0.0005). This is unsurprising, given that there was no significant revision of pre-advice judgment, post-advice.

Table 8.6 Multiple regression – ultimate estimation accuracy

Experimental condition	variable	B	SE	Beta	*t*	*p*
Road fatalities	constant	0.14	0.022		6.16	0.0005
	LN E1 initial accuracy	0.02	0.003	0.63	5.71	0.0005
	LN advice accuracy	0.002	0.002	0.08	0.71	0.49
Triplet maternities	constant	0.20	0.02		12.77	0.0005
	LN E1 initial accuracy	0.03	0.003	0.78	9.40	0.0005
	LN advice accuracy	0.01	0.002	0.21	2.53	0.02

Next, the determinants of ultimate estimation accuracy (natural log transformations of pre-advice initial estimation accuracy, and accuracy of advice) were entered into the regression model for participants in receipt of the triplet maternities problem (N = 39). The model was a good fit (R^2_{Adj} = 0.78), and the overall relationship was significant ($F_{(2, 36)} = 69.09$, $p < 0.0005$).

Inspection of the standardized beta weights for participants in receipt of the triplet maternities problem (see Table 8.6) lead to the formulation of a regression equation that could predict a 1% change in ultimate estimation accuracy –

1% change in ultimate estimation accuracy = 0.78 LN Initial accuracy + 0.21 LN Advice accuracy.

The regression equation can be interpreted such that, a single unit decrease in initial estimate accuracy increases ultimate estimation accuracy by 0.78% (holding LN Advice accuracy constant). This is because high values of initial estimate accuracy are indicative of *less* accuracy. Similarly, each unit decrease in the accuracy of advice increases ultimate estimation accuracy by 0.21% (holding LN Initial accuracy constant). LN Initial accuracy (t = 9.40, df = 38, $p < 0.0005$), and LN Advice accuracy (t = 2.53, df = 38, $p < 0.02$), were both significant predictors of ultimate estimation accuracy. Whilst the finding that initial estimation accuracy accounted for variance in ultimate estimation accuracy is not surprising, the finding that advice accuracy accounts for some of the variance in ultimate estimation accuracy (for people in receipt of the triplet maternities problem) is. This is because participants largely ignored advice and did not significantly revise their initial pre-advice judgments. It is possible that, as the accuracy of advice ($\bar{x}$ = -5.57, sd = 2.13) was no more accurate than participant's pre-advice intuitions ($\bar{x}$ = -5.31, sd = 1.71), (t = 0.76 df = 38, $p < 0.45$), those few participants who did act upon advice, did so no more beneficially than those who did not.

Next, the determinants of judgment change for participants in receipt of the road fatalities problem were sought through logistic regression. A logistic regression was performed with judgment change as the DV and initial estimation accuracy, accuracy of advice, and perceived utility of advice as sub-scales of the LSI-R, as predictor variables. A total of 52 cases were analyzed and the full model was significantly reliable (χ^2 = 10.55, df = 3, $p < 0.01$), and the

model fitted the data adequately (Hosmer and Lemeshow $\chi^2 = 10.00$, $df = 8$, $p < 0.27$). This model accounted for between 18.4% and 29.4% of the variance in judgment change, with 30% of the participant's who changed judgment post-advice successfully predicted. However, 90.5% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 78.8% of predictions were accurate. Only the perceived utility of advice significantly predicted judgment change. Table 8.7 shows that for every one unit increase in the perceived utility of advice, the likelihood of judgment change increases by a factor of 0.39. This finding is intuitive – judgment change is more likely when people perceive advice to be useful.

Table 8.7 Logistic regression – predictors of judgment change

Experimental condition	Variable	B	SE	Wald	*df*	*p*	Exp(B)
Road fatalities	constant	1.68	1.39	1.46	1	0.23	5.35
	LN Advice accuracy	-0.095	0.12	0.66	1	0.42	0.91
	LN E1 accuracy	-0.22	0.15	1.96	1	0.16	0.81
	Advice rating	-0.93	0.39	5.61	1	0.02*	0.39
Triplet maternities	constant	66.76	43.00	2.41	1	0.12	9.89
	LN Advice accuracy	9.00	6.38	1.99	1	0.16	0.76
	LN E1 accuracy	-0.24	0.56	0.19	1	0.67	0.79
	Advice rating	-3.47	1.55	5.03	1	0.03*	0.03

* $p < 0.05$

Consideration was next given to the determinants of judgment change for participants in receipt of the triplet maternities problem. A logistic regression was performed with judgment change as the DV and initial estimation accuracy, accuracy of advice, and perceived utility of advice as sub-scales of the LSI-R, as predictor variables. A total of 39 cases were analyzed and the full model was significantly reliable ($\chi^2 = 26.82$, $df = 3$ $p < 0.0005$), and the model fitted the data adequately (Hosmer and Lemeshow $\chi^2 = 2.53$ $df = 8$ $p < 0.96$). This model accounted for between 49.7% and 81.5% of the variance in judgment change, with 85.7% of the participant's who changed judgment post-advice successfully predicted. However, 96.9% of predictions for the group of participants who did not change judgment post-advice were accurate. Overall, 94.9% of predictions were accurate. Table 8.6 gives coefficients and the Wald statistic and associated degrees of freedom and probability values for each of the predictor variables.

Table 8.7 shows that for every one unit increase in the perceived utility of advice, the likelihood of judgment change increases by a factor of 0.03 for participants in receipt of the triplet births estimation problem. Although this finding might appear intuitive (judgment change is more likely when people perceive advice to be useful), it is clearly less influential than the weight that participants attach to their perceptions of advice utility when considering the output-of-algorithm in the road fatalities problem. It is difficult to ascertain why this might be so given the low *N* of participants who did change judgment in this study, but one explanation may include the differences in task characteristics between estimation problems. Road traffic accidents are often newsworthy both at a local and national level in the UK, so people may weight their perceptions of these events more so than less common (or newsworthy) events such as triplet births. Put another way, all pedestrians (irrespective of gender) may be susceptible to road safety issues at one time or another, but it is only women who personally experience childbirth. Further, it is a small proportion of maternities that result

in triplets. Hence, on average people may weight the seriousness of road traffic accidents more so, than the chance of any single maternity resulting in a triplet birth.

8.4 Discussion

The results reported here do not support the idea that knowledge of the process of advice generation benefits estimates of both extremely large, *and* extremely small, numerical quantities (Armstrong and MacGregor, 1994). Nor do the preceding results provide strong evidence for the idea that people indiscriminately conform to (self-generated) advice. Instead, this study demonstrated that irrespective of either of the two algorithmic decompositions presented, participants who change judgment do so because they *perceive* advice to be useful (even though advice in this study could potentially degrade estimation accuracy post-advice, on average). This finding is particularly disconcerting as far as interventions such as algorithmic decomposition are concerned, as without knowing *a priori* the beneficial effects of this approach, it is unlikely that by using the decomposition method, a participant will be able to make estimation accuracy improvements, beyond those available by guessing. In part, this finding highlights the advantage of regression based analysis in JAS studies, as opposed to measures derived from difference scores (Cronbach and Furby, 1970). These findings also echo those of Lichtenstein and Weathers (1998), who found that, even when people were trained to create their own algorithmic decompositions (of estimates of uncertain quantities), performance was not improved to any great degree. Further, these authors also found that, on some occasions, the use of algorithmic decomposition could inhibit accurate estimation (Study 2). Similarly, Henrion *et al.* (1993) compared the accuracy of estimates of subjective probability distributions produced by (i) people generating their own decompositions and (ii) people working through algorithmic decompositions provided by an experimenter. They found no differences in the accuracy of the outputs of these two conditions. Instead, Henrion *et al.* found that algorithmic decomposition 'inflated' people's estimates, in comparison to a control

condition. Hence, people who underestimated a target quantity initially, and persisted in underestimating (to a lesser degree) a target quantity after working through an algorithmic decomposition can be expected to improve their estimation accuracy. Conversely, people who overestimate a target quantity, initially, can expect to diminish their estimation accuracy after working through an algorithmic decomposition.

The 'inflationary' nature of algorithms used to decompose estimation problems involving extremely large numerical quantities was *not* observed in this study, however. Instead, no significant effect upon participants' post-advice estimates was observed, in comparison to pre-advice estimates. However, when people did revise their judgment post-advice, the shift in judgment was negligible. This finding may point towards the notion that, having invested cognitive resources in 'working through' an algorithmic decomposition that 'aids' estimation of an uncertain quantity, people who revise their judgment do not always do so by accepting the veracity of the output of the algorithm (Bonaccio and Dalal, 2006).

In this study regression modelling was able to show some of the determinants of ultimate estimation accuracy, and judgment change. Analysis of the data from participants in receipt of the road fatalities algorithm, and the triplet maternities algorithm, indicated that pre-advice initial estimation accuracy is influential in terms of ultimate estimation accuracy. This finding is unsurprising as there was no significant difference between participant's pre-advice initial estimates, and post-advice estimates in either experimental condition. The data from both estimation problems suggests that the perceived utility of advice is influential in judgment revision.

This study shows that, in conditions where people make estimates of extremely small numerical quantities (expressed as relative frequencies), they are unable to use an algorithmic decomposition to mechanically combine sub-component estimates into a useful estimate of the quantity in question. This may imply that in such circumstances people are more likely to utilize peripheral environmental cues, rather than invest cognitive resources, in what potentially maybe a fruitless effort, to evaluate the quality of advice on offer (in conditions of less uncertainty it is possible that people may be able to successfully scrutinize the quality of advice). Moreover, where people have little or no task relevant knowledge, and are unable to input more accurate sub-component estimates into an algorithmic decomposition than what may be achievable by a global holistic evaluation of an estimation problem, then algorithmic decomposition may not be of any greater utility than holistic estimation. Further, given the circumstances outlined above, it is likely that people are insufficiently motivated (at least in this study) to commit cognitive resources to a potentially fruitless evaluation of advice. This is because where people have little or no task relevant knowledge it is difficult to evaluate the quality of advice. Also, evaluations of advice based upon 'world knowledge' are insufficient for the purposes of distinguishing between poor and good advice. Lichtenstein (1990) makes a similar point in reporting the results of her second study, where she argues that algorithmic decomposition is of little utility where it is (i) poorly understood, and (ii) 'noisy' due to the poor quality information generated by participants, coupled with a lack of arithmetic skills - which are necessary to combine the various component steps of the algorithm. She goes on to argue that the conditions for the successful use of algorithmic decomposition are likely to involve the careful and deliberate design of algorithms, and various computational aids for participants.

In sum, the findings from this study suggest that people tend to overestimate the chances of the occurrence of rare events (at least in the two estimation tasks completed by participants here).

Further, algorithm decompositions of estimates of small quantities produce values that exceed unaided estimates (much in the same way that algorithms inflated pre-advice estimates in Chapter 7). However, where the results differ from those of the study reported in Chapter 7 pre-advice estimate inflation due to the use of algorithmic decompositions of estimation problems was not carried over to post-advice estimates. Here, only a small number of participants revised their post-advice estimates, and such revision was not sufficient to create a significant difference between the mean values of estimates pre-advice, and the mean values of estimates post-advice. However, people valued the algorithmic decompositions equally, on average, irrespective of condition. Moreover, it was participants perceptions of the utility of the algorithms that was the only significant variable (determined by logistic regression) that influenced their propensity to revise their judgment post-advice. It is also interesting to speculate over the influence of the different characteristics of each task in participants' deliberations. However, such speculations will need to be developed and investigated in future work as there is insufficient data in this study to form even tentative conclusions.

The preceding results add to my understanding of some of the influences upon human judgment revision, in so far as future studies will have to consider knowledge of the process of advice generation as a multi-dimensional construct – rather than as the unitary construct presented here, and in Chapter 7. It has become clear that people may be able to assess the utility of the process of advice generation, but are constrained in their abilities to successfully use this knowledge to formulate beneficial judgment revision. Algorithmic decompositions may inflate pre-advice estimates, leading to either an illusion of enhanced estimation (or enhanced estimation error). An individual's cognitive evaluation of advice is further mediated by the task domain from which estimation problems are drawn, and the specific characteristics of a particular estimation problem. It may be that the mechanical nature of an algorithmic decomposition of some troublesome estimation problem is not a sufficient cue to effortful and

relevant cognition – despite its mathematical tractability. This idea has some credence given the nature of the tasks presented here – people are aware of the nature of traffic accidents and of complications in childbirth - possibly to a greater extent than the number of pieces of mail handled by the US postal service, or the number of forested square miles in the US state of Oregon. Given the limitations that this study has highlighted about the utility of knowledge of the process of advice generation when estimating relative frequencies, and the difficulty of sustaining the symmetry of the algorithmic approach to estimation, future studies could fruitfully explore whether people are able to distinguish between objectively 'good' and 'poor' advice. If people are sensitive to appropriate cues in the way that advisor(s) formulate advice, this may serve as proxy cues to the accuracy and utility of such advice.

Chapter 9: Distinguishing between objectively 'good' and 'poor' advice - perceptions of infant mortality rates.

9.1 Introduction

In a situation where an individual is engaged in making a quantitative estimate to some question or quantity of interest, and precise numerical information is not available and/or retrievable from memory, such an individual is theorized to be making 'plausible inferences' between their own world knowledge, and the target item; further, in order to produce a meaningful estimate, a (numerical) value must, ultimately, be attached to the target item (Collins, Warnock, Aiello and Miller, 1975; Collins, and Michalski, 1989). The dilemma for a decision maker in these circumstances, is whether an unaided intuitive estimate should be revised when additional information, (in the form of quantitative advice), becomes available. In sum, the issue of interest here is whether people are able to adequately discriminate between poor, and good advice, or whether conformity to advice obtains, when forming an estimate under conditions of poor, or at best incomplete domain knowledge.

Why attend to 'poor' advice?

It is unremarkable to assert that people occasionally attend to poor advice - however, what is noteworthy is that they may do so for very good reasons. Consider a scenario where the advice available to a judge maybe poor (i.e. complying with the advice results in sub-optimal estimation accuracy), but does contain information or alternatives that are unfamiliar to a judge - potentially enriching the basis for any subsequent estimate (Heath and Gonzales, 1995). Alternatively, interacting with advice may cause a judge to think of the decision problem in a different, and potentially more beneficial, way (Schotter, 2003; Druckman, 2001). Further,

improvements in estimation accuracy can be obtained by integrating (numerical) advice from independent and/or uncorrelated sources (Budescu and Rantilla, 2000; Yaniv 2004a; Yaniv and Kleinberger 2000). Moreover, people seek out, and pay attention to advice, not only to improve the accuracy of their decisions, but to share responsibility for those decisions; also judges are reticent to completely reject advice that is freely given – irrespective of its quality. Here, a 'token' shift by judges towards the position of the advisor(s) has been observed (Sniezek and Buckley, 1995; Harvey and Fischer, 1997; Yaniv, 2004a; Yaniv 2004b). In sum, people are susceptible to the influence of inappropriate advice under conditions of uncertainty.

Can people differentiate between 'poor' and 'good' advice?

The preceding discussion outlines why people may choose to attend to poor advice under conditions of uncertainty. It is equally true that people may choose to attend to good advice under conditions of uncertainty, but the question of interest addressed in this study is whether people are able to adequately differentiate between poor and good advice, when attempting to make an accurate estimate of some quantity. Moreover, where people hold little confidence in their ability to make accurate estimates of target quantities, they may be disposed to attend to any advice (irrespective of its quality) from a credible source in attempting to improve estimation accuracy (Yaniv and Kleinberger, 2000; Yaniv, 2004b). These findings suggest that people find distinguishing between potentially beneficial, and poor advice, cognitively demanding. People may seek to relax the constraints upon cognition, by recognizing environmental cues that potentially serve as substitutes for direct feedback about the veracity, and accuracy of their unaided estimate(s). Previous research has shown the influence of trust (Van Swol et al. 2001), advisor confidence (Sniezek et al. 1995), and source credibility (Reinard, 1988), as some of the cues that are inferred to relax cognitive demands, and potentially influence the estimation performance of judges. Lee (2007, unpublished) sought to

develop the idea that people utilize 'proxy' cues when attempting to make quantitative estimates under conditions of uncertainty. This author reasoned that people may generate quantitative estimates on the basis of ecologically valid proxy cues, and that people will successfully distinguish between potentially beneficial and unhelpful cues. Such a rationalization was operationalized in terms of a task where participants sought to estimate the Infant Mortality Rate (hereafter IMR) of several countries. IMRs are defined as the number of children who die (irrespective of cause) prior to reaching one year old, per every one thousand live births. Hence, an IMR of 35.8 would indicate that, on average, in every 1000 live births, 35.8 children under the age of twelve months die. Some participants were aided in this task through the provision of numerical information of each of the target countries' per capita (US$) Gross Domestic Product (hereafter GDP). It was theorized that 'national wealth' is an objectively good cue to national IMRs as, on average a nation's GDP is negatively correlated with its IMR. Other participants were similarly aided in making estimates of countries' IMRs, by the provision of accurate population estimates of each of the target countries. Such an operationalization facilitated a comparison in terms of post-advice estimation accuracy between people in receipt of GDP advice, and people in receipt of Population estimate advice. Various measures of estimation performance, or estimation accuracy, have been utilized in JAS research; the following section discusses two such measures – metric and ordinal accuracy.

Measures of estimation accuracy

Typically, in JAS research, improvements in estimation accuracy are derived from the absolute difference between an unaided point estimate and the true value, prior to accessing advice, and a point estimate and the true value post-advice (hereafter 'metric' accuracy). If the latter measurement is less than the former measurement, then it is not unreasonable to conclude that

a participant has improved their estimation accuracy. However, Gigerenzer, Todd and the ABC Research Group (1999), propose a further measure of estimation accuracy for participants attempting to make estimates of (numerical) quantities, under conditions of imperfect domain knowledge. They argue that people are able to recognize predictive information in an estimation task, and recall cues that enable individuals to rank order target items (hereafter 'ordinal' accuracy). This is a useful distinction as, on average, people find estimating numbers and statistics such as IMRs difficult, as the exact values of such statistics are rarely widely known. Hence, the ability of people to recognize predictive information, and recollect cues sufficient to rank order target items – as Gigerenzer *et al.* argue - is clearly ecologically useful.

The significance of recognizing the differences between metric and ordinal accuracy is that these two measures may well be independent – people might simultaneously make poor metric estimates of a country's IMR, while accurately rank ordering several countries by IMR. Lee (2007 unpublished) tasked his participants with rank ordering 58 countries by IMR, whilst simultaneously providing highly predictive information (each country's GDP per capita US$) as an aid (Study 4). Here, people improved their ordinal estimation accuracy. In a second condition, Lee's participants were provided with information that is not predictive of a country's IMR (i.e. the population figures for each of the same 58 representatively sampled countries). Here, there were negligible ordinal estimation accuracy improvements. Participants received both GDP information, and Population estimate information in a third condition, here, ordinal accuracy did not surpass that achieved by participants who only received GDP information in condition 1. However, the critical issue for the purposes of this study, is that estimation accuracy improvements for participants in receipt of 'predictive' advice, in conjunction with negligible estimation accuracy improvements for participants in receipt of 'non-predictive' advice is entirely consistent with the idea that participants are

constrained in their abilities to evaluate 'good' or 'poor' advice, and revise post-advice estimates by conforming (to a degree) with any available advice.

Conformity to advice

The findings from previous chapters have strongly suggested that people (i) prefer reasons-based evidence in addition to numeric advice (ii) are disposed to attend to advice from a credible source (iii) egocentrically discount solely numeric advice, and reason-based advice to the same degree (iv) under conditions of uncertainty are disposed to conform to advice. The latter finding, is largely due to the relative positions of a participant's unaided estimate, the true (but unknown to the participant) value that an estimate is sought for, and the available advice. Where a participant revises an unaided estimated, by reducing the absolute numerical distance between the estimate and the available advice, that simultaneously reduces the absolute numerical distance between the estimate and the true answer to an estimation problem, improvements in estimation accuracy can be expected.

Given the preceding exposition of conformity effects, the idea that people report estimation accuracy improvements when in receipt of 'predictive' advice, and no such estimation accuracy improvements when in receipt of 'non-predictive' advice (Lee, 2007, unpublished) is not inconsistent with a conformity effect. Recall that Lee tasked his participants with estimating the IMRs of 58 countries, and provided GDP advice, Population estimate advice, or GDP and Population estimate advice, as an aid to participants (a participant was exposed to no more than one of the advice types). This implies that should people conform to advice in such a task, estimation accuracy improvements can be expected for participants in receipt of GDP

advice, and GDP and Population estimate advice – but not for participants in receipt of solely Population estimate advice.

This reasoning leads to the hypothesis –

> H1: Average ordinal and metric accuracy will improve from pre to post-advice for participants in receipt of GDP information alone, or in combination with Population estimate information, but participants in receipt of Population estimate information only, will not report improvement in estimation accuracy.

Consideration is next given to the issue of advice quality. The notion of advice quality is difficult for JAS theorists for a number of reasons. Firstly, few studies in the extant JAS literature have operationalized advice as anything other than numeric or statistical. Second, at an individual level, advice quality is difficult to isolate from an individual's specific domain knowledge, and thirdly cognitive limitations imply that people may imperfectly both assess, and use, advice in estimation (Harvey et al. 2000). However, the estimation of quantities such as IMRs may offer a way through some of these difficulties. Recall that Lee's (2007 unpublished) study not only involved the metric and ordinal estimation of various countries IMR figures, but in three conditions offered GDP information, Population estimate information, or GDP and Population estimate information as advice. Here, both GDP and Population estimates were numerically accurate, so at one level this advice was equivalent in terms of its quality. However, GDP information is highly predictive of IMRs, whereas Population estimate advice is not. Participants were not made aware of the predictiveness of the advice in Lee's study, but were made aware of the accuracy of the advice. If participants had been asked to rate the utility of advice (where advice was made available to participants), it might be expected that there would be no differences between participants in receipt of GDP advice, and participants in receipt of Population estimate advice, as long as people were focused on advice accuracy as an aid in resolving the estimation problem. If, however, people were able to recognize the predictive superiority of GDP advice over Population estimate

advice, in resolving the estimation problem, then Population estimate advice would have been rated less highly than GDP advice. That Lee did not ask his participants to rate the utility of the GDP, Population estimate, and GDP and Population estimate information in aiding the formation of an estimate of a specific country's IMR is unfortunate, as the 'perceived utility' of such advice is likely to be a factor in participants' estimation metric and ordinal accuracy. The preceding discussion leads to the following hypothesis –

> H2: If participants can distinguish between 'good' and 'poor' advice, then participants will value GDP information more highly, than participants in receipt of Population estimate advice. Participants in receipt of both GDP and Population estimate information, will value advice more highly than participants in receipt of solely Population estimate information, on average.

9.2 Method

Participants

Participants (N =118) for the study were drawn from staff and students (age $\bar{x}$ = 24.5 years, *sd* = 8.37 years) attending the University of Durham. The sample consisted of 79 males (age $\bar{x}$ = 24.78 years, *sd* = 9.22 years), and 38 females (age $\bar{x}$ = 23.89 years, *sd* = 6.32 years) – one participant did not provide age data. Members of staff at Durham Business School were recruited by the author via a personal approach; however the majority of participants were students at the University of Durham. Students were recruited by the author (with the permission of staff colleagues) in a lecture theatre at the University of Durham, where students had gathered to attend their studies. Participation in the experiment was voluntary, and participants had the right to withdraw at any time – fortunately none chose to do so. All participants took approximately the same amount of time to complete the questionnaire. No payments were made for participation in the experiment.

Materials and Stimuli

Participants were tasked with completing questionnaire booklets, where the experimental task was to estimate the IMRs of ten countries. Four versions of the questionnaire booklet were constructed so that three booklets contained advice (GDP advice, Population advice, GDP and Population advice – no booklet contained more than one advice type), and one booklet did not contain advice. The booklet that did not contain advice instead contained a distractor task.

The front covers of all four questionnaire booklets were identical, and informed participants that the study 'examines the perceptions people may hold about infant mortality rates (IMR) in various countries throughout the world' (see Appendix VI). The inside cover of the questionnaire booklet was identical for all four questionnaire booklets - a paragraph of text explained the nature of IMRs and how they are calculated (see below).

> IMR is the ratio of infant mortality per 1000 live births (e.g. an IMR of 5.5 is representative of 5.5 infant deaths per 1000 live births on average; an IMR of 500 would indicate that 1 in 2 children die within the first 12 months of life; whilst an IMR of 995 per 1000 live births would indicate that nearly all children die within the first 12 months of life). Whilst statistics such as these are compiled by governments and various international health organisations, most non-expert people have only a vague idea of the true rate of infant mortality in their own, or in other peoples' countries.

In order to prevent participants 'anchoring' upon the numerical information in the paragraph above, three other versions were created that randomised the presentation order of the numerical information (see below).

> IMR is the ratio of infant mortality per 1000 live births (e.g. an IMR of 995 per 1000 live births would indicate that nearly all children die within the first 12 months of life; an IMR of 500 would be indicate that 1 in 2 children die within the first 12 months of life, whilst an IMR of 5.5 is representative of 5.5 infant deaths per 1000 live births, on average). Whilst statistics such as these are compiled by governments and various international health organisations, most non-expert people have only a vague idea of the true rate of infant mortality in their own, or in other peoples' countries.

> IMR is the ratio of infant mortality per 1000 live births (e.g. an IMR of 5.5 is representative of 5.5 infant deaths per 1000 live births on average; an IMR of 500 would indicate that 1 in 2 children die within the first 12 months of life; an IMR of 995 per 1000 live births would indicate that nearly all children die within the first 12 months of life). Whilst statistics such as these are compiled by governments and various international health organisations, most non-expert people have only a vague idea of the true rate of infant mortality in their own, or in other peoples' countries.
>
> IMR is the ratio of infant mortality per 1000 live births (e.g. an IMR of 995 per 1000 live births would indicate that nearly all children die within the first 12 months of life; an IMR of 500 would indicate that 1 in 2 children die within the first 12 months of life; an IMR of 5.5 is representative of 5.5 infant deaths per 1000 live births, on average). Whilst statistics such as these are compiled by governments and various international health organisations, most non-expert people have only a vague idea of the true rate of infant mortality in their own, or in other peoples' countries.

The preceding randomization of numerical information in the explanation and examples of IMRs was designed to ensure that any 'anchoring' effects that participants engaged in would be mitigated by the random allocation of participants between experimental conditions. Participants were asked to provide details of their age and gender on the same page.

The third page of the questionnaire booklet tasked participants with rank ordering ten countries in order of highest IMR, then providing a point estimate of IMR for each country.

A fourth page in the questionnaire booklet provided participants with advice. In one version of the questionnaire booklet GDP advice was provided. A paragraph of text informed participants that IMRs are often not widely understood, and that accurate GDP information for each of the ten countries they had previously estimated IMRs for, would now be provided. A similar paragraph of text was provided for questionnaires that contained Population estimate advice, and GDP and Population estimate advice (see Appendix VI). Participants were

instructed to copy the information they had entered on page 3 into the appropriate text boxes on page 4. Subsequently, participants were able to choose (or not) to revise their original estimates in the light of advice. Finally, participants were asked to evaluate the advisor's advice on a 5-point Likert-type scale anchored at one end by '1' (no use at all), and at the other by '5' (absolutely invaluable). Participants were then thanked for their participation.

Where participants did not receive advice, they received instead (commencing on page 4 of the questionnaire), a distractor task - the 18-item NFC Scale (Petty, Kao and Caccioppo, 1982) (see Appendix VI). Participants were encouraged to believe that the distractor task was meaningful (and approximately as meaningful, on average, as the tasks participants completed in the advice conditions) by an introductory sentence at the top of page four of the questionnaire – 'This next section is concerned with your individual thinking style. Please complete all questions'. On completion of the distractor task, participants were directed to re-visit their unaided estimates of IMRs by the sentence, 'since you have now had a short period of time upon which to reflect about the infant mortality rates (IMR) of 10 different countries, please re-evaluate your previous estimate of each IMR for the 10 countries below'. The purpose of this direction was to establish a baseline 'think again' condition.

Procedure

Participants were tasked with estimating the IMRs of ten countries. The selected countries were drawn from the CIA Factbook (2007), on the basis of their GDP – five countries with the highest GDP, and the five countries with the lowest GDP. Presentation order of these countries was randomized. The introductory remarks in the questionnaire booklet informed participants of the purpose of the study (to estimate IMRs), and explained how IMRs were derived, and also informed participants of the actual range of IMRs (2-185) across the ten

selected countries. Participants were then asked to provide both their age, and their gender. Participants were then tasked with estimating the IMR of each of the ten selected countries. The estimation process had two phases. In Phase I participants were tasked with ranking the ten countries in order of IMR (a rank of '1' = country with lowest IMR figure, and a country given the rank of '10' = the country with the highest IMR figure), and then estimating the actual rate of IMR in each of the ten countries.

In Phase II participants were randomly allocated between the experimental conditions of control group (no advice), and conditions in which participants were exposed to GDP advice (accurate numerical estimates of each countries' GDP); Population estimate advice (accurate numerical estimates of each country's population); or GDP and Population estimate advice (where participants were exposed to accurate GDP information, and accurate Population estimate information for each of the ten countries). An individual participant was exposed to only one of the three types of advice information. Participants were asked to copy the estimates they had provided in Phase I of the estimation process, into the spaces provided in the questionnaire booklet prior to commencing Phase II of the experiment. In Phase II of the experiment, participants were tasked with revisiting their original estimates in the light of the advice information (or reconsidering their estimates after completing a distracter task in the case of the control group), and were able to revise their unaided estimates if they so chose. Finally, participants were asked to rate how useful they found the available advice, in forming estimates of national IMRs on a 5 point Likert type scale (anchored at one end by 0 = no use at all; anchored at 5 = absolutely invaluable). Participants were thanked for their participation, but were not paid for their participation.

9.3 Results

Prior to analysis the number of responses from participants was tabulated (see Table 9.1).

Table 9.1 Number of responses

Participants *N* = 118	Pre-advice estimate	Pre-advice Rank order	Post-advice estimate	Post-advice Rank order
Responses *N*	188	188	139	167

The primary hypothesis addressed by this study, involved determining the ordinal and metric post-advice accuracy of participants in receipt of advice. Specifically, should people be constrained in their abilities to discriminate between potentially beneficial advice, and poor advice, and subsequently indiscriminately conform to any available advice, it was predicted that participants in receipt of GDP advice would improve both their ordinal and metric accuracy, more so, than participants in receipt of Population estimate advice. Participants in receipt of GDP and Population estimate advice are also expected to improve their ordinal and metric accuracy, more so, than participants in receipt of Population estimate advice, but not to exceed the post-advice accuracy achieved by participants in receipt of GDP advice. For the purposes of analysis, ordinal accuracy is defined as the strength of association between participants' rank orderings of ten target countries (in terms of IMR), and the true rank order of the same 10 countries. If participants improved in terms of ordinal accuracy, then we might expect that there would be a stronger association between participants' rank orderings post-advice and the true rank order (of the 10 target countries by IMR), than pre-advice. For ordinal accuracy to improve it is necessary for some participants to revise their rank ordering of countries in terms of national IMRs. Of interest then, is the number of judgment revisions made by participants.

In order to examine the number of judgment revisions participants made in terms of ordinal accuracy, the number of judgment revisions was tallied for each participant, and for each estimation question. Subsequently, these data were aggregated for the purpose of comparison (see Table 9.2).

Table 9.2 Ordinal accuracy revisions

Condition	Number of ordinal revisions
No advice Controls (*N* = 44)	127/439 (28.92%)
GDP advice (*N* = 36)	224/360 (62.22%)
Population estimate advice (*N* = 16)	67/160 (41.88%)
GDP & Population estimate advice (*N* = 22)	132/220 (60%)
All (N = 118)	550/1179 (46.65%)

Table 9.2 shows that, on average, approximately half of my participants (46.65%) revised their judgments post-advice. Whilst this finding in isolation is not particularly meaningful, it is clear from Table 9.2 that participants in receipt of advice (57.16%) proportionately changed judgment more often than people who did not receive advice (28.92%). Further, it appears that participants in receipt of Population estimate advice proportionately revised their judgments to a lesser extent (41.88%) than either GDP advice (62.22%), or GDP and Population estimate

advice (60%). Possibly, this finding may be attributable to the low *N* in the Population estimate advice condition. However, in order to determine whether there were differences between conditions in terms of ordinal accuracy, it was necessary to ascertain the strength of association between participants pre-advice ordinal accuracy, and compare it with participants' post-advice ordinal accuracy. Hence, Spearman's r correlation coefficients were calculated between the true rank order of countries (by IMR), and those reported by participant's pre-advice, and post-advice (see Table 9.3).

Table 9.3 Ordinal accuracy: Strength of association between pre-advice rankings and true rank order of 10 countries (IMR), and post-advice rankings and true rank order of 10 countries (IMR).

Advice	Participants *N*	Strength of association between pre-advice ranks and true rankings	*p*	Strength of association between post-advice ranks and true rankings	*p*
Control (no advice)	44	r_s = - 0.286	0.0005	r_s = - 0.432	0.0005
GDP	36	r_s = - 0.193	0.0005	r_s = 0.263	0.0005
Population estimate	16	r_s = 0.012	0.877	rs = 0.221	0.0005
GDP and Population estimate	22	r_s = - 0.071	0.298	rs = 0.022	0.747

Table 9.3 reveals some interesting results, in so far as participants' rankings of the ten countries (by IMR) appear to vary prior to exposure to any advice. Participants allocated to the GDP advice condition, for example, showed a significant, but weak, negative correlation between participants' rankings pre-advice and the true rankings. Participants allocated to the Population estimate advice condition, and the GDP and Population estimate advice condition, report non-significant correlations between their pre-advice rankings and the true rank order of

the ten countries (by IMR). Post-advice, participants allocated to the GDP advice condition, and the Population estimate advice condition, showed significant correlations between their post-advice rankings, and the true rank order of the ten countries (by IMR). However, people allocated to the GDP and Population estimate advice condition, did not report a significant correlation between their rankings, and the true rank order of countries by IMR, post-advice.

To determine if ordinal accuracy improves for any of our participants, it was necessary to test whether the correlation between participants' rankings of countries by national IMRs, and the true rank ordering of countries by national IMRs, remained statistically non-significant post-advice. Traditionally, this test of non-independent correlation coefficients has been addressed by Hotelling's 't' (1931). However following the work of Williams (1959), Dunn and Clark (1969), and Steiger (1980), it is clear that 't' can be overestimated when computing Hotelling's statistic in some circumstances. Instead, the statistic suggested by Steiger is utilized – Z_1. This is a test of whether the correlation coefficient (pre-advice rankings compared to true rank order), is the same as the correlation efficient (post-advice rankings compared to true ranks). Table 9.4 below reports the results of the significance tests.

Table 9.4 Significance test of correlations pre and post advice.

Advice	Z_1
Control (no advice) ($N = 44$)	3.40606
GDP advice ($N = 36$)	- 7.301668
Population estimate advice ($N = 16$)	- 2.590401
GDP and Population estimate advice ($N = 22$)	- 1.228691

Examining the critical values from a Z distribution shows that for significance of $p < 0.05$ to be reached, the test statistic must exceed plus/minus 1.96 for a two-tailed test, or plus/minus 1.65 (in the hypothesized direction) for a one-tailed test. Clearly, Table 9.4 shows that the correlations between participants' rankings of countries (by IMR) and the true rank order, pre-advice was not the same as the correlation post-advice - for people who did not receive advice, for people in receipt of GDP advice, and people in receipt of Population estimate advice. For participants not in receipt of advice, it is possible to infer that, as the correlation coefficient between pre-advice rank ordering of countries by IMR, and the true rank order, is greater than the correlation coefficient post-advice, the strength of association between participants' rankings of countries by IMR, and the true rank order, became weaker for people who did not receive any advice (controls). This is because the correlation coefficient between pre-advice rank ordering of countries by IMR, and the true rank order is negative, and post-advice the correlation coefficient remains negative but larger (see Table 9.3). However, the strength of association between participants' rankings of countries by IMR, and the true rank order, became stronger for participants in receipt of GDP advice, and participants in receipt of

Population estimate advice. It is not appropriate to infer that the correlation between pre-advice rank orderings, and the true rank order, and the correlation between post-advice rank orderings and the true rank order, are significantly different for participants in receipt of GDP and Population estimate advice, as the test statistic (Z_1 = - 1.228691) does not exceed the critical value for Z_1.

In sum, the strength of association between participants' rank ordering of countries by IMR, and the true rank order became stronger (i.e. ordinal accuracy improved) for participants in receipt of GDP advice, and participants in receipt of Population estimate advice. Recall that the hypothesis being tested concerns whether people in receipt of GDP advice are able to improve their ordinal accuracy to a greater extent than people in receipt of Population estimate advice. This prediction was tested by comparing the post-advice correlation coefficient (i.e. the strength of association between participant's post-advice rankings and the true rank order of countries by IMR) of people in receipt of GDP advice, with the post-advice correlation coefficient of people in receipt of Population estimate advice (i.e. to determine which set of participants' rankings are more closely associated with the true ranking). This test revealed that it is inappropriate to reject the null hypothesis that the two independent correlations are equal (z = 0.465, $p < 0.321$, 1 tail). In sum, the hypothesis concerning ordinal estimation accuracy is partially supported. Participants who were in receipt of no advice (controls) appear to have become more error prone, in terms of ordinal accuracy, by revisiting the estimation problems. Participants in receipt of GDP advice, and participants in receipt of Population estimate advice, significantly improved their ordinal accuracy by about the same amount, on average, whilst participants in receipt of GDP and Population estimate advice did not significantly improve (or degrade) their ordinal estimation accuracy. Consideration is next given to metric accuracy.

Recall that metric accuracy can be defined as the absolute numerical difference between a participant's unaided estimate of some quantity and the true value of the quantity, and a participant's post-advice estimate, and the true value of the quantity. Should the absolute numerical difference between a participant's unaided estimate of some quantity and the true value of the quantity be greater than the latter, then it is possible to infer that estimation accuracy has improved. It was hypothesized that participants in receipt of solely Population estimate advice would not significantly improve their post-advice accuracy; that participants in receipt of GDP and Population estimate advice would achieve a greater degree of post-advice accuracy than participants in receipt of Population estimate advice, but not exceed the accuracy achieved by participants solely in receipt of GDP advice. To test our hypothesis, two measures of metric accuracy were utilized (Lee, 2007, unpublished). First, Signed Order of Magnitude (SOME) error was calculated for each participant. SOME was derived as the Log_{10} of each participants IMR estimate, after dividing the estimate by the actual IMR of the country being estimated. Positive SOME values indicate overestimation, and negative SOME values indicate underestimation, on average. A second measure of metric accuracy was derived as Order of Magnitude Error (OME). Here, OME is the absolute SOME score for each participant. Mean OME is a measure of absolute accuracy, so that an OME = 1.0 describes an error of one order of magnitude. SOME scores satisfied the assumption that our data were normally distributed, but the OME scores were not normally distributed. Hence, non-parametric tests are reported for data that is not normally distributed. First, I examined the proportion of participants who revised their metric estimate (SOME scores) post-advice.

Table 9.5 Metric accuracy revisions (SOME) scores

Condition	Number of metric revisions
No advice Controls (N = 44)	283/438 (64.61%)
GDP advice (N = 36)	262/360 (72.78%)
Population estimate advice (N = 16)	84/151 (55.63%)
GDP & Population estimate advice (N = 22)	158/220 (71.82%)
All (N = 118)	550/1179 (67.32%)

Table 9.5 shows that overall approximately two-thirds (67.32%) of participants revised their post-advice point estimate of a countries' IMR. Interestingly, even participants who did not receive advice revised their estimates (64.61%) more often than not. Participants in receipt of GDP advice revise their judgment (72.78%), marginally more so than participants in receipt of GDP and Population estimate advice (71.82%), but both conditions contain more judgment revisions than participants in receipt of Population estimate advice (55%). On average, inspection of the data presented in Table 9.4 and 9.5 shows that participants were more disposed to revise their judgment when considering a point estimate of a countries' IMR, than when considering the rank order of countries according to their IMR. Whilst this finding may be anticipated due to the 'fine grained' nature of point estimation, in comparison to the 'blunt' nature of rank ordering by IMR, it is of interest to determine if the accuracy improvements

suggested by the analysis of the ordinal data are mirrored in the analysis of the metric data. Analysis of SOME scores and OME scores is tabulated in Table 9.6 and 9.7.

Table 9.6 and Table 9.7 show that participants were unable to significantly reduce their estimation error, and hence improve their metric estimation accuracy. On average, all participants overestimate the IMRs of the ten target countries, both prior to the receipt of advice, and post-advice. Participants in the no advice (control) condition make an even larger overestimate of the IMRs of the ten target countries post-advice than they did pre-advice (although this difference did not reach significance). Where participants were in receipt of advice, they made a smaller overestimate of the ten target countries IMRs post-advice than pre-advice, but not significantly so. Inspection of the mean OME scores in Table 9.4 shows that participants non-significantly increased their error pre to post advice when in receipt of GDP and Population estimate advice, but all other participants non-significantly reduced error pre to post advice. In sum, taking into account measurements of ordinal accuracy and metric accuracy, there is only partial support for this hypothesis. It is clear that participants in receipt of GDP advice are able to significantly improve their ordinal, but not their metric accuracy. However, participants in receipt of GDP and Population estimate advice were unable to improve either their ordinal or metric accuracy, whilst participants in receipt of Population estimate advice were able to improve their ordinal accuracy, but not their metric accuracy.

Table 9.6 Metric accuracy – SOME scores

Advice	N	SOME $\bar{x}$ (Pre-advice)	SOME $\bar{x}$ (Post-advice)	*t*	*df*	*p*
None (control)	44	0.05 (sd = 0.77)	0.07 (sd = 0.65)	- 0.52	43	0.61
GDP	36	0.19 (sd = 0.38)	0.17 (sd = 0.41)	0.76	35	0.45
Population	16	0.19 (sd = 0.31)	0.20 (sd = 0.23)	0.21	15	0.84
GDP & Population	22	0.17 (sd = 0.39)	0.18 (sd = 0.41)	0.28	21	0.78

Table 9.7 Metric accuracy – OME scores

Advice	OME $\bar{x}$ (Pre-advice)	OME $\bar{x}$ (Post-advice)	N	Z	*t*	*df*	*p*
None (control)	0.71 (sd = 0.49)	0.66 (sd = 0.35)	44	1.71			0.09
GDP	0.52 (sd = 0.21)	0.49 (sd = 0.26)	36		1.65	35	0.11
Population	0.58 (sd = 0.32)	0.51 (sd = 0.32)	16	- 1.15			0.25
GDP & Population	0.51 (sd = 0.25)	0.52 (sd = 0.26)	22		0.38	21	0.71

The results reported so far paint an inconsistent picture so far as ordinal and metric accuracy are concerned. All participants overestimate the actual IMR of each target country on average, both prior to, and post-advice. No one group of participants was able to improve metric accuracy, and only participants in receipt of GDP advice, or Population estimate advice, were able to improve their ordinal accuracy. These results are somewhat puzzling, as the cue validities for GDP advice and Population estimate advice, are clearly different. When advice was linearly regressed upon the true IMRs for the ten countries utilized in the experimental task, ($F_{(2, 7)}$ = 30.57, p < 0.0005; Adjusted R^2 = 0.87), only GDP advice emerged as a significant predictor (beta = - 0.004, t = - 7.55, p < 0.0005). Put another way, there is a significant correlation between GDP advice and the true value of the ten target countries' IMRs, on average (r_s = - 0.63, p < 0.05). However, there is no significant correlation between Population estimate advice and the true IMRs of the target countries, on average (r_s = - 0.067, p < 0.85).

Further support for the idea that participants were insensitive to the cue validities of advice was found when consideration was given to the amount of judgment change between pre and post-advice IMR estimates. The absolute numerical difference between a participant's pre-advice IMR estimate, and post-advice IMR estimate was calculated for each question, and then aggregated for the purposes of comparisons between conditions (see Table 9.8). These scores did not satisfy the assumption that the data was normally distributed, hence the Log_{10} transformed scores were entered into tests of significance.

Table 9.8 Mean absolute difference scores (IMR)

Condition	*N*	$\bar{x}$	*sd*	Log_{10} $\bar{x}$	*sd*
No advice (control)	38	25.45	36.08	0.97	0.70
GDP	31	33.84	51.28	1.23	0.60
Population estimate	12	28.27	16.80	1.31	0.47
GDP and Population estimate	21	34.34	43.96	1.24	0.55

A one-way ANOVA was performed on the data in Table 9.8 in order to determine if there were any significant differences in the amount of judgment change attributable to experimental condition – no evidence of a main effect of condition was found ($F_{(3, 98)} = 1.63$, $p < 0.19$). Further, no evidence of differences in the absolute amount of judgment change was found when follow-up planned comparisons were performed on the data.

One resolution to the difficulty of explaining the preceding results lies in the possibility that participants become more error prone post-advice in terms of metric accuracy. This might be so, as participants in this study were informed of the true range of possible IMR values at the outset. In order to investigate this possibility, the absolute difference between a participant's unaided estimate, and the true IMR, was calculated for each question and aggregated for the purposes of comparison. Similarly, the absolute difference between a participant's post-advice estimate, and the true IMR was calculated, and aggregated for the purposes of comparison. Marginal (non-significant) increases in estimation error appear to be common to participants in receipt of advice (see Table 9.9).

Table 9.9 Pre-advice and post-advice IMR accuracy

	Observations N	Error prior to receipt of advice $\bar{x}$	*sd*	Error post-advice $\bar{x}$	*sd*
No advice (N = 44)	439	96.35	149.52	88.32	130.81
GDP advice (N = 36)	360	55.38	67.83	60.07	95.44
Population estimate advice (N = 16)	160	56.43	75.36	56.66	76.29
GDP and Population estimate advice (N = 22)	220	67.54	107.91	69.23	121.49

The table above shows that participants in receipt of advice increased their estimation error post-advice (albeit such increases in estimation error are statistically non-significant). This may suggest that participants were susceptible to a conformity effect, where judgment revision is towards the available advice (metric) even though estimation revision is not beneficial. To investigate this issue further, the correlation between advice and rank order of countries prior to, and post-advice was considered.

If people are unable to adequately discriminate between good and poor advice, it might be expected that participants in receipt of advice will strengthen the association between advice, and participants rank orderings of the ten target countries post-advice. The correlation

coefficients between advice and participants rank order of the ten target countries is shown in Table 9.10. Inspection of the data in Table 9.10 shows support for the idea that people are conforming to advice irrespective of its quality. This is because the correlation coefficient representing the strength of association between advice and the rank order (by IMRs) of the ten target countries is greater post-advice, in all advice conditions.

Table 9.10 Correlations between rank order of question items and advice

	N	r_s
Rank order pre-advice and GDP advice	36	0.674
Rank order post-advice and GDP advice	36	0.891
Rank order pre-advice and GDP and Population estimate advice	22	0.645
Rank order post-advice and GDP and Population estimate advice	22	0.777
Rank order pre-advice and Population estimate advice	16	0.168
Rank order post-advice and Population estimate advice	16	0.228

Next, a test of related correlation coefficients was performed (Steiger's Z_1), which tests the hypothesis that the correlation between the pre-advice rank order of question items, is the same as the correlation between the post-advice rank order of question items and advice.

Table 9.11 Significance tests of related correlation coefficients

Advice	Z_1
GDP advice (N = 36)	- 3.479758
Population estimate advice (N = 16)	- 1.30168
GDP and Population estimate advice (N = 22)	- 0.3792976

Although all participants in receipt of advice strengthened the association between rank order of the ten target countries and advice post-advice, only participants in receipt of GDP advice did so significantly. This is because the test statistic (Z_1) must exceed plus/minus 1.96 for a two-tailed test, or plus/minus 1.65 (in the hypothesized direction) for a one-tailed test, from the critical values in a Z distribution for significance of $p < 0.05$ to be reached. In part, the non-significant results of the preceding significance tests of related correlation coefficients may be explainable by the different numbers of participants in the three advice conditions, which could have resulted in low statistical power. However, it is of interest to determine why participants in receipt of Population estimate advice were able to improve their ordinal accuracy (despite a relatively low N = 16).

One remaining issue is to determine why participants in receipt of non-predictive advice (Population estimate advice), were able to improve their ordinal accuracy, to the same extent as participants in receipt of 'predictive advice' (GDP advice). One possible explanation for this finding concerns individual differences in the degree of 'world' knowledge that participants held prior to the experiment. Although no measure of prior knowledge held by participants of

the target countries, utilized in the experiment, was taken a priori, inspection of the distribution of age ranges between the experimental conditions revealed that the participants in receipt of Population estimate advice were older, on average, than participants in the other experimental advice groups (see Table 9.12). This may suggest that this group was able to bring a greater degree of 'world' knowledge to bear on the estimation task (if one accepts that greater life experience results in greater world knowledge), than other participants, and subsequently be able to improve ordinal estimation on the basis of Population estimate advice.

Table 9.12 Distribution of age ranges

Advice	N	Age $\bar{x}$	Age sd	Age Median	Age Range
None (control)	44	23	2	23	9
GDP	36	20	7	19	40
Population	16	32	11	32	28
GDP and Population	22	29	12	27	45

The age data did not meet the assumptions of a normal distribution, hence a non-parametric significance test was performed in order to determine if the differences in age ranges between advice groups were non-random ($\chi^2 = 18.81$, $df = 2$, $p < 0.0005$). Consideration was next given to whether the advice groups differed in shape or location in terms of age. Table 9.13 shows that there were significant differences in the shape and location of the distribution of age data between groups ($\chi^2 = 12.44$, $df = 2$, $p < 0.002$). Inspection of the data in Table 9.13 shows that

the main difference was between people in receipt of GDP advice, and people in receipt of Population estimate advice.

Table 9.13 Frequency of age data

	GDP advice	Population estimate advice	GDP and Population estimate advice
> median	9	11	13
<= median	27	4	9

This account is leant further support when consideration is given to the frequency of judgment change data. If participants were sensitive to the cue validities of each type of advice that was provided to them, it might be expected that judgment change would be a function of advice type. To explore this possibility the number of judgment changes reported by participants were calculated, and it is clear from Table 9.14 that participants in receipt of advice changed judgment more often than not. Moreover, people in receipt of GDP advice, and GDP and Population estimate advice, changed judgment more often than people in receipt of solely Population estimate advice ($\chi^2 = 14.80$, $df = 2$, $p < 0.001$). In conjunction with the age range data, and the frequency of judgment change data for people in receipt of Population estimate advice, suggests that participants were changing judgment approximately 50% of the time (i.e. participants were no more likely to change judgment, than not change judgment); hence it is difficult to discern the influence of advice upon judgment change. Rather, this group may have been able to improve ordinal accuracy through possession of a higher degree of 'world' knowledge than the other experimental advice groups.

Table 9.14 Number of judgment changes

	Change Judgment	Do not change judgment
No advice (N = 44)	266	174
GDP advice (N = 36)	255	105
Population advice (N = 16)	86	74
GDP and Population advice (N = 22)	149	71

Further support for a 'conformity to advice' effect is found when the perceived utility of advice measure was examined. Recall that it was hypothesized that should participants be able to adequately distinguish between 'good' and 'poor' advice, they should be able to differentially rate the available advice on the dimension of perceived utility. Should participants be sensitive to the cue validities of objectively good and poor advice, then this sensitivity should be reflected in the ratings participants attribute to the perceived utility of the advice, on average.

Table 9.15 Frequencies of perceived utility of advice scores

	GDP advice	Population advice	GDP and population advice
> median	17	3	12
<= median	19	13	10

Table 9.15 shows the frequency of median scores above, or below, the overall median (= 3), irrespective of advice group membership. Inspection of Table 9.14 reveals that there are a greater number of scores that below the overall median value, than above it, for people who were in receipt of Population estimate advice. This finding could suggest that people valued Population estimate advice to a significantly lesser extent than they valued either GDP, or GDP and Population estimate advice. However, a Median test confirmed that there were no significant differences in location or shape, in the median ratings attributed by participants between groups ($\chi^2 = 5.29$, $df = 2$, $p < 0.071$). This finding suggests that participants were largely insensitive to the cue validities of the available advice.

9.4 Discussion

The focus of this study was to determine if people are able to differentiate between objectively 'good' and 'poor' advice, in order to improve ordinal and metric estimation accuracy. Previous research (Gigerenzer, Todd and the ABC Research Group (1999); Lee [unpublished], 2007) theorized that people are able to recognize predictive information in an estimation task, and recall cues that enable individuals to rank order target items (ordinal accuracy). These authors go on to argue that the cognitive processes involved in ordinal estimation are independent from the processes that facilitate estimation that ultimately resolves in measures of metric accuracy. Lee (unpublished, 2007) supported this position, finding that participants tasked with estimating IMRs were able to improve their ordinal accuracy when in receipt of predictive information (accurate GDP information for target countries, or accurate GDP and Population estimate information for target countries), but participants were not able to improve ordinal accuracy when in receipt of non-predictive information (accurate Population estimate information for target countries). Participants in this study reported no improvements in metric accuracy, but simultaneously reported improvements in ordinal accuracy, both for participants

in receipt of predictive advice (GDP information), and for participants in receipt of non-predictive advice (Population estimate information). One possible explanation for the findings reported here is that conforming to advice is a good strategy for accurate estimation of national IMRs when the advice is predictive of the target quantity. Not conforming to advice appears to be a good strategy for ordinal estimation when an individual can bring 'world knowledge' to bear on the issue.

It may be true that the results of the analysis of ordinal accuracy may be confounded in this study by the fact that a concentration of older people (arguably with more developed world knowledge) appeared in the experiment. However, as discussed in the introduction to this study, participants may attend to poor advice for good reasons. It is possible that participants in receipt of Population estimate advice, found such advice unfamiliar, but were able to utilize the advice to enrich the basis for ordinal estimation (Heath and Gonzales, 1995). Alternatively, interacting with Population estimate advice may have caused participants to think of the estimation problem in a different, and potentially more beneficial, way (Schotter, 2003; Druckman, 2001). Moreover, people seek out, and pay attention to advice, not only to improve the accuracy of their decisions, but to share responsibility for those decisions; and hence participants may have been reticent to reject Population estimate advice that they perceived to be freely given (Sniezek and Buckley, 1995; Harvey and Fischer, 1997; Yaniv, 2004a; Yaniv 2004b).

Of interest then is why the results of this study differ markedly from the results obtained by Lee (2007, unpublished). Participants in this study did not show improvements in either metric or ordinal accuracy when in receipt of GDP and Population estimate advice (N= 22). Recall that Lee's participants reported improvements in ordinal accuracy (Study 4) for participants in

receipt of GDP and Population estimate advice (N = 24); such ordinal accuracy improvement did not surpass the accuracy improvement of participants solely in receipt of GDP advice. The main difference between the two experimental procedures is that the actual range of IMRs (2-147 per 1000 live births) was unknown to participants in Lee's studies (where metric accuracy was recorded in Study 1-3). Lee's participants were instructed to respond on a scale anchored at one end by zero (no infant mortalities), and at the other by 1000 (all children die). Hence, it is not surprising that (i) Lee predicted that people would generally overestimate IMRs, and (ii) people generally did overestimate IMRs. Under these conditions metric accuracy will almost inevitably be poor – any estimation strategy based upon central tendency (mean, median) will invariably lead to overestimation - as the true range of IMRs for the countries included in Lee's study, is in the extreme lower 20% of the response range that participants were informed of (i.e. 0-1000). Conversely, at one level ordinal accuracy is unaffected by the response scale used to determine metric accuracy (participants were tasked with ordering a list of countries by IMR). Lee argues that his results demonstrate that people are unable to define an ecologically valid range of values for target items where domain knowledge is weak, and this largely accounts for poor metric accuracy.

However, an alternative account might posit that participants may have been inhibited from forming an ecologically valid response range for target items, because participants were cued by the experimenter that the range of true values for each country's IMR was somewhere between 0 and 1000 – a far wider range than the actual range of values (2-147). Where people have incomplete domain knowledge and perhaps little confidence in their own intuition; they may be susceptible to the demand characteristics of the experimental situation. Hence, being told that the response range of quantities to be estimated is 0-1000 could induce participants to believe, that the values of the target items to be estimated, are normally distributed about the arithmetic mean (i.e. 500) of the known range. Under these circumstances non-ordinal

accuracy measures capture both metric accuracy, and an experimenter bias. In contrast, participants in this study were informed of the true range of IMRs of target countries, hence relaxing the constraints upon the formation of an ecologically valid response range. However, participants in the present study overestimated the actual IMRs of target countries both prior to the receipt of advice, and post-advice. Hence, if the instruments designed to capture measurements of accuracy are not well calibrated, the disassociation between ordinal and metric accuracy, may not be so clear-cut. This is because participants in this study strengthened the association between the rank order of target countries, with advice, post-advice (although only participants in receipt of GDP advice did so significantly). This finding shows that participants in the current study utilized advice to inform their deliberations of the rank order of the target countries, but did not have to formulate a valid response range (this was provided to participants). This difference in experimental design between the current study and Lee (unpublished 2007), may account for the fact that in the current study participants in receipt of GDP and Population advice did not show improvements in metric or ordinal accuracy, while participants in Lee's experiment (Study 4) did show improvements in metric and ordinal accuracy (group *N*s were similar). Support for this idea is found when the data from participants in receipt of GDP and Population estimate advice was analyzed – people appeared to be insensitive to the asymmetries of advice information, and hence were not able to recognize, (or if they did recognize were not able to use) the benefit of GDP advice, in comparison to Population estimate advice. This idea appears consistent with the work of Harvey *et al*. (1999) who make an identical argument that people may be able to assess the utility of advice, but may not be able to use such assessments to beneficially utilize advice. Such a position is consistent with the idea of bounded cognition, and points towards the practical benefits of decision aids, or interventions, that lighten the demands on cognition in estimation tasks. The issue of perception of asymmetries in the quality advice is considered next.

Central to the argument that people are able to differentiate between objectively 'good' and 'poor' advice, is the idea that people recognize the predictiveness, or utility of advice, and are subsequently able to either incorporate good advice into their deliberations, or at least reject poor advice. However, there was little support for this position found in this study. Participants were tasked with providing a rating of the perceived utility of advice on a 5-point Likert-type scale, and no differences in perceived utility of advice ratings were found between experimental advice conditions. In sum, the findings from this study show that the people were not sensitive to asymmetries in the quality of advice.

Perhaps the most important limitation to this study (and one that is shared by Lee's Study) is that currently, there is no comprehensive paradigm through which ordinal accuracy can be studied. In this study, the conventional estimate–advice-estimate format, that much JAS research is predicated upon, has been utilized. However, it is difficult to develop a study that captures the nuances of ordinal estimation in such a paradigm, and subsequently current understandings of the cognitive processes that underpin ordinal estimation are incomplete. Clearly, further work in this area is necessary if a fuller understanding of the cognitive architectures that facilitate ordinal estimation is to be developed.

A further limitation of this study is that individual differences between participants have not been sufficiently controlled for in this study, and/or the random allocation of participants to experimental treatment conditions may not have been successful. Recall, that ordinal estimation accuracy improved for participants in receipt of Population estimate advice, despite the fact that such advice was not predictive of national IMRs. This finding suggests that some measure of 'world' knowledge, or specific domain knowledge should be recorded prior to exposure to advice, in order to exert greater experimental control. This may suggest a

potentially fruitful avenue, in terms of directions for future studies, as individual differences may mediate the cognitive mechanisms that facilitate both metric and ordinal estimation.

In conclusion, under conditions of uncertainty people are susceptible to the influence of both good and poor advice, and may conform to either. This may imply that further research could fruitfully be employed in discovering if, how, and to what extent, judges and estimators search out information about the potential utility that advisors may offer, above that of credibility of the source of information.

Chapter 10: Conclusions, suggestions for methodological improvements, and directions for future research.

10.1 The central problem addressed in the current work

The central problem addressed in the current work concerns the fallibility of human judgment, and understanding some influences upon human judgments that potentially result in beneficial judgment revision. Here, a fundamental assumption is that judgmental fallibility is based upon the limitations of cognition. Further, it is important to distinguish between the cognitive processes that underpin choice, and the processes that facilitate judgment, as there is strong evidence that each of these processes is supported by different cognitive architectures (Billings and Scherer, 1988). The focus of the current work is human judgment.

Some researchers have taken a pessimistic view of human judgment, seeing people as 'cognitive misers' (Simon, 1982), who base judgment upon a sub-optimal 'satisficing' strategy. Here, both due to the limitations on individual cognitions and the practicalities of comprehensive information search (resources and time), people often accept the first acceptable solution to any particular task involving judgment. Other researchers have catalogued how people making judgments are susceptible to cognitive biases, because of the use of cognitive shortcuts, or heuristics (Gleitman *et al*. 2004; Tversky and Kahneman, 1982). Much of the time the individual use of heuristics results in acceptable judgmental outcomes that satisfy the environmental context, but in some circumstances using heuristics may result in errors that have important consequences (safety, medicine, commerce etc.).

Not all researchers take such a pessimistic view of human judgment however. Instead of judgment being prone to systematic cognitive biases, these researchers argue that judgment

should be conceived as an adaptive system, measured in terms of its 'fit to reality' (Gigerenzer, 2006). Under this rubric measuring human judgment in terms of probabilistic reasoning is mistaken, as probability theory cannot apply to single events. Moreover, these researchers see this adaptive system of judgment as 'fast and frugal' - judgments are fast because they are frugal in terms of what information is taken into account.

Given the preceding discussion over the fallibility of human judgment, it is perhaps encouraging to acknowledge the advantages of simple interventions designed to enhance human judgment in comparison to holistic estimation (Dawes, Faust and Meehl, 1989; Grove *et al*. 2000; Grove and Meehl, 1996). This programme of research suggests that the condition of human judgment can be improved. Moreover, people intuitively seek to improve the quality of their judgments by the solicitation of advice. People often seek out advice in circumstances where the resource implications for comprehensive information search are prohibitive. People may accept and act upon advice both to improve the quality of judgment and for social reasons. The social context of advice giving and taking is an underdeveloped area of enquiry in the extant literature, and hence, the problem addressed in the current work is to determine some of the factors that are influential in beneficial judgment revision (or at least to point to circumstances where some factors are not influential in beneficial judgment revision) when an advisor gives advice to a judge. To that end, the current work empirically examined advice giving and taking (defined as information exchanged between an advisor and judge in a single episode), in the context of JAS research and algorithmic decomposition (knowledge of the process of advice generation), prior to attempting to ascertain whether, in fact, people are able to recognize and utilize objectively 'good' advice.

10.2 Summary of the main findings of this investigation

The results of the investigations conducted in the current work can be grouped under a number of headings. These are summarized below -

Advice preference

- People prefer reasons-based advice to solely numeric advice, and are sensitive to advice evidenced by authority (Chapter 4, p105-12, p115-118, p124-27).

Depth of information search

- People search for information in a non-random way (Chapter 4, p133), but do not hold preconceived ideas about the utility of advice messages (Chapter 4, p131-32).

- People search for decision relevant information to the same extent - irrespective of whether people consider advice or not. People also give greater consideration to reasons-based advice (Chapter 4, p133-34).

Confidence

- Attending to advice bolsters peoples' post-advice confidence in their estimates (Chapter 4, p134-35). There is some evidence that changes in the levels of confidence are attributable to the type of advice people attended to (Chapter 4, p136).

Perceived quality of advice

- People seem insensitive to the quality of the advice they prefer, and perceptions of the utility of advice are not predictive of subsequent judgment revision (Chapter 4, p117-9, Chapter 5, p165, Chapter 6, p211-13, Chapter 7, p261-62, Chapter 8, p288, Chapter 9, p326, p328).

People cannot easily discriminate between potentially beneficial and poor advice

- People are not able to easily tell the difference between 'good' and 'bad' advice (Chapter 4, p120, Chapter 5, p171-72, Chapter 6, p206-8, p213, Chapter 7, p264-65, Chapter 8, p288, Chapter 9, p328).

Cognitive weight placed upon advice

- People are resistant to advice and prefer their own judgments. However, where people do not discount advice, they tend to conform to advice indiscriminately (Chapter 5, p166-69).

Propensity to revise judgment

- People in receipt of advice revise their judgment more often than people who do not consider advice (Chapter 4, p110, Chapter 5, p165, Chapter 6, p201-5, Chapter 7, p260-61).

Direction of judgment revision

- People who preferred reasons-based advice tended to change judgment in the 'direction' of advice, on average - irrespective of the sub-type of advice (Chapter 5, p171-76, Chapter 6, p208-13, 221-22, Chapter 7, p263). People who preferred solely numeric advice did not significantly revise judgment in the 'direction' of advice (Chapter 4, p113-14).

Influence of knowledge of the process of advice

- Where people use algorithms to arrive at an estimate of some large quantity, their post-advice estimates are likely to be greater than their pre-advice estimates Chapter 6, p198-200, Chapter 7, p259-60). Where algorithms are utilized to estimate relative frequencies no estimate 'inflation' is evident (Chapter 8, p285-86).

- In terms of the extent of judgment revisions, people are not sensitive to the quality or complexity of the algorithms utilized (Chapter 6, p209-10, Chapter 7, p262-64).

- However, there is some evidence that people are sensitive to the proximity of the value of an individual's unaided estimate to the numerical value of advice, in terms of a propensity to revise judgment (Chapter 7, p266-67).

Determinants of accuracy

- Acting upon advice may not help people make more accurate estimates than people who did not attend to advice. This is because where an individual's estimate falls between advice and the true answer, people tend to revise their judgment in the direction of advice (irrespective of its quality), suggesting that people indiscriminately conform to advice (Chapter 4, p118-20).

- Factors which influence ultimate estimation accuracy include initial estimation accuracy, the accuracy of advice, and post-advice confidence (Chapter 6, p238-41, Chapter 7, p269-70, Chapter 8, p291-3).

Determinants of judgment revision

- Most predictor variables that may have predicted judgment revision proved to be non-significant, however there is some evidence that the perceived utility of advice is influential (Chapter 6, p242), particularly when estimating relative frequencies (Chapter 8, p293-97).

People are constrained in their abilities to assess and use advice in their deliberations even when advice is objectively 'good'.

- People in receipt of advice (irrespective of quality) revise judgment more so than people not in receipt of advice (Chapter 9, p315-16, p321).

- People overestimate target quantities expressed as relative frequencies (Chapter 9, p322).

- People are successful in improving accuracy when tasked with rank ordering target quantities (ordinal accuracy), and in receipt of a single advice type (irrespective of the quality of advice). Ordinal accuracy does not improve where an individual is in receipt of advice that is predictive of a target quantity *and* advice that is not predictive of a target quantity (Chapter 9, p322-25).

- Point estimates of quantities improve when advice is utilized. Where 'compound' advice (both predictive and non-predictive advice) is utilized, metric accuracy does not improve post-advice (Chapter 9, p325). However, there are no significant differences, on average, between advice types when the absolute amount of judgment change is considered (Chapter 9, p326).

10.3 Relationship between the results reported here and the extant advice giving and taking literature.

Advice preference

The result of the exploratory study in Chapter 4 of this thesis indicated that people prefer reasons-based advice to solely numeric advice, and are sensitive to advice evidenced by authority (Chapter 4, p105-12, p115-118, p124-27), in a single judge-advisor interaction. It is clear that devoid of preconceptions of the nature of any available advice, people prefer reasons-based advice, to solely numeric advice, when deliberating over an issue that may be troublesome. Research by Dalal and Bonaccio (2010) supports this position. In two experiments these authors presented participants with various fictional scenarios such as deciding which job to apply for. Participants were then offered various different permutations of advice (a recommendation for a particular job, a recommendation against a particular job,

advice on how to make the decision, sympathy over the predicament or information about the jobs under consideration). Participants consistently preferred information about the jobs under consideration. A second experiment introduced more varied decision-making scenarios, where advice could be solicited from an expert (with domain relevant knowledge). Results indicate that participants preferred information in addition to recommendations about the decision scenarios. These findings compliment my own, as my participants overwhelmingly preferred reasons-based advice, in preference to solely numeric advice when given the choice. The reasons why people may consistently prefer 'elaborated' advice are further discussed next.

People may prefer information rich and elaborated advice (i.e. a recommendation accompanied by reasons) when considering a troublesome estimation issue, because processing solely numeric information is cognitively demanding. In order to reduce the processing load upon human cognition, that consideration of solely numeric information may require, people may seek out environmental cues (Yalch and Elmore-Yalch, 1984). In other words, solely numeric advice may predispose people to attend to peripheral cues to determine the veracity and utility of advice, in preference to expending significant cognitive resources upon centrally processing the contents of the advice message (Petty and Cacioppo, 1986). The results reported from experimentation in this thesis are consistent with the insights provided by Yalch and Elmore-Yalch (1984), in so far as an important environmental factor in judgment revision highlighted in Chapter 4, was the credibility of the source of advice (Reinard, 1988). This factor loomed large in the deliberations of participants, and is significant in so far as it can be seen to relax the cognitive demands of estimation - in part by allowing participants to share responsibility for the outcome of the estimate through the perceived expertise of the advisor (Harvey and Fischer, 1997). This insight suggests that judgment revision maybe unlikely when participants are exposed to advice that is high in quantification. This has implications for JAS researchers, who up till now have operationalized advice largely in solely numeric terms.

Depth of information search

In terms of depth of information search, participants in Chapter 4 of this thesis indicated that they held no preconceptions about the utility of advice messages initially, but gave considerable attention to reasons-based advice, irrespective of whether they subsequently revised judgment post-advice. Indirectly, this finding may support the work of Harvey *et al*. (2000), as participants appear to have assessed advice, but have been unable to use the advice to successfully improve performance. I argue a similar point here – my participants may have found the task in Chapter 4 of this thesis cognitively demanding, and despite a thorough information search may not have been able to subsequently use advice to improve performance. In sum, my participants may have applied significant cognitive resources to information search and evaluation, which left them without sufficient cognitive resources to combine advice with their unaided intuitions in order to improve performance. This finding is partially echoed in the extant literature, in so far as task complexity has been shown to decrease information search and induce simplification strategies in order to conserve individual's cognitive processing resources (Schrah *et al.* 2006; Johnson and Bruce, 2008). However, participants in my study were sensitive to peripheral cues (such as the credibility of the source of advice) which may indicate that my participants *were* adopting simplifying strategies as a means of reducing task related cognitive demands. Further, evidence from the extant literature of consumer psychology may suggest that, where people are not highly motivated to critically evaluate the central elements of an attempt at persuasion, numerate advice is likely to be perceived as cognitively demanding. Hence, people may seek to conserve cognitive resources by attending to peripheral environmental cues, in place of centrally processing advice, in order to determine the veracity and utility of such advice (Yalch and Elmore-Yalch, 1984; Petty and Cacioppo, 1986).

Confidence

Participants' levels of confidence were bolstered by attending to advice (Chapter 4, 134-45), which is consistent with the findings of Cooper (1991), who observed that less confident judges have been observed to seek advice to a greater extent, than confident judges, prior to making a final determination in a decision task. There is some evidence, however, that post-advice confidence is not well calibrated with post-advice accuracy (Heath and Gonzales, 1995; Savadori *et al.* 2001). Indeed, it is well established in the extant literature that a perfect match between judgmental accuracy and confidence is not the default position of human cognition (Lichtenstein, Fischhoff, and Phillips, 1982; Alpert and Raiffa, 1982; McClelland and Bolger, 1994; Brenner, *et al.* 1996; Teigen and Jorgenson, 2005). This is particularly so in decision scenarios where judges have limited domain relevant knowledge – as is the case in the current work. Hence, it is perhaps not surprising that post-advice confidence did not reflect the observed level of accuracy achieved by my participants, and that my participants were susceptible to becoming overconfident post-advice (Alpert and Raiffa, 1982, Sniezek and Buckley 1995). Unlike the decision dyads formulated in the current work, overconfidence may be reduced in a situation where advice is solicited from multiple advisors who disagree over the veracity of the advice (Sniezek and Buckley 1995). Ultimately, interactions between judges and advisors involving the exchange of advice may not optimally increase the ultimate quality of judgments, but may increase confidence in those judgments, and hence overconfidence (Heath and Gonzales, 1995).

Perceived quality of advice

One of the most significant findings of the current work is that participants appeared to be insensitive to the quality of advice (irrespective of whether the advice was reasons-based, solely numeric, or algorithmic). Furthermore, perceptions of the quality of advice do not appear to be strong predictors of subsequent judgment revision (Chapter 4, p117-9, Chapter 5,

p165, Chapter 6, p211-13, Chapter 7, p261-62, Chapter 8, p288, Chapter 9, p326, p328). This finding is deleterious to people seeking accurate estimates of some quantity when 'one-shot' advice is poor. However, JAS research has shown that people do have abilities to differentiate between potentially beneficial and deleterious advice where advisors are *repeatedly* solicited for advice. Under this scenario, Yaniv and Kleinberger (2000), investigated the formation of advisors' reputations and found that decision makers update appraisals of advisors asymmetrically. Specifically, a good appraisal was quickly downgraded after bad advice—advisors were susceptible to losing a 'good' reputation. Yaniv and Kleinberger found that decision makers can judge advisors' performance accurately, and purchase information only from good advisors. Luan *et al.* (2004) also found that judges are able to correctly judge the performance of different sources of advice. Further, experiments by Celen, Kariv and Schotter (2005), Kameda and Nakanishi (2003), and Yaniv (Yaniv, 2004a), show that even naïve participants tend to give useful advice, and that social learning generally improves performance. These findings suggest that decision-makers should be particularly wary of non-routine decisions, but if the necessity to solicit advice arises, judges should access advice from advisor(s) with domain relevant knowledge and an established reputation for accuracy. Clearly, non-routine and novel situations are instances of when advice is *most* likely to be sought.

Propensity to revise judgment

Given that people are likely to seek advice in order to improve the quality of their judgments, to reduce the complexity of the decision, and to share responsibility for decision outcomes, it is unsurprising that people in receipt of advice revised judgement post-advice more often than people who did not choose to attend to advice (Chapter 4, p110, Chapter 5, p165, Chapter 6, p201-5, Chapter 7, p260-61). However, this propensity to revise judgment following receipt of

advice did not optimally result in accuracy improvements. These findings – across different task domains – suggest that post-advice judgment revision is dominated by issues of social conformity at the expense of estimation accuracy. Similar conclusions were reached by Harvey and Fischer, (1997), Lim and O'Connor, 1995, Yaniv, (2004), and Yaniv & Kleinberger, (2000) where judges were observed to revise judgement by a sub-optimal 'token shift' of 20-30% in the direction of the position held by advisor(s).

Cognitive weight placed upon advice

Significantly, the data reported here suggests that some people weight their own judgments far more than those of an advisor (Chapter 5, p166-69). In other words, some people are *not* disposed to revise their judgment, on average. Krueger (2003) and Harvey and Harries (2004) explain such reticence on the part of judges to revise their judgments in the light of advice as 'egocentrism' (a long term belief that one's own judgments are superior to those of others – including those of advisors). Egocentrism, according to Krueger (2003), occurs when people are engaged in making judgments about novel situations, which implies that judges are unable to adequately evaluate advice, and adjust judgment accordingly. This explanation is consistent with the results I have reported in this thesis, in so far as people who did not revise their judgments appear to have reported higher levels of pre-advice confidence, on average, than participants who subsequently did revise their judgments post-advice.

Chapter 5 of this thesis considered the Weight of Advice (WOA) measure (Yaniv, 2004a) as a means of quantifying the cognitive weight people placed upon both their own judgments, and the judgments of an advisor. In response to the 14-item historical quiz that participants were tasked with completing in Chapter 5, people discounted advice and preferred their own

judgments on average. However, comparing each experimental condition to the situation where people placed equal weight upon their own judgments and advice (0.5), only participants in receipt of inaccurate numerical advice and very weak reasons significantly discounted advice. This may indicate a 'boundary condition' where the limited plausibility of spurious advice is exceeded. Yates *et al.* (1996) report similar findings in their discussion of overconfidence - there are limits to the extent that confidence can dominate advice quality. Here, judges perceived advisors' expressions of confidence that were close to 100%, in a probabilistic forecast (dichotomous choice) task, as reckless, rather than confident. Hence, such advice is perceived as exceeding some latitude of plausibility.

A further finding in Chapter 7 of this thesis was that some participants may have revised their judgments on the basis that the advice they considered was similar to their own unaided intuitive estimates. In other words, my participants may have placed greater cognitive weight upon advice that was similar to their unaided judgments, than advice that was dissimilar. Evidence for this contention was found in Chapter 7 of this thesis where the absolute numerical difference between participants' pre-advice estimates and the advice (output-of-algorithm) was considered. A significant difference was found between the scores of participants who revised their judgment and those who did not. This finding supports the idea that people in receipt of algorithmic decompositions (US Mail and Forested Miles estimation problems) changed judgment, on the basis that the absolute numerical difference between a participant's pre-advice estimates, and the available advice, was *less* for participants who subsequently revised their judgment, than for participants who did not revise their judgment.

Judgment revision on the basis of the proximity of advice to an individual's unaided intuition may be facilitated by the same processes that underpin attitude revision. Social cognition

research has established that attitude revision is more likely where the 'reforming' attitude is less divergent from an individual's initial judgment (Kunda and Oleson, 1997). This view is further supported by the studies reported by Yaniv (2004), and Yaniv and Harries (2004), who tested this idea empirically in the context of JAS. Specifically, Yaniv (2004) examined the mediating role of the numerical 'distance' between a judge's initial unaided intuition, and advice, upon an individual's cognitive weighting policy of advice (Study 2). Here, it was hypothesized that an individual judge would weight advice that was numerically distant from advice, *less* so, than advice that was numerically close to a judge's initial unaided judgment. Analysis of the advice weighting policies of participants indicated that high knowledge participants weighted advice *less* as the distance between a judge's unaided estimate and advice increased. Contrastingly, low knowledge participants did not report any systematic pattern in their advice weighting policies. The preceding manipulation was repeated (Study 3), where advice was constituted as a constant that either 'helped' judges (acceptance and acting upon advice could improve estimation accuracy), or hindered judges (accepting and acting upon advice would be deleterious to estimation accuracy). Similar results obtained as those found in Study 2. These data can be explained by the idea that high knowledge participants may be constrained in terms of cognition and motivation to expend significant resources evaluating advice messages that are highly discrepant from their own. In constrast, low knowledge participants are constrained in evaluating advice because of limits on cognition and motivation, but also have little in the way of a pre-existing knowledge base from which to facilitate discrimination between spurious and useful advice. Here, it is perhaps no surprise that any advice is evaluated as potentially useful. In part, this may explain the findings presented in the current work, in so far as many participants chose not to revise their judgments post-advice (i.e. the numerical distance between a participant's pre-advice judgement and advice was large, on average, particularly in Chapters 6, 7, 8 and 9). Hence, participants in receipt of advice that was not dissimilar to their own judgment pre-advice, may have been

more pre-disposed to revise judgment as the demands of elaborated cognition may have been less, than where a participant would have to 'bridge' a larger cognitive 'gap' between pre-advice judgment and advice. Similarly, Yaniv and Harries (2004), report that judges are less likely to accept the recommendations of advisors whose advice is highly atypical of advice from other advisors in a multiple advisor-judge JAS polyad. The preceding discussion suggests that similarity between a judge's unaided intuition and advice may be influential in the judge's decision to revise judgment.

Determinants of judgment revision

Multiple regression analysis was utilized in this thesis to determine which variables significantly predict judgment revision (Chapters 6-8). However, most predictor variables that may have predicted judgment revision proved to be non-significant, with the exception of the perceived utility of advice. Here, this variable proved to be influential upon a participant's disposition to revise judgment, on average (Chapter 6, p242), particularly when estimating relative frequencies (Chapter 8, p293-97). Intuitively, advice that is evaluated positively is likely to be more heavily weighted in a person's deliberations over some issue than advice that is evaluated less so. However, the measure of perceived utility of advice used in the studies reported in Chapters 6-8 of this thesis could be held to be overly reductionist, as there is much research to suggest that perceived quality of advice is a complex construct (Feng and Burleson, 2008). Moreover, the processes that individuals employ to evaluate advice positively, or negatively is currently not well understood. Feng and Burleson comment that, "we still know little about the features of advice messages that lead to positive advice outcomes such as positive evaluations of advice quality and strong intentions to implement the advice" (p850). Hence, the findings reported in this thesis highlight the importance of perceptions of advice quality in determining judgment revision, whilst simultaneously pointing to the

underdeveloped nature of current understanding over how people form evaluations of advice - a potentially fruitful avenue for future research.

Direction of judgment revision

Whilst perceptions of the quality advice may be influential in an individuals tendency to revise judgment, and people who attend to advice revise judgment post-advice more often than people who do not attend advice; the issue of the 'direction' of judgment revision remains important. The findings reported in the current work suggest that people who preferred reasons-based advice tended to change judgment in the 'direction' of advice, on average - irrespective of the sub-type of advice (Chapter 5, p171-6, Chapter 6 p208-13, Chapter 7, p263). In contrast, people who preferred solely numeric advice did not significantly revise judgment in the 'direction' of advice (Chapter 4, p113-14). These findings indicate that people conform (to a degree) to reasons-based advice irrespective of its quality.

The notion that conformity to advice is an important determinant of the direction of judgment revision for individuals is not universally recognized in JAS research. Notably, Salvadori *et al.* (2001) have argued that the dynamics of a JAS interaction work *against* conformity to advice, and that, 'JAS dynamics have lower conformity pressure than other group decision dyads/polyads' (p742). This idea is antithetical to the findings presented here, as participants revised judgment such that the absolute numerical difference between a participant's pre-advice initial estimate and the available advice was greater than the absolute numerical difference between participants post-advice and the available advice. This pattern of results can be explained by the relative numerical positions of participants' pre-advice initial estimates, the numerical component of advice, and the true numerical value of the estimate.

Where advice and the true answer are numerically greater than a participant's pre-advice initial estimate, then any revision of judgment in the direction of advice or the true answer will be beneficial on average. Where a participant's numerical pre-advice initial estimate falls between advice, and the true answer, judgment revision is in the direction of advice on average. This pattern of results obtained across the experiments reported here, with the exception of the study reported in Chapter 8. Here, algorithms were provided to participants as an aid to estimate relative frequencies, and no significant effect post-advice was reported. Conformity to advice appears to be an important aspect of my participant's estimation behaviour that is robust over a number of different task domains, (where relative frequencies did not constitute the target quantity), irrespective of whether experimental tasks were operationalized in terms of a JAS, or an algorithmic decomposition of some estimation problem.

Influence of knowledge of the process of advice

One influence upon judgment, and revisions of judgment, is the idea that knowledge of the process of advice generation should allow people to evaluate advice. Should advice be evaluated positively (on the basis that the process by which the advice has been formulated is logical and coherent) then the advice *should* benefit a judge. However, the findings reported in Chapters 6-8 indicate that limited improvements in estimation accuracy are available for the tasks that my participants completed. Instead, my findings suggest that where people use algorithms to arrive at an estimate of some large quantity, their post-advice estimates are likely to be greater than their pre-advice estimates (Chapter 6, p198-200, Chapter 7, p259-60), but where algorithms are utilized to estimate relative frequencies no estimate 'inflation' is evident (Chapter 8, p285-6). These findings echo those of (Lichtenstein *et al.* 1982), in that participants in my studies tended to overestimate small quantities, and underestimate large quantities. Relying upon 'advice' in this context, by using an algorithm to relieve some of the

processing demands upon human cognition, could (and often did) result in sub-optimal post-advice estimation. Further, in terms of the extent of judgment revisions, people are not sensitive to the quality or complexity of algorithms utilized (Chapter 6, p209-10, Chapter 7, p262-64). However, there is some evidence that people are sensitive to the proximity of the value of an individual's unaided estimate to the numerical value of advice, in terms of a propensity to revise judgment (Chapter 7, p266-67). In part, MacGregor and Armstrong (1994) acknowledged the limitations of the algorithmic approach to estimation, by setting out the limited conditions under which improvements in estimation accuracy can be expected. The findings reported in this thesis reinforce the idea that algorithmic decomposition (knowledge of the process of advice generation), is only likely to be an effective cue leading to enhanced estimation accuracy, where the algorithm utilized is known to be effective *a priori,* where the quantity to be estimated is 'large' (six or more digits), and the quantity to be estimated is able to be decomposed into sub-component estimates. The responses to the sub-component estimates should also not be correlated – otherwise error will exponentially increase when the sub-component responses are combined by the algorithm to produce advice. These conditions are difficult to meet in a general sense – hence the value of algorithmic decomposition (knowledge of the process of advice) in terms of its general utility to estimators is limited.

Determinants of accuracy

Acting upon advice may not help people make more accurate estimates than people who do not attend to advice. This is because where an individual's estimate falls between advice and the true answer, people tend to revise their judgment in the direction of advice (irrespective of its quality), suggesting that people indiscriminately conform to advice (Chapter 4, p118-120). Multiple regression analysis in Chapters 6-8 sought to determine the determinants of estimation accuracy. Here, participants' performance was sensitive to the influence of initial

estimation accuracy, the accuracy of advice, and post-advice confidence (Chapter 6, p238-41, Chapter 7, p269-70, Chapter 8, p291-93). These findings do little to suggest that algorithmic decomposition is generally effective in relaxing the cognitive demands upon people when they conjure with a troublesome estimation problem. Instead, these findings highlight the domain and task specific nature of the existing findings in the extant literature, where it is proposed that decomposition, in some form, facilitates rigorous consideration of each attribute in complex estimation problems (Fischer, 1977; Kleinmuntz 1990; MacGregor, 1988, 1991, 2001). Indeed the findings reported in the current work are more aligned to the conclusions of Lichtenstein and Weathers (1998), who found that, even when people were trained to create their own algorithmic decompositions (to estimate uncertain quantities), performance was not improved to any great degree. Moreover, these authors also found that on some occasions the use of algorithmic decomposition could inhibit accurate estimation (Study 2). In short, knowledge of the process of advice generation was not a sufficient cue to participants in the studies reported here, to facilitate significant improvements in estimation accuracy. The findings reported in this thesis instead point to the robust influence of the conformity to advice account. Where estimators' unaided intuitive estimates were more accurate than advice from an advisor, any revision of judgment in the direction of advice will be deleterious in terms of ultimate estimation accuracy. For accuracy to be beneficial for estimators, it clearly must be more accurate than the level of performance that estimators are able to achieve unaided. This account does not support the idea that even inaccurate advice is spur to enhanced 'cognitive activation' (Brehmer and Hagafors, 1986) that may be a route to improved ultimate estimation accuracy.

People are constrained in their abilities to assess and use advice in their deliberations even when advice is objectively 'good'.

In the absence of feedback that enables people to beneficially revise their intuitions of some estimate, or quantity, people are constrained and limited in their abilities to separate out advice that is potentially beneficial (if acted upon), and spurious advice. Instead, the studies reported here indicate that people may be inclined to act upon any advice irrespective of its quality - if they choose to revise their own intuitions in a 'one-shot' advice giving exchange. This idea is evidenced by the observation that people are not able to easily tell the difference between 'good' and 'bad' advice (Chapter 4, p120, Chapter 5, p171-72, Chapter 6, p206-8, Chapter 7, p264-65, Chapter 8, p288, Chapter 9, p315-16, p321).

However, research by Soll and Larick (2009), suggests that at least in part, people are able to mitigate the worst effects of constraints upon advice evaluation by employing a repertoire of weighted averaging and choosing strategies. For these authors, the gap between a normative benchmark of accuracy and sub-optimal estimation accuracy following advice, can be explained by people's inability to discern when to apply a weighted averaging strategy, or a choosing strategy in response to advice offered by advisors. The data in this thesis does not address this issue directly, but does show that people are insensitive to the quality of advice across various domains and tasks. Soll and Larick show experimentally that an averaging strategy should dominate a choosing strategy in a variety of environmental contexts. Indeed, Herzog and Hertwig (2009) concur that an averaging strategy over the judgments of a judge and advisor is desirable in terms of extracting the 'good' portion of advice from a 'one-shot' advice giving scenario. These authors go further however, by arguing that an averaging strategy is also possible for individuals - providing that an individual forms at least two judgments about some estimation issue, and that each judgment is based on a different source

of information. Under this account errors and biases largely cancel each other out, leading to improved estimation accuracy.

Chapter 9 in this thesis however, tasked people with estimating the IMRs of ten countries. Here, advice was constituted as GDP advice (which was predictive of the target quantity), Population advice (which was not predictive of the target IMR) and GDP and Population advice. Both Population estimate advice and GDP advice were accurate, and taken from the CIA Factbook. Hence neither, averaging or choosing was an optimal strategy for estimating IMRs. Under these conditions where people revise judgment post-advice, the influence of GDP advice is likely to be beneficial in terms of accurately estimating IMRs, whilst Population estimate advice was not. Consistent with the results of reported in Chapters 4-8, people in receipt of advice (irrespective of quality) revised judgment more often than people not in receipt of advice (Chapter 9, p315-16), and overestimated IMRs pre-advice, on average (p312) (Lichtenstein *et al.* 1982).

Where the results of the study reported in Chapter 9 differed markedly from the previous studies (Chapters 4-8), was the analysis of participants' ability to rank order countries by IMRs – ordinal accuracy. Here, results suggested that people *are* successful in improving accuracy when tasked with rank ordering countries by IMRs (ordinal accuracy), and in receipt of a single advice type (irrespective of the quality of advice). This was somewhat surprising given that Population estimate advice was not predictive of the target quantity. It was suggested that this might be explained by a failure to successfully randomly allocate participants between experimental conditions – participants in the Population estimate advice group were older than those in the GDP advice group, on average, and may have possessed greater world knowledge. Moreover, ordinal accuracy does not improve where an individual is in receipt of both advice

types (Chapter 9, p322-5). This may suggest that due to cognitive limitations, people were unable to assess and use the potentially beneficial GDP component of advice to improve ordinal accuracy. This position is further supported in so far as metric accuracy for participants in the GDP and Population estimate advice group, did not improve post-advice (Chapter 9, p323). These findings do not fully support the idea that people are able to recognize predictive information in an estimation task, and recall cues that enable individuals to rank order target items (ordinal accuracy). (Gigerenzer, *et al.* 1999), because where advice was constituted as a combination of predictive and non-predictive information (GDP and Population estimate advice), people were unable to separate out the predictive (GDP element) of advice to improve metric estimation accuracy. Instead, there is limited support for the idea that by conforming to advice that *is* predictive of the target quantity (solely GDP advice), people are able to improve estimation accuracy. This account was leant further support when the absolute amount of judgment change data was examined - there were no significant differences, on average, between advice types (Chapter 9, p326). These findings strongly suggest that people are unable to adequately differentiate between potentially beneficial advice, and poor advice in a 'one-shot' advice exchange.

10.4 Implication of the findings from the current work for the real world

The preceding discussion leads to consideration of what implications these findings might have in an applied sense. Perhaps the most important theme here is that human judgment is influenced by a number of factors, and should not be examined outside of its social context. The preceding discussion has also shown that the psychological construct of advice is multi-faceted, and that people are not well equipped to differentiate between potentially beneficial and poor advice across a range of task domains (in a single advice exchange episode). However, the findings reported in this thesis can make some contributions to how individuals

can improve their judgments, and also show how some organizational decisions and judgments (employment decisions, mentoring and socialization) can be viewed as analogous to the judge-advisor informational exchange.

Advice for judges

In order to fully utilize the cognitive capacity for judgment, judges should avoid (where possible) single judge-advisor information exchanges (particular where a judgment concerns a non-routine, or novel decision). This is because people are ill equipped to differentiate between potentially beneficial and poor advice under these circumstances, and often seek to lighten the demands of cognitive processing by over-weighting peripheral cues - such as the credibility of the advisor(s). Second, if circumstances dictate that a single judge-advisor information exchange is necessary judges should aggregate their own judgments with those of the advisor (where a clear solution to the problem of judgment is not evident). Judges should discount unelaborated advice (i.e. solely numeric advice), in favour of (at least) statistical, or elaborated reasons-based advice. If a judge finds it necessary to enter a single judge-advisor interaction, where a clear solution to a problem of judgment is not evident, and unelaborated advice is not available, but numeric advice is available, judges should rank order multiple pieces of advice by some target criterion variable. Where advice is constituted in solely numeric terms, judges should revert to a simple averaging strategy, as this likely to be the least deleterious course of action, on average.

Advice for Advisors

Advisors who wish to influence judges should provide advice that is elaborated and reasons-based, low in quantification, with information evidenced by a credible source. Advisors may also wish to 'pitch' their advice close to the unaided intuitive judgments of the judge

(proximity between a judge's judgment and that of an advisor may be more likely to result in revisions of judgment on the part of the judge). Algorithmic decompositions of estimation problems may only be persuasive under limited and defined circumstances (MacGregor and Armstrong, 1994). Here, advisors should not provide algorithms for estimates of extremely small quantities, or quantities expressed as relative frequencies. The following sections focus upon some implications for the organizational sciences based upon the findings of the current work.

Employment decisions

The extant literature of the organizational sciences has typically viewed employment decisions from the perspective of the organization, yet the results presented in this thesis potentially speak to employment decisions from the perspective of the employee. This is because as Slaugher and Highhouse (2003, p12) argue that, "choices among job alternatives are almost never made in isolation. Individuals choosing among jobs are likely to consult those with whom they have social contact, such as friends, and those individuals for whom the decision will have indirect yet important consequences, such as family members". This position invites many issues that this thesis does not address, yet the results presented here do suggest that employment decisions maybe influenced by elaborated, or reasons-based advice, and that less confident potential employees may seek out more job related advice than more confident potential employees. Furthermore, potential employees who face a complex employment decision task, maybe susceptible to seeking more advice and being inappropriately influenced by peripheral cues within that advice, in order to reduce the complexity of the decision. A particularly salient scenario might be an employee's decision to leave an organization. Here, advice from a variety of sources is likely to be sought, including co-workers, family and

friends. Inappropriately, weighting advice in this context, or inappropriately discounting advice, could prove significantly costly to an individual.

Mentoring

It is easy to recognise the outlines of a prototypical JAS dyad (or polyad) when considering the role that mentors or protégés assume in an organizational context. Here, protégés can be seen in the role of a judge, and mentors can be viewed as an advisor. The psychological or emotional content that potentially characterizes a mentor-protégé advice exchange is not within the parameters of the present work (but perhaps *should* be considered in conceptualizing the psychological construct of advice). However, the advisory exchange (concerning strategies for achieving work related objectives in an organizational context), can encompass recommendations for particular courses of action. Given the social nature of the relationship between mentor and protégé, it perhaps should not be expected that protégés should, or would, egocentrically discount advice to any great extent. However, the findings presented in this thesis suggest that the propensity of a judge to accept and act upon the advice of an advisor is likely to be mediated by the perceived confidence of the advisor, or the similarity between the judge's unaided judgment and the advice proffered by the advisor.

Socialization

Although currently there is no general model (either prescriptive or descriptive) for how new recruits become socialized within an organization, it is possible to discern how new recruits might seek out advice from members of an organization - both formally and informally. Dalal and Bonaccio (2006) make the observation that such an information search is akin to judges seeking advice from advisor(s). Further, JAS research might be the lens through which the degree and extent of judgment revisions on the part of new recruits to an organization could be examined (when presented with advice from both confident and less confident advisors – who

provide elaborated or unelaborated judgments/advice). Further, judgment revision could be influenced by the extent a new recruit possesses previous work related, or domain related, experience and is hence resistant to the influence of advice.

10.5 Caveats, limitations and some perils of conducting research in the area of advice giving and taking.

This section will discuss certain caveats to the findings of the current research, before considering some of the methodological pitfalls that researchers may face when undertaking similar research in the future.

The findings reported in this thesis should be accepted with a number of caveats in mind. Firstly, the studies reported here address the situation of a single judge-advisor informational exchange. Here, a strong theme that emerges is that people are constrained in their abilities to differentiate between potentially beneficial advice and poor advice. In part this can be attributed to the limitations of human cognition, and it could be easy to conclude that this might be the central message from this research. However, it is clear from the extant literature that people *can* differentiate between potentially beneficial and poor advice where advice is sought and considered from advisor(s) *repeatedly* (Yaniv and Kleinberger, 2000). Here, judges weight the advice from accurate advisors more so than advice from less accurate advisors. Further, evidence emerged in Chapter 9 of this thesis that the constraints upon people to differentiate between potentially beneficial and poor advice, may not be absolute. People retain some ability to rank order target quantities by some criterion variable even where the judge-advisor exchange is limited to a single episode. Hence, the results reported in this thesis are limited in so far as they address only a single judge-advisor exchange of information.

A second caveat concerns the operationalization of the advice construct used in the current work. A major limitation in the extant JAS literature is that advice has been operationalized by experimental considerations, rather than in an ecological sense. Defining advice in solely numeric terms (which is common in JAS research) is clearly unsatisfactory because this does not fully acknowledge the social interaction implicit in real world advice giving and taking scenarios (Heath and Gonzales, 1995; Gibbons, 2003). There are many elements of advice not captured by solely numeric advice - the provision of social support to a judge, or the endorsement of a judge's decision in addition to the provision of information or rationales for alternatives not originally considered by a judge. A step towards a broader conceptualization of the nature of advice, and operationalizations for experimental purposes, has been attempted in this thesis where advice has been constituted as reasons-based (in argumentative form, in variance in the quality of supporting rationales, and in the provision of process information). However, the current work can only be accepted if the limitations of the operationalizations of advice presented here are acknowledged.

In a similar vein, it is necessary to acknowledge the possible limitations to the use of solely numeric advice as a meaningful form of advice. Arguably, such an advice form lacks ecological validity, and is perhaps not likely to be offered or accepted without additional contextual information - given the prevailing social mores in the UK. Further, solely numeric advice may prime people to attend to peripheral environmental cues. This is because people are predisposed to conserve cognitive resources, instead of expending cognitive resources centrally processing the message content of advice high in quantification. Statistical information constituted as advice may be a more meaningful way to convey a numeric advice message. This is because statistical information may suggest associations about the population from which the statistic is drawn beyond the actual figures conveyed.

Further caveats need to be acknowledged in terms of the operationalization of reason-based advice in the current work. All the reasons-based advice was formulated by the author and none was provided by any advisor(s). Clearly, this deception raises the issue over whether participants actually believed that the reasons-based advice in Chapter 4 and 5, emanated from real advisors, and whether the methods proposed by advisors in Chapters 6, 7 and 8 for solutions to the estimation problems were actual solutions from real people, or merely the product of the author. Whilst there appears to be little evidence to favour one alternative over another, missing data might be indicative of the idea that participants were not totally convinced of the 'cover story'. These considerations raise further questions of the appropriateness of the use of almanac 'quiz' type tasks, and laboratory experiments generally. Clearly, tasks such as these, under typically artificial circumstances may not produce data that can meaningfully address potential influences upon human judgment. This is a potentially serious issue for both the current work, and for research in the area of advice giving and advice taking generally.

There are numerous potential pitfalls researchers may face when attempting empirical work in the field of human judgments. Discussion of some of the methodological shortcomings evident in the current work may serve as a guide to future researchers about pitfalls they may wish to avoid. Perhaps the most obvious issue that is apparent from a reading of Chapter 4 in this thesis concerns the issue of counterbalancing the experimental stimuli. Here, the numerical value of advice (both reasons-based and solely numerical) was not adequately counterbalanced. This meant that reasons-based advice was less error prone (35 years, on average), than solely numerical advice (48 years, on average) (see Table 4.1). Although this counterbalancing scheme does not appear to have significantly biased the results of the study, the manipulation could have been better formulated so that the aggregate numerical component of each advice type matched. Further, that the numerical component of each piece of reasons-

based advice was identical to a piece of numeric advice. In a wider sense this issue highlights the pitfalls researchers may face when constructing experimental materials. This issue is significant as a participant's propensity to accept and act on advice is mediated by the nature of the estimation task. One example of the mediating influence of the experimental task, upon a person's propensity to attend to and act upon advice, is the perceived complexity of the experimental task. The idea that task complexity may mediate judgment revision should be further scrutinized. This is because 'complexity' has been defined in several different ways. Gino and Moore (2007) for example, defined 'complexity' on the basis of clear or blurred photographs utilized for the purpose of experimental stimuli. Clearly, the construct of 'complexity' can be (and perhaps should be) defined in a multiplicity of ways, and the effects upon judgment revision tested empirically.

A further idea explored in Studies 6-8 in this thesis was that knowledge of the process of advice generation (algorithmic decompositions of estimation problems), *should* be a sufficient cue to allow participants to evaluate advice. However, it is clear from the results reported here that knowledge of the process of advice is not an effective cue for distinguishing quality of advice, when advice is constituted as an algorithmic decomposition of known (and unknown) effectiveness. However, this does not necessarily imply that knowledge of the process of advice is *never* a successful environmental cue to the quality of advice. The results from the studies reported in Chapters 6-8 are of limited generalizability – because the population from which participants were drawn consisted of mostly students from a specific socio-economic background, tasked with completing novel and unfamiliar estimation problems, for which they may not have been adequately motivated. Clearly, this does not imply all estimation aids based upon syllogistic logic, or Socratic intervention, are by implication as limited as the algorithmic approach utilized in these studies. On the contrary, commercial airline pilots are trained to rely upon instructional algorithms (that are high in ecological validity) to control

aircraft, and avert potential dangers to safe travel. Here, knowledge of how the instructional algorithms for controlling aircraft were produced is clearly important to the acceptance of the advice that such algorithms provide to aircrew (Scriven, 2000). This observation implies that knowledge of the process of advice generation is an important issue that requires further investigation (i.e. the formulation and manipulation of algorithms that are ecologically valid, and meaningful to participants), in order that the effects can observed in terms of judgment revision (or not). Researchers may wish in future to pre-test the effectiveness of a repertoire of algorithms, and match these algorithms to different task domains prior to designing studies.

A related issue to the development of estimation tasks for experimentation, concerns validity. A major pitfall for researchers is that laboratory based estimation problems (usually of almanac type questions), may not simulate either the task domains, or the conditions, under which people make estimates of quantities in their everyday experience. This has significant implications for understanding the cognitive processes that facilitate decisions and estimates. It is parsimonious to suppose that in everyday life many judge-advisor(s) exchanges are loosely structured, informal, often (but not always) face-to-face, and based upon a history of trust between a judge and his/her advisors. Capturing this reality within the confines of an experiment is likely to be problematic, yet it should be possible for researchers to at least develop experimental materials that are more centrally situated within the everyday experience of participants – to a far greater extent than has been accomplished in this thesis, or in the extant literature.

The ecological validity of experimental materials must also address the issue of participant motivation in future studies. Often this issue has been addressed by some nominal financial payment for participation (Budescu and Rantilla, 2000; Yaniv, 2004), rather than performance,

in the extant advice giving and advice taking literature. Clearly, the motivation of participants in the studies reported in the current work may have been variable, given the number of missing data entry values in some cells of the experiments. This issue could possibly be mitigated by the formulation of more valid experimental tasks which may intrinsically motivate participants, and by linking payments only to levels of performance (extrinsic motivation).

The vast majority of participants in the studies reported in this thesis were students, and hence it is difficult to make the argument that findings generalize to populations other than students. Whilst this is a perennial debate in social science research, researchers in the field of advice giving and advice taking should at least consider careful representative sampling of target populations. For reasons of practicality, sampling issues were dominated by concerns about accessibility to participants in this thesis.

Moreover, the use of IT, and online surveys should be carefully evaluated during the design of studies. Two of the studies in this thesis were developed so that participants could respond online. Future researchers should approach online methods of data capture with caution, as some knowledge of IT hardware (the servers on which the survey is to be hosted, the programming language in which the survey will be written, and the means by which data will be stored and retrieved) is essential for success in this area. The author naively assumed that these issues were easily surmountable – when they were not. Hence, some of the missing data reported in each study could be attributable to IT issues such as validation of data entries. Further, the author's lack of IT skills may have not allowed a full exploration of the capabilities of the MouseLabWeb software used in Chapter 4. Here, the experimental design

could have been extended in ways that made more meaningful use of process tracing methods to investigate judge-advisor interactions.

A further issue that is clearly important, yet not fully acknowledged in the extant JAS literature concerns the measures that researchers have used to analyze judgment revision. Many studies of advice utilization in JAS research have used measures based upon absolute difference scores, where the judge's final decision is the combination of an initial intuition mediated by some form of advice from an advisor. Harvey and Fischer (1997), for example defined advice taking as a ratio of two differences – the difference between a judge's post-advice estimate and the same judge's pre-advice estimate, divided by the difference between an advisors recommendation and the judge's pre-advice estimate. I used a similar measure in Chapter 5 – WOA (Yaniv, 2004). Here, WOA can be defined as $|f - i| / |a - i|$, where *i, f,* and *a,* represent initial, final and advice respectively. Further, it is argued that the weight of advice is well defined if the final estimate falls between the initial estimate and the advice. Expressed as a proportion, WOA reflects the weight that a participant places upon advice (it is inversely related to the extent that the advice is discounted). Hence, advice weighted as '0' indicates that a participant has completely discounted the advice and is fully convinced of the veracity of their own judgment, whereas a WOA of '1' indicates that a participant has jettisoned their own initial judgment and entirely adopted the advice available to them in determining a final estimate. A WOA of 0.5 would indicate that both the advice and a participant's initial judgment were weighted equally. Similarly, Yaniv and Kleinberger (2000), defined a further measure – weight of own estimate (WOE). Here, WOE can be defined as $|a - f| / |a - i|$, where *i, f,* and *a,* represent initial, final and advice respectively. There are a number of issues of concern to JAS researchers, that are implied by the use of measures of absolute difference scores discussed below.

Perhaps an obvious problem for researchers when analyzing data derived from the measures of absolute difference, outlined in the preceding paragraph, is that where advice is the same as a judge's initial estimate, the denominator in each of the ratios is zero. However, it may not be the case that zero indicates absolutely no consideration of advice, and hence the product of the measure is ill-defined. A similar problem arises in the case of both Harvey's (1997) advice taking measure, and Yaniv's (2004), WOA, if a judge's final estimate is no different from the same judge's initial estimate. Here, the product of the measure – zero, may not be indicative of absolutely no consideration of advice. In these circumstances it might be that the advice served only to confirm the judge in the veracity of his/her initial intuition. Further, the WOA measure does not adequately differentiate situations where a judge's final estimate approaches advice, from situations where a judge's final estimate moves away from advice. Possibly, such a scenario may be captured by measures of confidence - which may increase as a result. Where a judge's final estimate is identical to advice when the WOE is considered, this indicates that no weight is attached to the judge's initial intuition. Further, this measure does not distinguish situations where a judge's final estimate approaches the numerical value of advice, or 'overshoots' the numerical value of advice. Both WOA and WOE are bounded at the lower extremity by zero, yet have no upper bound – 'overshoots' are typically characterized by values greater than 1.00 for WOA, or moves away from advice (in numerical terms) for WOE. Lastly, these measures fail to treat relative changes differently. A judge who revises his/her initial intuition from 50 to 75 on receipt of advice of 100, is no different from a different judge who revises their initial intuition from 20 to 40 on receipt of advice of 100. Both judges are measured as revising their judgment by 50%. This is problematic in so far as the study reported in Chapter 6 found that the propensity of a judge to revise their judgment is linked to the proximity of a judge's initial intuition to advice. One alternative to the problems associated with measures based upon absolute difference scores is approaches based upon regression, which I now turn to.

Regression based modelling was utilized in the studies in Chapters 6, 7 and 8 in this thesis, and is an approach that could be further explored by JAS researchers. Essentially, the percentage variance in an endogenous criterion variable is predicted by multiple exogenous variables. In other words, a judge's final deliberation is regressed upon multiple sources of information simultaneously. Further, where judges' face several decision problems, future studies could contemplate the computation of advice utilization indices *within* judges, rather than *between* judges which is implied by the use of measures based upon absolute difference scores. Hence, there is a strong argument for the development of regression based approaches in future JAS studies. One further issue of measurement is considered next, that of ordinal accuracy.

The study reported in Chapter 9 raised the possibility that ordinal (i.e. rank order of target items) accuracy, and metric (i.e. point estimate) accuracy, may be dependent on independent cognitive processes. Whilst the results of this study did not fully support this proposition, JAS experimental studies have not, to date, examined how judges utilize advice to facilitate ordinal estimates. This appears to be a potentially fruitful line of enquiry that, in conjunction with some or all, of the recommendations made here could produce insights into the cognitive architecture that makes judgment revision possible.

10.6 Future directions

Perhaps the main difficulty in setting out future directions for research can be attributed to the current lack of a unifying comprehensive theoretical framework for studies in the area of advice giving and advice taking. This is unfortunate, as comparisons between studies, and across studies, are problematic due to the hetereogeneity of approaches, methods, and levels of analysis reported by researchers. What has been attempted in the current work is to examine

influences upon judgments where advice is provided in a single judge-advisor exchange. Whilst this may appear limiting, people often base their judgments on such an interaction (e.g. when accessing advice from government bodies, medical and legal professionals). Clearly, the preceding scenarios are not comprehensive in terms of any taxonomy of judge-advisor interactions, nor do they preclude instances of repeated judge-advisor interactions where judgments may be revised and updated in the light of new information. However, it is parsimonious to suppose that the single judge-advisor episode is not uncommon and hence deserves serious consideration. Future research could usefully further examine the circumstances in which beneficial judgment revision is the outcome of such an exchange. A significant first step might be to address the issue of what constitutes *advice*, and then move on to discuss operationalizations of advice for experimental purposes.

Investigations under a wider rubric of advice will no doubt operationalize advice as a multi-dimensional construct, where 'advice' is analyzed in terms of its many component factors (Dalal and Bonaccio, 2006). The appeal for research of this nature was motivated by the work of Cross, Borgatti and Parker (2001), who identified five types of advice ordered along a single dimension: solutions, meta-knowledge, problem reformulation, validation, and legitimization. These authors argue that an advisor was likely to provide advice characterized by solutions, followed by the remaining four advice types, when advice was sought in an organizational or workplace context. Such a conceptualization of advice is somewhat removed from the current operationalization of advice in JAS research. Hence, future JAS studies should acknowledge advice as a multi-dimensional construct, and operationalize measures for the purpose of experimentation. A useful example of the utility of a re-conceptualization of advice in a judge-advisor interaction is provided by Dalal and Bonaccio (2006), who drawing on the work of Goldsmith (2000), argue that advice may overcome a judge's reticence to revise an unaided judgment, if it was to follow emotional support (p47). Hence, a re-conceptualization and

operationalization of exactly what constitutes 'advice'; in the context of JAS research is a necessary precursor to a greater understanding of the cognitive architecture which facilitates beneficial judgment revision.

A second issue for future research concerns the formulation of ecologically valid experimental tasks, and the recruitment of participants whose experience encompasses familiarity with such tasks. The author was aware of some of these issues prior to embarking on the current project, but was constrained in terms of time and resources in my ability to fully address these issues in this thesis. Unfortunately, this leaves many potential avenues of interest unexplored. The ecological validity of the experimental tasks participants carried out in the current work is seriously limiting. These tasks were novel and unfamiliar to participants and clearly do not reveal the full extent of influences upon judgment revision. Future research could beneficially focus upon experimental tasks that more fully simulate real world activities that people commonly engage in. This may be achievable by collaborations with 'real world' organizations, where researchers could possibly conduct experiments disguised as work related activities. This possibility could also simultaneously limit experimental demand characteristics, achieve acceptable levels of participant motivation, and more realistically simulate judge-advisor interactions that may provide both theoretical and applied knowledge about the cognitive processes that underpin advice solicitation, evaluation of advice, and combining advice with a judge's unaided intuition.

A related issue concerns the population from which participants for the current work were drawn. The majority of participants were undergraduate students enrolled at the University of Durham (UK). Hence, the findings from the current work are limited in terms of generalizing to other populations (in terms of age, experience, social background etc.). Engaging with real

world organizations (as previously discussed) offers one way to relax some of the constraints upon the generalizability of findings. Organizations offer researchers potential populations of participants with a wider range of characteristics (in terms of age, experience, motivation and social background), and is currently an underdeveloped area of research from the perspective of single judge-advisor exchange episodes – which would be an appropriate area for future research.

Participants in the current work were randomly allocated between experimental conditions (although there may have been a problem with this experimental control procedure in Chapter 9). However, no attempt was made to match participants on important characteristics between conditions – no measure of task relevant knowledge was taken prior to the experimental procedure, and hence individual differences between important characteristics of participants may not have been normally distributed across experimental conditions. This is an important issue that should be incorporated into future experimental designs.

A range of experimental design issues were encountered by the author in the current work, and perhaps the most obvious concerns control. The counterbalancing scheme involving the accuracy of the numerical component of advice in Chapter 4 was not well thought out, and future research should ensure that stricter experimental control procedures are adhered to. Furthermore, a tension appears in Chapter 9 between the author's attempt to ensure random allocation of participants between experimental conditions and the overall balance of numbers of participants in each experimental condition. Ultimately, this resulted in some cells in the subsequent analysis having such low *N*'s that it was impossible to have confidence that any significant differences that were found by statistical testing were real, or merely random. Clearly, the issues of random allocation of participants and an approximate balance of numbers

of participants in each cell of the analysis should be seriously considered during the data collection stage of subsequent research.

A further important direction for future research is to establish the cognitive processes that allow people to make *evaluations* of advice. This thesis has established that people prefer elaborated rationales as an aid that relaxes some of the cognitive demands upon centrally processing advice messages, yet it has not yet been established what attributes of elaborated, or reasons-based advice are most significant in this respect. Clearly, individual differences between people in terms of the degree of knowledge that can be mustered to address estimation problems, in conjunction with the motivation and confidence of individuals to estimate accurately, coupled with the similarity of advice to an individual's unaided intuition may interact to facilitate the process of evaluation. However, in what proportion and in what configuration remains an empirical question.

Moreover, an important and under-researched issue is the influence of a judge upon an advisor's propensity to provide appropriate advice (Jonas and Frey, 2003). One of the few extant studies to consider this issue conducted by Gibbons *et al.* (2003) examined the influence of a judge's unaided intuition upon an advisor's subsequent recommendation and confidence. Here, advisors were able to offer advice to a judge where the judge's initial unaided intuition conflicted with the judgment of a judge. In these circumstances, advisors were more disposed to offer advice. Here, research is necessary to determine the mediating variables of the confidence of a judge (in terms of self-efficacy), the confidence of the judge in the abilities of the advisor(s) to provide useful advice, the influence upon the advisor of the elaborated nature of information flows in committees, and the influence of organizational hierarchy upon both the propensity, and content of advice advisor(s) may offer.

Ultimately, future research should acknowledge the influence of advice timing upon judgment revision post-advice. Some JAS studies have precluded judges from forming initial intuitions in advance of an experimental task and associated advice (Budescu and Rantilla, 2000; Harvey *et al.* 2000, Sniezek and Van Swol, 2001). Participants in these studies were prevented from accessing their own autonomous cognitions prior to receiving advice, and hence are 'cued' towards the recommendation of the advisor. One result of 'cueing' is that participants may differentially process information, spending less cognitive resources considering 'uncued' information, and generally becoming overconfident in the value of the information that *is* processed (Sniezek and Buckley, 1995). A further consequence of preventing judges forming a judgment prior to the advent of advice is that judges may be unaware of the influence of advice, upon the formation of their 'own' judgment (Koehler and Beauregard, 2006). These authors suggest that judges are either unaware of the influence of advice, when forming their 'own' judgments; or unable to separate advice from their own cognitions and deliberations when making judgments (Koehler and Beauregard, 2006). Clearly, the interaction between the temporal presentation of advice and other variables is likely to influence the outcome of judgment.

In sum, the preceding experimental studies have framed, explored, and operationalized, significant questions of interest of concern to JAS researchers. Further, recommendations for methodological improvements have been made in the design, operationalization, and execution of future JAS studies. Ultimately, this contribution is consistent with the appeal of theorists such as Payne *et al.*(1993) in so far as, 'the social context of decisions has been a neglected part of decision research . . . and is a area worthy of much greater study' (p255).

Appendix 1

MouseLabWeb interface screens

Screen 1

Start of experiment

Subject name:	

Screen 2

Introduction and Instructions

In this questionnaire you will be asked to estimate the date of notable events that have occurred within the last 300 years. Below, we ask you to make estimates to three questions about which you may have very little knowledge. Even though you may find these questions difficult, we ask that you do your best to make the most accurate assessment that you can. Please answer all the questions that follow, and avoid the use of the backspace key in empty text boxes.

Next Page

Screen 3

Below, we ask you to make an estimate to a question concerning a specific historical event that has occurred in the last 300 years (1700 - 2006) - about which you may have very little knowledge. Even though you may find the question difficult, we ask that you do your best to make the most accurate assessment that you can.

In what year was the first Japanese railway opened?

Put your answer here

Since you may not know the exact answer to the question, we now move on to assess how confident that you are in the answer of you have just given.

We ask that you put both a lower year and a higher year, such that you are fairly sure that the true year of the opening of the first Japanese railway lies between these lower, and higher bounds.

Please provide a lower bound, and a higher bound between which you are 95% certain that the true answer – whatever that might be – lies.

Lower bound . . . . . .true answer . . . . . Upper bound

. . . . . .

Next Page

Screen 4

Often when people make estimates they can think of a reason for their estimate. The reasons that people provide for their estimates can be placed in to four separate categories. We illustrate the four types of reasons below.

Analogous reason – here the reason a person gives makes use of our general knowledge of relationships between two events in dissimilar situations. For example, if someone is trying to estimate the time it will take to drive to a nearby airport, they may reason that, "the airport is roughly the same distance away as the shopping mall. Therefore, the time it will take to get to the airport will be approximately the same as it is to travel to the shopping mall – about 30 minutes".

Parallel Case reason – here the reason a person gives makes use of their knowledge of a previous experience of a near identical situation. For example, if someone is trying to estimate the time it will take to drive to a nearby airport they may reason that, "it will take about 30 minutes to drive to the airport because it took 30 minutes at the same time of day last month".

Authoritative reason – here the reason a person gives makes use of expert knowledge. For example, "the radio announcer has said that traffic to the airport is heavy today and so I estimate that you should add 20 minutes to our journey time".

Motivational reason – here the reason a person gives makes use of specific insights about people's motivations or desires. For example, "since you will be in a hurry then I reckon that you will cut 5 minutes off our journey time".

In the boxes below are eight responses by other people to the question of which year the first Japanese railway opened. The correct answer may or may not be included. You are free, if you wish to look at the answers given by the eight others. Four of the answers contain simple numerical estimates, and four of the answers contain both a numerical estimate and the person's reasoning behind the estimate. The reasons match the four types of reasons that we have described above. You are free to look at either all eight estimates, a few of the estimates, or none of the estimates at all.

You can reveal an estimate by a simple mouse click on the appropriate box. After you have looked at as many estimates as you wish, then you will have an opportunity to change the estimate that you made a few minutes ago.

Recall that your estimate of the year that the first Japanese railway opened was - ****

Recall that you were 95% certain that the true answer (whatever that might be) was a number between ** and ******

The first railway was opened in 1890, in the same way that compulsory elementary education was introduced in Japan – by central govt. Person E: This box contains a number & an analogy	1861 Person A: This box contains a number
The construction of the first Japanese railway in 1902 was motivated by the desire to adopt European technology. Person H: This box contains a number & a motivation	1895 Person D: This box contains a number
According to the historian Dr. John Edwards, the first Japanese railway was opened in 1835. Person G: This box contains a number & an authority	1883 Person C: This box contains a number
The first Japanese railway opened in 1869 – the same year that the trans-continental railroad was completed in the USA. Person F: This box contains a number & a parallel case	1849 Person B: This box contains a number

Recall that your estimate of the year that the first Japanese railway opened was - ****

Recall that you were 95% certain that the true answer (whatever that might be) was a number between ** and ******

Which person's answer did you find most helpful?

- (•) None
- ( ) Person A
- ( ) Person B
- ( ) Person C
- ( ) Person D
- ( ) Person E
- ( ) Person F
- ( ) Person G
- ( ) Person H

In light of the above responses, do you wish to amend your previous answer?

- ( ) No
- ( ) Yes

If you have answered 'YES' to the above question, you MUST put your revised estimate here

[ ]

Next Page

Screen 5

Recall that your estimate of the year that the first Japanese railway opened was - **

Recall that you were 95% certain that the true answer (whatever that may be) was a number between ** and ****

Recall that your new best estimate of the year that the first Japanese railway opened was -

Recall that the advice that you found useful was -

None of the advice was useful to you

Please tell us the reasons why you changed your judgment in the box below.

If you did not change your judgment, please tell us the reasons why you did not change judgment in the box below.

Next Page

Screen 6

Recall that your estimate of the year that the first Japanese railway opened was - **

Recall that you were 95% certain that the true answer (whatever that may be) was a number between ** and ****

Recall that your revised estimate of the year that the first Japanese Railway opened was -

Recall that the advice that you found useful was -

None of the advice was useful to you

How confident are you that your revised estimate is accurate? If you have not revised your estimate, then how confident are you that your unrevised estimate is accurate?

Please provide a lower bound, and a higher bound between which you are 95 % certain that the true answer – whatever that may be – lies.

Lower bound . . . . . .true answer . . . . . .Higher bound

[ ] . . . . . [ ]

Next Page

Thanks!

Thanks for joining this experiment.
You can close the browser window.

Appendix II

Appendix II: 14 questionnaire items and correct dates

1. In what year was the telephone invented by Alexander Bell? (1876)
2. In what year did India gain independence from Britain? (1947)
3. In what year did Henry Ford first produce the Model T car? (1908)
4. In what year did Mao Zedung and Zhu De lead communists on the 'Long March' through China? (1934)
5. In what year were diamonds first discovered at Kimberley in South Africa? (1867)
6. In what year was the first general election held in Japan? (1890)
7. In what year did Egypt nationalise the Suez Canal? (1956)
8. In what year was Rev. Martin Luther King assassinated? (1968)
9. In what year was the 'Boxer Rebellion' in China? (1900)
10. In what year was the first Camp David summit between Egypt and Israel hosted by the United States? (1978)
11. In what year did Thomas Eddison invent the light bulb? (1883)
12. In what year did Britain receive a mandate to govern Palestine? (1920)
13. In what year did Thomas Jefferson become president of the United States? (1801)
14. In what year was the first Japanese railway opened? (1872)

Appendix II: Advice available to participants

	Correct date & accurate advice	Inaccurate numerical advice	Very weak reasons	Weak reasons	Strong reasons
In what year was the telephone invented by Alexander Bell?	1876	1895	The telephone was invented sometime before WWI, probably the Victorian age.	Bell's invention of the telephone developed out of his Study to improve the telegraph system, so it must have been late nineteenth century.	I was at a pub quiz last week and this was one of the questions that no one got right — that's why the date stuck in my mind.
In what year did India gain independence from Britain?	1947	1955	It was definitely shortly after WW II.	Indian independence was negotiated by Lord Mountbatten — the last British Viceroy — and the various political parties in India following WW II. This may have taken some years to finalise.	My grandfather returned from India shortly before the granting of Independence.
In what year did Henry Ford first produce the Model T car?	1908	1915	I don't know if the Model T was the first car to be produced, but it must have been amongst the first.	Ford probably wasn't the first car manufacturer in the automotive business. Ford made cars that ordinary people could afford, so it must have been at the start of mass production around the time of WW I in Europe	I'm a bit of a Ford enthusiast — unfashionable as it is nowadays. Purely out of curiosity I did a few internet searches to find out about the company and its history. That's how I know when the Model T was first produced.

In what year did Mao Zedung and Zhu De lead communists on the 'Long March' through China?	1934	1925	China only became fully communist after WW II, but Mao's Long March was before that.	The Chinese communists didn't seize power until after WWII, Japan invaded China in the late 1930's so the Long March must have been during a civil war period some time before the Japanese invasion.	The origins of communist China are part of my undergraduate studies. That's how I know when the 'Long March' occurred.
In what year were diamonds first discovered at Kimberley in South Africa?	1867	1895	Intuitively, the South African diamond industry feels quite old, probably during Victorian times. So I'm sort of guessing here.	Cecil Rhodes is associated with diamond mining and the formation of De Beer's diamond company. The diamonds were discovered shortly before Rhodes went to South Africa.	My father worked for DeBeer's and bored me to death with facts and figures about nineteenth century mineral discoveries. I never thought that such useless information would come in useful until now.
In what year was the first general election held in Japan?	1890	1930	Even under the Emperor they had some form of cabinet government, so a general election may have taken place before WWII.	Japan adopted western institutions in the nineteenth century, so the first general election couldn't have occurred before the late nineteenth century.	I have studied Japanese government. Western institutions were adopted in the time of Emperor Meji in 1868. Constitutional government was established in 1885 and the general election followed some years later.
In what year did Egypt nationalize the Suez Canal?	1956	1950	I think the Suez Canal was in the hands of the British & French after WW II, and I don't think it was after the assassination of JFK in '63.	The rise of Arab nationalism in the late 1940's and 50's, lead to intervention by the ex-colonial powers to hold on to the canal which was ultimately unsuccessful.	British, French and Israeli forces attempted to intervene to stop the Egyptian leader Nasser from nationalizing the canal, but were forced to withdraw by the American President Eisenhower. A particularly memorable date.

In what year was Rev. Martin Luther King assassinated?	1968	1964	In the 60's. At the height of the civil rights protests.	Martin Luther King is celebrated as the leader of the US civil rights movement, this was a turbulent time in the US and dates King's death to the 1960's.	This year was notable not only for the assassination of Martin Luther King, but also for the assassination of Bobby Kennedy and the suppression of the 'Prague Spring' by the USSR.
In what year was the'Boxer Rebellion' in China?	1900	1885	Wasn't there a film starring Charlton Heston about this? From what I recall it was set around the turn of the century.	This sounds like a reaction against European penetration of Chinese society, so I would imagine it took place in the latter years of the nineteenth century, and certainly before WWI.	The 'Boxer rebellion' was a reaction against European demands for territorial, railroad and mining concessions in China. I've read a book about it, so I know when the events took place.
In what year was the first Camp David summit between Egypt and Israel hosted by the United States?	1978	1985	This was on the news years ago, so in my lifetime - probably during the 1970's or 80's.	I think this famous summit ended formal hostilities between Egypt and Israel, despite Egyptian territorial losses in the various Arab-Israeli wars. It's got be at least 20 years ago. Either during the Carter or Reagan years.	This was the controversial summit between Sadat and Begin, hosted by President Carter. The summit led to the later Egyptian-Israeli peace treaty.

In what year did Thomas Eddison invent the light bulb?	1883	1863	Electric lighting has been around for some time – at least one hundred years.	Eddison is also associated with lighthouses. I think they were introduced around the middle of the nineteenth century, or maybe a little later. However, the light bulb would have been later still.	There's some controversy over who invented it first, but most people credit Eddison with it. It's a general knowledge thing that I just happen to know.
In what year did Britain receive a mandate to govern Palestine?	1920	1930	In the interwar years. Britain got out shortly after WWII.	Shortly after WWII the British got out, so it must have granted by the League of Nations in between the two world wars.	Britain received a mandate from the League of Nations to govern territories seized in the break-up of the Ottoman empire after WW I. I know this as I have been following the TV coverage and analysis of the Gaza pullout by Israel recently.

In what year did Thomas Jefferson become president of the United States?	1801	1785	Well, I know that the American Declaration of Independence was 1776. George Washington was the first president but I don't know how long he served for or who was the next president.	Thomas Jefferson was one of the founding fathers of the American Constitution, so I think he must have succeeded to the presidency after Washington.	Jefferson drafted the Declaration of Independence and was the 3rd President of the US following Washington and Adams.
In what year was the first Japanese railway opened?	1872	1892	They were probably building railways around the same time, as the US was, so maybe the late 1800's.	The Japanese rail system must have begun somewhat later than the great railroad network in America. As that coincided with the end of the Indian Wars, I think Japan's rail network begun at the end of the nineteenth century.	The introduction of western institutions came about in the Meji Restoration period of Japanese history. Britain successfully tendered for the contract to build the narrow gauge railway between Tokyo and Yokohama. It helps when you're a train spotter!

Appendix II: Permutations of reasons and advice

Reasons	Very weak	Weak	Strong
Advice	accurate	accurate	accurate
	inaccurate	inaccurate	inaccurate

Statistical advice (estimates of the dates of historical events are held constant across conditions) i.e. if the advice offered by advisor *K* to a question is '1830' in the statistical condition, '1830' is also the numerical advice offered in response to the same question in the reasons advice condition.

Appendix III

Appendix III: Experimental condition: control

Instructions

The following questions are intended to determine how you go about estimating uncertain quantities. You are asked to estimate a numerical answer to four questions involving uncertain quantities - it is highly unlikely that you will know the exact numerical answer. Nevertheless, we would like you to form your best estimate of the quantity in question, and tell us how confident you are in your answer. Also, you are asked to answer questions concerning how you think generally about estimation.

Please answer all questions in sequential order.

Many thanks for your participation

{Instructions were identical for all experimental conditions}

1. This question asks you to estimate the number of National Lottery tickets (main lotto game) that were sold in shops, garages, post offices etc., in the first week of June 2006 in the UK.

You are unlikely to know the actual number of National Lottery tickets sold each week in the UK, but you should be able to make an informed guess or estimate.

What is your best estimate? Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, that you wrote above.

For example, if your best estimate is 25 million, then you might be 90% confident that the true number lies between 20 and 30 million as shown below.

True value

Low = 20 million - - - - - - - - - ¦ - - - - - - - - - High = 30 million

Write your high and low values in the spaces provided below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Please enter your best estimate again here _____

Please enter the reasons for your best estimate (of the number of National Lottery tickets sold in the first week of June 2006) below –

As you have now had an opportunity to think about this issue do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no"

Yes No

If you now wish to change your best estimate do so now - otherwise move on to another question by. Please enter the changed estimate you are making here _____

We next move on to ask you to say how confident you are that your changed estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces, below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

2. This question asks you to estimate the number of people who enrolled for the first time at Higher Education Institutions in the UK in 2004/5.

You are unlikely to know the actual number of people enrolled in UK Higher Education institutions in 2004/5, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please enter below, where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously the low and the high estimates, should, between them, contain your best estimate, that you wrote above. Enter your high and low values in the spaces provided, below.

True value

Low = _____- - - - - - - - -¦- - - - - - - - - High = _____

Please enter your best estimate again here _____

Please enter the reasons for your best estimate (of the number of people who enrolled in Higher Education institutions in 2004/5) below –

As you have now had an opportunity to think about this issue do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no"

Yes No

If you now wish to change your best estimate do so now - otherwise move on to another question. Please enter the changed estimate you are making here _____

We next move on to ask you to say how confident you are that your changed estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces, below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

3. This question asks you to estimate the number of fatalities that resulted from UK road accidents in 2004.

You are unlikely to know the actual number of UK road accidents that resulted in fatalities in 2004, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please enter below, where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, Please write below where indicated, a low estimate and a high estimate, so that that you are 90% confident that the true answer lies between these points. You are 90% confident that the true answer lies between -

True value

Low = _____- - - - - - - - -¦- - - - - - - - - High = _____

Please enter your best estimate again here _____

Please enter the reasons for your best estimate (of the number of fatalities resulting from road accidents in the UK in 2004) below -

As you have now had an opportunity to think about this issue do you wish to change the best estimate that you gave in answer to the question?

Please indicate "yes" or "no"

Yes No

If you now wish to change your best estimate do so now - otherwise move on to another question.

Please enter the changed estimate you are making here _____

We next move on to ask you to say how confident you are that your changed estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces, below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

4. This question asks you to estimate the number of passengers per year that travel through Durham Tees Airport.

You are unlikely to know the actual number passengers that travel through Durham Tees Airport each year, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, that you wrote above. Write your high and low values in the spaces provided below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Please enter your best estimate again here _____

Please enter the reasons for your best estimate (of the number of passengers per year that travel through Durham Tees Airport) below -

As you have now had an opportunity to think about this issue do you wish to change the best estimate that you gave in answer to the question?

Please indicate "yes" or "no"

Yes No

If you now wish to change your best estimate do so now - otherwise move on to another question. Please enter the changed estimate you are making here _____

We next move on to ask you to say how confident you are that your changed estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces, below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Appendix III - Experimental condition: Process advice

1. This question asks you to estimate the number of National Lottery tickets (main lotto game) that were sold in shops, garages, post offices etc., in the first week of June 2006 in the UK.

You are unlikely to know the actual number of National Lottery tickets sold each week in the UK, but you should be able to make an informed guess or estimate.

What is your best estimate? Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, that you wrote above.

For example, if your best estimate is 25 million, then you might be 90% confident that the true number lies between 20 and 30 million as shown below.

True value

Low = 20 million - - - - - - - - - ¦ - - - - - - - - - High = 30 million

Write your high and low values in the spaces provided below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘Joe’ states that one way to form an estimate to this problem is to follow the instructions below. Please follow the instructions and write estimates where indicated –

(Step A) What is the average number of lottery terminals in a typical UK town/city?

Your answer _____

(Step B) How many towns/cities are there in the UK?

Your answer _____

(Step C) Multiply (A) times (B) to get the total number of lottery terminals in the UK.

(A) multiplied by (B) = _____

(Step D) How many tickets does each terminal issue per draw, on average?

Your answer _____

(Step E) Multiply (C) times (D) to get the total number of lottery tickets sold per draw for all lottery terminals.

(C) multiplied by (D) = _____

(Step F) How many main lottery draws are there per week?

Your answer _____

(Step G) Multiply (E) times (F) to get the total number of UK National Lottery tickets sold each week.

(E) multiplied by (F) = _____

This final figure is the outcome of Joe's advice to you. This advice may have helped you make a better estimate of the number of National Lottery tickets sold in the first week of June 2006 at outlets in the UK.

How useful was Joe's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following Joe's advice

here _____

Given Joe's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you answered 'No' above, then move onto the next question. If you answered 'Yes' please fill in below.

Please enter the changed estimate you are making in the light of the advice you received from Joe, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

2. This question asks you to estimate the number of people who enrolled for the first time at Higher Education Institutions in the UK in 2004/5.

You are unlikely to know the actual number of people enrolled in UK Higher Education institutions in 2004/5, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below, where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously the low and the high estimates, should, between them, contain your best estimate, that you wrote above. Write your high and low values in the spaces provided, below.

True value

Low = _____- - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘Kate’ states that one way to form an estimate to this problem is to follow the instructions below. Please follow the instructions and write estimates where indicated –

(Step A) What is the average number of institutions of Higher Education in each Higher Education region of the UK?

Your answer _____

(Step B) How many Higher Education regions are there in the UK?

Your answer _____

(Step C) Multiply (A) times (B) to get the total number of Higher Education institutions in the UK.

(A) multiplied by (B) = _____

(Step D) What is the average number of students applying to each institution of Higher Education?

Your answer _____

(Step E) Multiply (C) times (D) to get the total number of students applying for places in Higher Education 2004/5.

(C) multiplied by (D) = _____

(Step F) What proportion of (E) successfully applied for a place in Higher Education 2004/5?

Your answer _____

G. Multiply (E) times (F) to get the total number of people enrolled in Higher Education institutions in 2004/5

(E) multiplied by (F) = _____

This final figure is the outcome of Kate's advice to you. This advice may have helped you make a better estimate of the number of people who enrolled at Higher Education institutions in 2004/5 in the UK.

How useful was Kate's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following Kate's advice here _____

Given Kate's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below –

Yes No

If you answered ‘No’ above, then move onto the next question. If you answered ‘Yes’ please fill in below.

Please enter the changed estimate you are making in the light of the advice you received from Kate, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____- - - - - - - - - ¦ - - - - - - - - - High = _____

3. This question asks you to estimate the number of fatalities that resulted from UK road accidents in 2004.

You are unlikely to know the actual number of UK road accidents that resulted in fatalities in 2004, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below, where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, Please write below where indicated, a low estimate and a high estimate, so that that you are 90% confident that the true answer lies between these points. You are 90% confident that the true answer lies between –

True value

Low = _____- - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘David’ states that one way to form an estimate to this problem is to follow the instructions below. Please follow the instructions and write your estimate where indicated –

(Step A) How many people in the UK hold driving licences?

Your answer = _____

(Step B) What proportion of (A) actually drives?

Your Answer _____

C. Multiply (A) times (B) to get the total number of car drivers in the UK.

(A) multiplied by (B) = _____

(Step D) How many accidents does the average driver have per year?

Your answer _____

(Step E) Multiply (C) times (D) to get the total number of road accidents per year.

(C) multiplied by (D) = _____

(Step F) What proportion of (E) are fatal accidents?

Your answer _____

(Step G) Multiply (E) times (F) to get the total number of fatal road accidents in the UK in 2004.

(E) multiplied by (F) = _____

This final figure is the outcome of David's advice to you. This advice may have helped you make a better estimate of the number of fatalities that resulted from UK road accidents in 2004.

How useful was David's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following David's advice

here _____

Given David's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you answered 'No' above, then move onto the next question. If you answered 'Yes' please fill in below.

Please enter the changed estimate you are making in the light of the advice you received from David, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

4. This question asks you to estimate the number of passengers per year that travel through Durham Tees Airport.

You are unlikely to know the actual number of passengers that travel through Durham Tees Airport each year, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, that you wrote above. Write your high and low values in the spaces provided below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘Tina’ states that one way to form an estimate to this problem is to follow the instructions below. Please follow the instructions and write your estimates where indicated –

(Step A) How many passenger airlines fly to/from Durham Tees?

Your answer _____

(Step B) How many landings/takeoff flights a day does an average passenger airline make from Durham Tees Airport?

Your answer _____

(Step C) Multiply (A) times (B) to get the total number of flights per day utilizing Durham Tees Airport.

(A) multiplied by (B) = _____

(Step D) What is the average number of passengers on each flight that departs/lands at Durham Tees Airport?

Your answer _____

(Step E) Multiply (C) times (D) to get the average number of passengers per day that travel through Durham Tees Airport.

(C) multiplied by (D) = _____

(Step F) How many days in a year?

Your answer _____

(Step G) Multiply (E) times (F) to get the total number of passengers that travel through Durham Tees Airport per year.

(E) multiplied by (F) = _____

This final figure is the outcome of Tina's advice to you. This advice may have helped you make a better estimate of the number of passengers that travel through Durham Tees Airport per year.

How useful was Tina's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following Tina's advice

here _____

Given Tina's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you answered 'No' above, then move onto the next question. If you answered 'Yes' please fill in below.

Please enter the changed estimate you are making in the light of the advice you received from Tina, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____- - - - - - - - - ¦ - - - - - - - - - High = _____

Appendix III - Experimental conditions: Non-Process advice

1. This question asks you to estimate the number of National Lottery tickets (main lotto game) that were sold in shops, garages, post offices etc., in the first week of June 2006 in the UK.

You are unlikely to know the actual number of National Lottery tickets sold each week in the UK, but you should be able to make an informed guess or estimate.

What is your best estimate? Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, that you wrote above.

For example, if your best estimate is 25 million, then you might be 90% confident that the true number lies between 20 and 30 million as shown below.

True value

Low = 20 million - - - - - - - - - ¦ - - - - - - - - - High = 30 million

Write your high and low values in the spaces provided below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘Joe’ advises you that-

> Not everyone in the UK participates in the National Lottery, and there are other games such as scratchcards. The main draw is on a Saturday night, but there is another draw on a Wednesday. So, I would think that approximately 20 million people participated in the National Lottery in the first week of June 2006.

This final figure is the outcome of Joe's advice to you. This advice may have helped you make a better estimate of the number of National Lottery tickets sold in the first week of June 2006 at outlets in the UK.

How useful was Joe's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following Joe's advice

here _____

Given Joe's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you now wish to change your best estimate in the light of Joe's advice you can do so now - otherwise move on to another question. Please write the changed estimate you are making in the light of the advice you received from Joe, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

2. This question asks you to estimate the number of people who enrolled for the first time at Higher Education Institutions in the UK in 2004/5.

You are unlikely to know the actual number of people enrolled in UK Higher Education institutions in 2004/5, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below, where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously the low and the high estimates, should, between them, contain your best estimate, that you wrote above. Write your high and low values in the spaces provided, below.

True value

Low = _____- - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘Kate’ advises you that –

> I suppose you would have to take into account UK school leavers, international students, and adult returners. Also, I think the government had a target of half all school leavers going on to higher education. So, I think that it is approximately 50 million.

This final figure is the outcome of Kate's advice to you. This advice may have helped you make a better estimate of the number of people who enrolled at Higher Education institutions in 2004/5 in the UK.

How useful was Kate's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following Kate's advice

here _____

Given Kate's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below –

Yes No

If you now wish to change your best estimate in light of Kate's advice you can do so now - otherwise move on to answer the next question. Please write the changed estimate you are now making in the light of the advice you received from Kate, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____- - - - - - - - - ¦ - - - - - - - - - High = _____

3. This question asks you to estimate the number of fatalities that resulted from UK road accidents in 2004.

You are unlikely to know the actual number of UK road accidents that resulted in fatalities in 2004, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below, where indicated, a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, Please write below where indicated, a low estimate and a high estimate, so that that you are 90% confident that the true answer lies between these points. You are 90% confident that the true answer lies between -

True value

Low = _____- - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘David’ advises you that –

> Deaths on the road are quite commonplace – you hear about them on the news all the time. Over a year, the figures must be in there thousands, so, I would say approximately 10,000.

This final figure is the outcome of David's advice to you. This advice may have helped you make a better estimate of the number of fatalities that resulted from UK road accidents in 2004.

How useful was David's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following David’s advice

here _____

Given David's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you now wish to change your best estimate in light of David's advice you can do so now - otherwise move on to another question. Please write the changed estimate you are making in the light of the advice you received from David, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

4. This question asks you to estimate the number of passengers per year that travel through Durham Tees Airport.

You are unlikely to know the actual number of passengers that travel through Durham Tees Airport each year, but you should be able to make an informed guess or estimate.

What is your best estimate?

Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, that you wrote above. Write your high and low values in the spaces provided below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘Tina’ advises you that –

> Durham Tees Airport is not one of the major UK airports, and there are other airports such as Newcastle fairly close by. You would have to account for passenger arrivals and departures each day, and adjust for seasonal variations. So, I would say approximately 500,000.

This final figure is the outcome of Tina's advice to you. This advice may have helped you make a better estimate of the number of passengers that travel through Durham Tees Airport per year.

How useful was Tina's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following Tina’s advice

here _____

Given Tina's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you now wish to change your best estimate in light of Tina's advice you can do so now - otherwise move on to another question. Please write the changed estimate you are making in the light of the advice you received from Tina, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____- - - - - - - - - ¦- - - - - - - - - High = _____

Appendix IV

Appendix IV: US mail estimation problem

Instructions

The following question is intended to determine how you go about estimating an uncertain quantity. However, it is highly unlikely that you will know the exact numerical answer. Nevertheless, we would like you to form your best estimate of the quantity in question, and tell us how confident you are in your answer.

Question: How many pieces of mail were handled by the US postal service in 1987?

What is your best estimate? Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, that you wrote above.

For example, if your best estimate is 25 million, then you might be 90% confident that the true number lies between 20 and 30 million as shown below.

True value

Low = 20 million - - - - - - - - - ¦ - - - - - - - - - High = 30 million

Write your high and low values in the spaces provided below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor 'Joe' states that one way to form an estimate to this problem is to follow the instructions below. Please follow the instructions and write estimates where indicated –

A. What was the average number of post offices per state in the USA in 1987? Put your estimate here _____

B. What is the number of states in the USA?

Put your estimate here _____

C. Multiply (A) times (B) to get the total number of post offices in the USA in 1987. (A) x (B) = _____

D. How many pieces of mail per day were handled by the average post office in 1987? Put your estimate here _____

E. Multiply (C) times (D) to get the total pieces of mail per day for all post-offices in the USA in 1987. (C) x (D) = _____

F. How many days are there in a year?

Put your estimate here _____

G. Multiply (E) times (F) to get the number of pieces of mail handled in a year by the US postal service in 1987.

(E) x (F) = _____

This final figure is the outcome of Joe's advice to you. This advice may have helped you make a better estimate of the number of pieces of mail that the US postal service handled in 1987.

How useful was Joe's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following Joe's advice

here _____

Given Joe's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you now wish to change your best estimate in the light of Joe's advice you can do so now. Please write the changed estimate you are making in the light of the advice you received from Joe, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Thanks for your participation

Appendix IV: Forested Miles in Oregon estimation problem

Instructions

The following question is intended to determine how you go about estimating an uncertain quantity. However, it is highly unlikely that you will know the exact numerical answer. Nevertheless, we would like you to form your best estimate of the quantity in question, and tell us how confident you are in your answer.

Question: How many Forested Miles are there in the US state of Oregon?

What is your best estimate? Put that amount here _____

We next move on to ask you to say how confident you are that your best estimate, above, is correct. Please write below where indicated, a low estimate and a high estimate, such that that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should, between them, contain your best estimate, that you wrote above.

For example, if your best estimate is 25 million, then you might be 90% confident that the true number lies between 20 and 30 million as shown below.

True value

Low = 20 million - - - - - - - - - ¦ - - - - - - - - - High = 30 million

Write your high and low values in the spaces provided below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Anonymous advisor ‘Kate’ states that one way to form an estimate to this problem is to follow the instructions below. Please follow the instructions and write estimates where indicated –

A. What is the distance in miles between the Oregon/California border and the Oregon/Washington border?

Put your estimate here _____

B. What is the distance in miles between the Oregon coast and the Oregon/Idaho border? Put your estimate here _____

C. Multiply (A) times (B) to get an estimate of the area of Oregon in square miles. (A) x (B) = _____

D. What proportion of Oregon is forested?

Put your estimate here _____

E. Multiply (C) times (D) to get the number of Forested Miles in Oregon. (C) x (D) = _____

This final figure is the outcome of Kate's advice to you. This advice may have helped you make a better estimate of the number of Forested Miles in the US state of Oregon. How useful was Kate's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your previous best estimate here _____

Please write the estimate you calculated following Kate's advice

here _____

Given Kate's advice, do you wish to change the best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you now wish to change your best estimate in the light of Kate's advice you can do so now. Please write the changed estimate you are making in the light of the advice you received from Kate, here _____

We next move on to ask you to say how confident you are that your changed estimate, above is correct. Please write below a low estimate and a high estimate, such that you are 90% confident that the true answer, whatever it is, lies between these points. Obviously, the low and the high estimates should contain your changed estimate. Write your high and low values in the spaces below.

True value

Low = _____ - - - - - - - - - ¦ - - - - - - - - - High = _____

Thanks for your participation

Appendix V

Appendix V: Road fatalities estimation problem

Instructions

Road safety has become an increasingly important issue both for the UK government, and the general public in recent years - given the increases in the numbers of motorists, year on year. Regrettably, fatal accidents occur throughout the year on Great Britain's road network involving both motorists and pedestrians. In the absence of accurate statistics, or where the cost of obtaining such information is prohibitively high, how would road safety managers assess the chance of a pedestrian being killed in a motor vehicle accident during the 12 months of 2006? This brief questionnaire is designed to elicit how *you* would go about such a problem.

Before you begin, please indicate your age and gender below –

Age: _____

Gender: _____ (male or female)

One way of expressing chance is by measuring a specific event's historical *relative frequency*. This is the ratio between the historic incidence of the target event, and the total number of incidences in which the target event *could have* occurred (e.g. a relative frequency of 1 in 1 is an absolute certainty that an event will happen; whilst a relative frequency of 1 in 2 is greater than a relative frequency of 1 in 10,000 that an event will happen).

The following question is intended to determine how you go about estimating the relative frequency of a specific event. However, it is highly unlikely that you will know the exact numerical answer. Nevertheless, we would like you to form your best estimate of the relative frequency in question.

Question: What was the chance (expressed as a relative frequency) that any individual pedestrian would die in a motor vehicle traffic accident in Great Britain in 2006?

So, your task is to determine the relative frequency of a (random) pedestrian being killed in a motor vehicle accident in 2006. Your best estimate of the chance (expressed as a relative frequency) that any individual pedestrian is killed in a motor vehicle accident in 2006 is: 1 in _____ (Put your answer in the space provided).

Anonymous advisor ‘Joe’ states that one way to form an estimate to this problem is to follow the instructions below. Sometimes you will be asked to estimate proportions, to do so please write these proportional quantities as numbers between 0 and 1 in the spaces provided. (Hence one-tenth is expressed as 0.1; ½ is expressed as 0.5; and nine-tenths is expressed as 0.9) Please follow the instructions and write estimates where indicated –

What was the chance (expressed as a relative frequency) that any individual pedestrian would die in a motor vehicle traffic accident in Great Britain (GB) in 2006?

A. What was the total population of GB in 2006? _____

B. What proportion of A owned a licensed motor vehicle in 2006? ____

C. Multiply A x B to calculate the total number of motor vehicles in GB in 2006 _____

D. For every 1000 motor vehicles on the roads in GB in 2006, how many were involved in a road traffic accident? _____

E. Divide D by 1000 _____

F. Multiply C x E to calculate the number of motor vehicle accidents in GB in 2006 _____

G. For every 1000 motor vehicle accidents that occurred in GB in 2006, how many resulted in fatalities? _____

H. Divide G by 1000 _____

I. Multiply F x H to calculate the total number of motor vehicle accident fatalities in GB in 2006 _____

J. Of the number you have calculated in step I, how many were pedestrians? ____

K. Divide A by J, to calculate the probability that a pedestrian died in a motor vehicle accident in GB in 2006: 1 in ____

This final figure is the outcome of Joe's advice to you. This advice may help you to make a better estimate of the chance (expressed as a relative frequency) of any individual pedestrian being killed in a motor vehicle accident in 2006. How useful was Joe's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your original best estimate of the chance (expressed as a relative frequency) that any individual pedestrian would be killed in a motor vehicle accident in 2006 here: 1 in _____

Please write the estimate of the chance (expressed as a relative frequency) you calculated following Joe's advice here: 1 in _____

Given Joe's advice, do you wish to change the original best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you now wish to change your original best estimate of the chance (expressed as a relative frequency) that any individual pedestrian is killed in a motor vehicle accident in 2006, in the light of Joe's advice, you can do so now. Please write the changed estimate you are making in the light of the advice you received from Joe, here: 1 in _____

Thanks for your participation

Appendix V: Multiple maternities problem

Instructions

Clinicians generally regard multiple maternities as more problematic and potentially dangerous for mothers, and babies, than single births. These maternities often require different medical resources than single births, so a salient question for health managers is how likely it is that a pregnancy will result in a triplet birth In the absence of accurate statistics, or where the cost of obtaining such information is prohibitively high, how are managers able to determine the chance of a triplet birth? This brief questionnaire is designed to elicit how *you* would go about such a problem.

Before you begin, please indicate your age and gender below –

Age: _____

Gender: _____ (male or female)

One way of expressing chance is by measuring a specific event's historical *relative frequency*. This is the ratio between the historic incidence of the target event, and the total number of incidences in which the target event *could have* occurred. (e.g. a relative frequency of 1 in 1 is an absolute certainty that an event will happen; whilst a relative frequency of 1 in 2 is greater than a relative frequency of 1 in 10,000 that an event will happen).

The following question is intended to determine how you go about estimating the relative frequency of a specific event. However, it is highly unlikely that you will know the exact numerical answer. Nevertheless, we would like you to form your best estimate of the relative frequency of the event in question.

Question: What is the chance (expressed as a relative frequency) that, in 2005, a pregnancy would result in a triplet birth in the UK?

So, your task is to determine the relative frequency of a triplet birth resulting from a pregnancy in 2005 in the UK.

Your best estimate that the chance (expressed as a relative frequency) of a pregnancy in 2005 resulting in a triplet birth in the UK is:

1 in _____ (Put your answer in the space provided).

Anonymous advisor 'Kate' states that one way to form an estimate to this problem is to follow the instructions below. Sometimes you will be asked to estimate proportions, to do so please write these proportional quantities as numbers between 0 and 1 in the spaces provided. (Hence one-tenth is expressed as 0.1; ½ is expressed as 0.5; and nine-tenths is expressed as 0.9) Please follow the instructions and write estimates where indicated –

What is the chance (expressed as a relative frequency) that, in 2005, a pregnancy would result in a triplet birth in the UK?

A. How many females were there in the UK in 2005? _____

B. What proportion of A were females able to have children (i.e. aged 16 – 40 years)? _____

C. Multiply A x B to calculate the number of potential maternities in 2005 _____

D. What proportion of C were pregnant in 2005? _____

E. Multiply C x D to calculate the number of maternities in 2005 _____

F. How many births in E were multiple maternities? _____

G. Divide E by F _____

H. The chance (expressed as a relative frequency) of a pregnancy resulting in a triplet birth in 2005 was 1 in _____ (insert the number you calculated in step G in the space provided here).

This final figure is the outcome of Kate's advice to you. This advice may help you to make a better estimate of the chance (expressed as a relative frequency) of a pregnancy in the UK in 2005 resulting in a triplet birth.

How useful was Kate's advice to you in estimating the correct answer to the question?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

Please write your original best estimate of the chance (expressed as a relative frequency) of a pregnancy resulting in a triplet birth in the UK in 2005 here: 1 in _____

Please write the estimate of chance you calculated following Kate's advice here: 1 in _____

Given Kate's advice, do you wish to change the original best estimate that you gave in answer to the question?

Please circle "yes" or "no" below -

Yes No

If you now wish to change your original best estimate of the chance (expressed as a relative frequency) of a pregnancy in the UK in 2005 resulting in a triplet birth, in the light of Kate's advice, you can do so now. Please write the changed estimate you are making in the light of the advice you received from Kate, here: 1 in _____

Thanks for your participation

Appendix VI

Perceptions about international rates of infant mortality

This study examines the perceptions people may hold about infant mortality rates (IMR) in various countries throughout the world.

On the following pages you will be asked to make simple quantitative estimates about IMR's in ten different countries.

No complex calculations are necessary, you need only use your existing knowledge (and maybe a bit of guesswork!) to complete the questionnaire.

Introduction & Instructions

This questionnaire is concerned with the perceptions people may hold about infant mortality rates (IMR) in various countries around the world.

IMR is the ratio of infant mortality per 1000 live births (e.g. an IMR of 5.5 is representative of 5.5 infant deaths per 1000 live births on average; an IMR of 500 would indicate that 1 in 2 children die within the first 12 months of life; whilst an IMR of 995 per 1000 live births would indicate that nearly all children die within the first 12 months of life). Whilst statistics such as these are compiled by governments and various international health organisations, most non-expert people have only a vague idea of the true rate of infant mortality in their own, or in other peoples' countries.

The purpose of this study is to determine the extent of your knowledge of international infant mortality rates (IMR). On the following page you will be asked to estimate the IMR for 10 countries.

Before you continue could you tell us a little about yourself by indicating your Age, and Gender below.

Age ____

Gender: Male ☐ Female ☐ (tick the appropriate box)

Your estimate of international infant mortality rates

Please estimate the infant mortality rates (IMR) of the 10 countries listed below. This task has two components.

First, rank order the countries below, by what you think is each individual country's infant mortality rate (IMR). Place a number between 1- 10 in the left column of the table below - '1' indicates the country with the LOWEST rate of infant mortality, and '10' indicates the country with the HIGHEST rate of infant mortality. No two countries have the same rate of infant mortality, so each number between 1-10 can only be used once.

Second, make an estimate of the actual rate of infant mortality in each of the countries below (place your estimate in the right column of the table below). Recall that the rate of infant mortality is calculated as the number of infant fatalities per 1000 live births.

Rank Order by IMR	Country	Estimate of actual IMR (infant deaths per 1000 live births)
	Mozambique	
	Niger	
	Liberia	
	Sierra-Leone	
	Angola	
	Iceland	
	Hong Kong	
	Japan	
	Sweden	
	Singapore	

Appendix VI: Control condition (including Need for Cognition scale)

This next section concerns your individual thinking style, please complete all questions.

For each of the statements below, please indicate to what extent the statement is characteristic of you. If the statement is extremely uncharacteristic of you (not at all like you) please circle "1"; if the statement is extremely characteristic of you (very much like you) please circle "5" under the question. Of course, a statement may be neither extremely uncharacteristic nor extremely characteristic of you; if so, please circle the number in the middle of the scale that describes the best fit. Please keep the following scale in mind as you rate each of the statements below: 1 = extremely uncharacteristic; 2 = somewhat uncharacteristic; 3 = uncertain; 4 = somewhat characteristic; 5 = extremely characteristic.

1. I would prefer complex to simple problems.

   1 2 3 4 5

2. I like to have the responsibility of handling a situation that requires a lot of thinking.

   1 2 3 4 5

3. Thinking is not my idea of fun.

   1 2 3 4 5

4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.

   1 2 3 4 5

5. I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something.

   1 2 3 4 5

6. I find satisfaction in deliberating hard and for long hours.

   1 2 3 4 5

7. I only think as hard as I have to.

   1 2 3 4 5

8. I prefer to think about small, daily projects to long-term ones.

   1 2 3 4 5

9. I like tasks that require little thought once I've learned them.

   1 2 3 4 5

10. The idea of relying on thought to make my way to the top appeals to me.

1 2 3 4 5

11. I really enjoy a task that involves coming up with new solutions to problems.

1 2 3 4 5

12. Learning new ways to think doesn't excite me very much.

1 2 3 4 5

13. I prefer my life to be filled with puzzles that I must solve.

1 2 3 4 5

14. The notion of thinking abstractly is appealing to me.

1 2 3 4 5

15. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

1 2 3 4 5

16. I feel relief rather than satisfaction after completing a task that required a lot of mental effort.

1 2 3 4 5

17. It's enough for me that something gets the job done; I don't care how or why it works.

1 2 3 4 5

18. I usually end up deliberating about issues even when they do not affect me personally.

1 2 3 4 5

Since you have now had a short period of time upon which to reflect about the infant mortality rates (IMR) of 10 different countries, please re-evaluate your previous estimate of each IMR for the 10 countries below.

First, you should copy the information you provided on page 3 into the first two columns on the left of the table below. Next, follow the same procedure as you did earlier by rank ordering the 10 countries by each state's infant mortality rate (IMR). Place a number between 1- 10 in the left column of the table below - '1' indicates the country with the LOWEST rate of infant mortality, and '10' indicates the country with the HIGHEST rate of infant mortality. No two countries have the same rate of infant mortality, so each number between 1-10 can only be used once. Should you decide not to revise your estimate, please copy your original estimate into the appropriate column in the right of the table below.

Second, revise your estimate of the actual rate of infant mortality in each of the countries below (place your estimate in the appropriate right column of the table below). Recall that the rate of infant mortality is calculated as the number of infant fatalities per 1000 live births. Should you decide not to revise your estimate, please copy your original estimate into the appropriate column in the right of the table below.

Your original rank ordering by IMR	Your original estimate of actual IMR (infant deaths per 1000 live births)	Country	Revised rank order by IMR	Revised estimate of actual IMR (infant deaths per 1000 live births)
		Mozambique		
		Niger		
		Liberia		
		Sierra-Leone		
		Angola		
		Iceland		
		Hong Kong		
		Japan		
		Sweden		
		Singapore		

That concludes our questionnaire – many thanks for your participation!

Perceptions about international rates of infant mortality

This study examines the perceptions people may hold about infant mortality rates (IMR) in various countries throughout the world.

On the following pages you will be asked to make simple quantitative estimates about IMRs in ten different countries.

No complex calculations are necessary, you need only use your existing knowledge (and maybe a bit of guesswork!) to complete the questionnaire.

Introduction and Instructions

This questionnaire is concerned with the perceptions people may hold about infant mortality rates (IMR) in various countries around the world.

IMR is the ratio of infant mortality per 1000 live births (e.g. an IMR of 995 per 1000 live births would indicate that nearly all children die within the first 12 months of life; an IMR of 500 would indicate that 1 in 2 children die within the first 12 months of life; an IMR of 5.5 is representative of 5.5 infant deaths per 1000 live births, on average). Whilst statistics such as these are compiled by governments and various international health organisations, most non-expert people have only a vague idea of the true rate of infant mortality in their own, or in other peoples' countries.

The purpose of this study is to determine the extent of your knowledge of international infant mortality rates (IMR). On the following page you will be asked to estimate the IMR for 10 countries.

Before you continue tell us a little about yourself by indicating your Age, and Gender below.

Age ____

Gender: Male ☐ Female ☐ (tick the appropriate box)

Your estimate of international infant mortality rates

Please estimate the infant mortality rates (IMR) of the 10 countries listed below. This task has two components.

First, rank order the countries below, by what you think is each individual country's infant mortality rate (IMR). Place a number between 1- 10 in the left column of the table below - '1' indicates the country with the LOWEST rate of infant mortality, and '10' indicates the country with the HIGHEST rate of infant mortality. No two countries have the same rate of infant mortality, so each number between 1-10 can only be used once.

Second, make an estimate of the actual rate of infant mortality in each of the countries below (place your estimate in the right column of the table below). Although the rate of infant mortality is calculated as the number of infant fatalities per 1000 live births, the infant mortality rates for all of the countries below are between 2 deaths per 1000 live births, and 185 deaths per 1000 live births.

Rank Order by IMR	Country	Estimate actual IMR (infant deaths per 1000 live births)
	Singapore	
	Sweden	
	Japan	
	Hong Kong	
	Iceland	
	Angola	
	Sierra-Leone	
	Liberia	
	Niger	
	Mozambique	

Appendix VI: Gross Domestic Product (GDP) ($ per capita)

Most people make poor estimates of infant mortality rates (IMRs), either because they are unfamiliar with such statistics, and/or they have insufficient knowledge of other countries to form an accurate estimate. It may be that you *would* have been able to make a more accurate estimate of a countries' IMR, if you had access to some accurate additional information. In this section we provide you with accurate per capita GDP ($) figures for each of the ten countries under consideration, (*Source*: 2007 CIA Factbook). Per capita GDP is the value of how much a person produces in a year. It is calculated by dividing the Gross Domestic Product for a country by the number of people who live there. Here, you will have an opportunity to re-consider your original responses in the light of this additional information, and make new estimates that are potentially more accurate than those you made initially.

First, you should copy the information you provided on page 3 into the first two columns on the left of the table below. Second, please re-assess the rank order that you assigned the 10 countries below in the light of the additional GDP information, and enter your revised estimate in the appropriate column on the right of the table below. Should you decide not to amend your estimate, please copy your original estimate into the appropriate column in the right of the table below.

Next, please re-evaluate your original estimate of each country's infant mortality rate (IMR) in the light of the additional GDP information, and enter your revised estimate in the appropriate column in the right of the table below. Although the rate of infant mortality is calculated as the number of infant fatalities per 1000 live births, the infant mortality rates for all of the countries below are between 2 deaths per 1000 live births, and 185 deaths per 1000 live births. Should you decide not to amend your estimate, please copy your original estimate into the appropriate column in the right of the table below.

Your original rank ordering by IMR	Your original estimate of actual IMR (infant deaths per 1000 live births)	Country	GDP Per capita ($)	Revised rank order by IMR	Revised estimate of actual IMR (infant deaths per 1000 live births)
		Singapore	$31400		
		Sweden	$32200		
		Japan	$33100		
		Hong Kong	$37300		
		Iceland	$38000		
		Angola	$4400		
		Sierra-Leone	$900		
		Liberia	$900		
		Niger	$1000		
		Mozambique	$1500		

Finally, we want to know how useful you found the additional information of each country's per capita GDP.

How useful was information of each countries' per capita GDP ($) to you in estimating the true infant mortality rate (IMR) of each country?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

That concludes our questionnaire – Many thanks for your participation!

Perceptions about international rates of infant mortality

This study examines the perceptions people may hold about infant mortality rates (IMR) in various countries throughout the world.

On the following pages you will be asked to make simple quantitative estimates about IMRs in ten different countries.

No complex calculations are necessary, you need only use your existing knowledge (and maybe a bit of guesswork!) to complete the questionnaire.

Introduction and Instructions

This questionnaire is concerned with the perceptions people may hold about infant mortality rates (IMR) in various countries around the world.

IMR is the ratio of infant mortality per 1000 live births (e.g. an IMR of 995 per 1000 live births would indicate that nearly all children die within the first 12 months of life; an IMR of 500 would indicate that 1 in 2 children die within the first 12 months of life; an IMR of 5.5 is representative of 5.5 infant deaths per 1000 live births, on average). Whilst statistics such as these are compiled by governments and various international health organisations, most non-expert people have only a vague idea of the true rate of infant mortality in their own, or in other peoples' countries.

The purpose of this study is to determine the extent of your knowledge of international infant mortality rates (IMR). On the following page you will be asked to estimate the IMR for 10 countries.

Before you continue could you tell us a little about yourself by indicating your Age, and Gender below.

Age ____

Gender: Male ☐ Female ☐ (tick the appropriate box)

Your estimate of international infant mortality rates

Please estimate the infant mortality rates (IMR) of the 10 countries listed below. This task has two components.

First, rank order the countries below, by what you think is each individual country's infant mortality rate (IMR). Place a number between 1- 10 in the left column of the table below - '1' indicates the country with the LOWEST rate of infant mortality, and '10' indicates the country with the HIGHEST rate of infant mortality. No two countries have the same rate of infant mortality, so each number between 1-10 can only be used once.

Second, make an estimate of the actual rate of infant mortality in each of the countries below (place your estimate in the right column of the table below). Although the rate of infant mortality is calculated as the number of infant fatalities per 1000 live births, the infant mortality rates for all of the countries below are between 2 deaths per 1000 live births, and 185 deaths per 1000 live births.

Rank Order by IMR	Country	Estimate actual IMR (infant deaths per 1000 live births)
	Singapore	
	Sweden	
	Japan	
	Hong Kong	
	Iceland	
	Angola	
	Sierra-Leone	
	Liberia	
	Niger	
	Mozambique	

Appendix VI: Population statistics

Most people make poor estimates of infant mortality rates (IMRs), either because they are unfamiliar with such statistics, and/or they have insufficient knowledge of other countries to form an accurate estimate. It may be that you *would* have been able to make a more accurate estimate of a countries' IMR, if you had access to some accurate information. In this section we provide you with accurate population estimates for each of the ten countries under consideration, (*Source*: 2007 CIA Factbook). Here, you will have an opportunity to re-consider your original responses in the light of this additional information, and make new estimates that are potentially more accurate than those you made initially.

First, you should copy the information you provided on page 3 into the first two columns on the left of the table below. Second, please re-assess the rank order that you assigned the 10 countries below in the light of the additional population estimate information, and enter your revised estimate in the appropriate column on the right of the table below. Should you decide <u>not</u> to amend your estimate, please copy your original estimate into the appropriate column in the right of the table below.

Next, please re-evaluate your original estimate of each country's infant mortality rate (IMR) in the light of the additional population estimate information, and enter your revised estimate in the appropriate column in the right of the table below. Although the rate of infant mortality is calculated as the number of infant fatalities per 1000 live births, the infant mortality rates for all of the countries below are between 2 deaths per 1000 live births, and 185 deaths per 1000 live births. Should you decide <u>not</u> to amend your estimate, please copy your original estimate into the appropriate column in the right of the table below.

Your original rank ordering by IMR	Your original estimate of actual IMR (infant deaths per 1000 live births)	Country	Population estimate	Revised rank order by IMR	Revised estimate of actual IMR (infant deaths per 1000 live births)
		Singapore	4,553,009		
		Sweden	9,031,088		
		Japan	127,433,494		
		Hong Kong	6,980,412		
		Iceland	301,931		
		Angola	12,263,596		
		Sierra-Leone	6,144,562		
		Liberia	3,195,931		
		Niger	12,894,865		
		Mozambique	20,905,585		

Finally, we want to know how useful you found the additional population estimate information for each country.

How useful was information of each country's population estimate to you in estimating the true infant mortality rate (IMR) of each country?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

That concludes our questionnaire – Many thanks for your participation!

Perceptions about international rates of infant mortality

This study examines the perceptions people may hold about infant mortality rates (IMR) in various countries throughout the world.

On the following pages you will be asked to make simple quantitative estimates about IMRs in ten different countries.

No complex calculations are necessary, you need only use your existing knowledge (and maybe a bit of guesswork!) to complete the questionnaire.

Introduction and Instructions

This questionnaire is concerned with the perceptions people may hold about infant mortality rates (IMR) in various countries around the world.

IMR is the ratio of infant mortality per 1000 live births (e.g. an IMR of 5.5 is representative of 5.5 infant deaths per 1000 live births on average; an IMR of 500 would indicate that 1 in 2 children die within the first 12 months of life; whilst an IMR of 995 per 1000 live births would indicate that nearly all children die within the first 12 months of life). Whilst statistics such as these are compiled by governments and various international health organisations, most non-expert people have only a vague idea of the true rate of infant mortality in their own, or in other peoples' countries.

The purpose of this study is to determine the extent of your knowledge of international infant mortality rates (IMR). On the following page you will be asked to estimate the IMR for 10 countries.

Before you continue could you tell us a little about yourself by indicating your Age, and Gender below.

Age ____

Gender: Male ☐ Female ☐ (tick the appropriate box)

Your estimate of international infant mortality rates

Please estimate the infant mortality rates (IMR) of the 10 countries listed below. This task has two components.

First, rank order the countries below, by what you think is each individual country's infant mortality rate (IMR). Place a number between 1- 10 in the left column of the table below - '1' indicates the country with the LOWEST rate of infant mortality, and '10' indicates the country with the HIGHEST rate of infant mortality. No two countries have the same rate of infant mortality, so each number between 1-10 can only be used once.

Second, make an estimate of the actual rate of infant mortality in each of the countries below (place your estimate in the right column of the table below). Recall that the rate of infant mortality is calculated as the number of infant fatalities per 1000 live births.

Rank Order by IMR	Country	Estimate actual IMR (infant deaths per 1000 live births)
	Mozambique	
	Niger	
	Liberia	
	Sierra-Leone	
	Angola	
	Iceland	
	Hong Kong	
	Japan	
	Sweden	
	Singapore	

Appendix VI: GDP per capita ($) and Population statistics

Most people make poor estimates of infant mortality rates (IMRs), either because they are unfamiliar with such statistics, and/or they have insufficient knowledge of other countries to form an accurate estimate. It may be that you *would* have been able to make a more accurate estimate of a countries' IMR, if you had access to some accurate information. In this section we provide you with accurate per capita GDP ($), and population estimates for each of the ten countries under consideration, (*Source*: 2007 CIA Factbook). Per capita GDP is the value of how much a person produces in a year. It is calculated by dividing the Gross Domestic Product for a country by the number of people who live there. Here, you will have an opportunity to re-consider your original responses in the light of this additional information, and make new estimates that are potentially more accurate than those you made initially.

First, you should copy the information you provided on page 3 into the first two columns on the left of the table below. Second, please re-assess the rank order that you assigned the 10 countries below in the light of the additional GDP and population estimate information, and enter your revised estimate in the appropriate column on the right of the table below. Should you decide <u>not</u> to amend your estimate, please copy your original estimate into the appropriate column in the right of the table below.

Next, please re-evaluate your original estimate of each country's infant mortality rate (IMR) in the light of the additional GDP and population estimate information, and enter your revised estimate in the appropriate column in the right of the table below. Although the rate of infant mortality is calculated as the number of infant fatalities per 1000 live births, the infant mortality rates for all of the countries below are between 2 deaths per 1000 live births, and 185 deaths per 1000 live births. Should you decide <u>not</u> to amend your estimate, please copy your original estimate into the appropriate column in the right of the table below.

Your original rank ordering by IMR	Your original estimate of actual IMR (infant deaths per 1000 live births)	Country	GDP Per capita ($)	Population estimate	Revised rank order by IMR	Revised estimate of actual IMR (infant deaths per 1000 live births)
		Mozambique	$1500	20,905,585		
		Niger	$1000	12,894,865		
		Liberia	$900	3,195,931		
		Sierra-Leone	$900	6,144,562		
		Angola	$4400	12,263,596		
		Iceland	$38000	301,931		
		Hong Kong	$37300	6,980,412		
		Japan	$33100	127,433,494		
		Sweden	$32200	9,031,088		
		Singapore	$31400	4,553,009		

Finally, we want to know how useful you found the additional information of each countries per capita GDP, and population estimate.

How useful was information of each countries' per capita GDP ($), and population estimate, to you in estimating the true infant mortality rate (IMR) of each country?

Please indicate your evaluation of the advice by ticking one box below –

No use at all absolutely invaluable

1	2	3	4	5

That concludes our questionnaire – Many thanks for your participation!

Bibliography

Adams, P. A. and Adams, J. K . (1960). Confidence in the recognition and reproduction of words difficult to spell. *American Journal of Psychology*, 73 pp. 544-552

Agar, M. (1996). *The Professional Stranger: An Informal Introduction to Ethnography*. Academic Press.

Allwood, C.M. and Granhag, P.A. (1996). The effects of arguments on realism in confidence judgments. *Acta Psychologica, 91,* 99-119.

Alpett, M., and Raiffa, H. (1982). A progress report on the training of probability assessors. In D.Kahneman, P.Slovic and A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases* 294-305. Cambridge University Press. Cambridge, UK.

Anderson, R.E. and Jolson, M.A. (1980). Technical wording in advertising: implications for market segmentation. *Journal of Marketing*, 44, 57-66.

Aronson, E., Turner, J. and Carlsmith, M. Communicator credibility and communicator discrepancy as determinants of judgment change. *Journal of Abnormal and Social Psychology, 67,* 31-36 (1963).

Armstrong, J.S., Denniston, W.B. and Gordon, M.M. (1975). The use of the decomposition approach in making judgments. *Organizational Behavior and Human Decision Processes, 14,* 257-263.

Ayres, I. (2007). *Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be smart.* Bantam Books. NY.

Azen, R. and Budescu, D.V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8,* 129-148.

Baron, J. (2008). *Thinking and Deciding*. Cambridge University Press.

Billig, M. (1996). *Arguing and Thinking: a rhetorical approach to social psychology.* Cambridge: Cambridge University Press.

Billings, R. S. and Scherer, L. L. (1998). The effects of response mode and importance on decision-making strategies. *Organisational Behaviour and Human Decision Making Processes*. 41: 1-19.

Bonaccio, S., and Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes,* 101, 127-151.

Bonner, S.E.; Libby, R.; and Nelson, M.W., "Using Decision Aids to Improve Auditors' Conditional Probability Judgments," *The Accounting Review* (April 1996), pp. 221- 240.

Brenner, L.A., Koehler, D.J., Liberman, V., Tversky, A. (1996), "Overconfidence in probability and frequency judgments: a critical examination", *Organizational Behavior and Human Decision Processes, 65,* 212-9.

Brockriede, W. and Ehninger, D. (1960). Toulmin on argument: an interpretation and application. *Quarterly Journal of Speech, 46,* 44-53.

Browne, G.J. and Curley, S.P. (1998). Reasoning with category knowledge in probability forecasting: Typicality and perceived variability effects. In G.Wright and P.Goodwin (eds). *Forecasting with Judgment* (pp169-200). John Wiley and Sons. UK.

Budescu, D.V. and Azen, R. (2004). Beyond global measures of relative importance. Some insights from dominance analysis. *Organizational Research Methods, 7,* 341-350.

Budescu, A.V. Rantilla, A.K., Yu, H.T., Krelitz, T.M. The effcts of asymmetry among advisors on the aggregation of judgments. *Organizational Behavior and Human Decision Processes,90, (1),* 178-94 (2003).

Budescu, D.V. and Rantilla, A.K. (2000). Confidence in aggregation of expert judgments. *Acta Psychologica, 104,* 374-98.

Budescu, D.V. and Yu, Hsiu-ting Yu. (2006). Aggregation of judgments based on correlated cues and advisors. *Journal of Behavioral Decision Making, 19,* 1-25.

Burrell, G. and Morgan G. (1979). *Sociological Paradigms and Organizational Analysis*. Heinemann. London.

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

Central Intelligence Agency. (2007). Fact Book

Celen, B., Kariv, S., & Schotter, A. (2005). *An experimental test of advice and social learning.* New York: Columbia Business School.

Chaiken, S. (1980). Heuristic versus systematic processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39,* 752-766.

Chalmers, A. F. (1999). *What is this thing called science*? (3rd ed.). Indianapolis, IN: Hackett.

Checkland, P. (1981). *Systems thinking, Systems practice*. Wiley, Chichester.

Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality, 40,* 359–376.

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of Web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instrument, & Computers, 34,* 177-180.

Collins, A., and Michalski, R. (1989). The logic of plausible reasoning. A core theory. *Cognitive Science, 13,* 1-49.

Collins, A., Warnock, E.H., Aiello, N and Miller, M.L. (1975). Reasoning from incomplete knowledge, in D.Bobrow and A.Collins (Eds.). *Representation and understanding.* New York: Academic Press.

Cronbach, L.J. and Furby, L. (1970). How should we measure "change" or should we? *Psychological Bulletin, 74Psychological Bulletin, 74,* 68-80.

Cross, R. Borgatti, S.P. and Parker, A. (2001). Beyond answers: Dimensions of the advice network. *Social Networks, 23,* 215-235.

Curley, S.P., Browne, G.J., Smith, G.F. and Benson, P.G. (1995). Arguments in practical reasoning underlying constructed probability responses. *Journal of Behavioral Decision Making, 8,* 1-20.

Czerlinski, J., Goldstein, D. G., and Gigerenzer, G. (1999). "How good are simple heuristics?" In Gigerenzer, G., Todd, P. M. and the ABC Group, *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Dalkey, N.C. and Helmer, O. An experimental application of the Delphi method to the use of experts, R17-127-PR, Santa Monica, CA RAND Corp. (1963).

Dalkey, N.C. (1969). The Delphi method: An experimental study of group judgment. USAF project RAND, RM-5888-PR.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582.

Dawes, R.M., Faust, D. and Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668-1673.

Denzin, N.K. and Lincoln, Y.S. (eds.) (1994). *Handbook of Qualitative Research.* Beverly Hills, CA: Sage.

Donaldson, L. (2003). Position statement for positivism in R. Westwood and S. Clegg (eds.) *Debating Organization: Point-counterpoint in Organization Studies*. Blackwell. London.

Druckman, J.N. (2001). Using credible advice to overcome framing effects. *Journal of Law, Economics, and Organization, 17,* 62-82.

Dunn, O.J. and Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association, 64*: 366-377.

Edwards, J.R. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes, 64,* 307-324.

Edwards, D. and Potter, J. (1992) *Discursive Psychology*. London. Sage.

Feng, B and Burleson, B. (2008). The effects of argument explicitness on responses toadvice in supportive interactions. *Communication Research 35, (6),* 849-874.

Fischer, G.W. (1977). Convergent validation of decomposed multi-attribute utility assessment procedures for risky and riskless decisions. *Organizational Behavior and Human Decision Processes, 18,* 295-315.

Flugstad, A.R. and Winschitl, P.D. (2003). The influence of reasons on interpretations of probability forecasts. *Journal of Behavioral Decision Making, 16,* 107-26.

Gardner, P.H. and Berry, D.C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9,* S55-S79.

Gigerenzer, G. and Goldstein, D. G. (1996). "Reasoning the fast and frugal way: Models of bounded rationality". *Psychological Review*, 103, 650-669.

Gigerenzer, G. (2006). Bounded and Rational in R.J. Stainton (ed.), *Contemporary Debates in Cognitive Science.* Blackwell.

Gigerenzer, G., Todd, P.M. and the ABC Research Group (1999). *Simple Heuristics that Make Us smart.* New York: Oxford University Press.

Gino, F. (2008). Do we listen to advice just because we paid for it? The impact of advice cost on its use. *Organizational Behavior and Human Decision Processes, 107, (2),* 234-245.

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making, 20, (1),* 21-35.

Gleitman, H., Fridlund, A.J., and Reisberg, D. (2004). Psychology (6th Ed), New York/London.

Goldsmith, D.J. (2000). Soliciting advice: The role of sequential placement in mitigating face threat. *Communications Monographs, 67,* 1-19.

Goldstein, D.G. and Gigerenzer, G. (1998). Recognition: How to exploit a lack of knowledge. In G.Gigerenzer and P.Todd (Eds.). *Simple heuristics that make us smart.* New York: Oxford University Press.

Goodwin, P., and Wright, G. (2000). *Decision Analysis for Management Judgment* (2nd ed). John Wiley and Sons. Chichester.

Green L, and Mehr D. (1997). What alters physicians' decisions to admit to the coronary care unit? *Journal of Family Practice*, 45(3), 219-26.

Grove, W., and Meehl, P. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293-323.

Grove, W., Zald, D., Lebow, B., Snitz, B and Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19-30.

Harmon-Jones, E. (1981). *A cognitive dissonance theory perspective on persuasion,* in 'The Persuasion handbook: developments in theory and practice'. James Price Dillard and Michael Pfaus (eds), 2002.

Harries, C., Yaniv, I, and Harvey, N. (2004) Combining advice: The weight of a dissenting judgment in the consensus. *Journal of Behavioral Decision Making, 17,* 333-48.

Harte, T.B. (1973). The effects of initial attitude and evidence in persuasive communications (Doctoral dissertation, University of Illinois Campaign-Urbana 1972). *Dissertation Abstracts International, 34,* 890A.

Harvey, N. Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes, 63, (3),* 247-263 (1995).

Harvey, N. and Fischer, I. (1997). Taking advice: accepting help, improving judgment and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70 (2),* 117-133.

Harvey, N., Koehler, D.J. and Ayton, P. Judgments of decision effectiveness: actor-observor differences in overconfidence. *Organizational Behavior and Human Decision Processes, 70, (3),* 267-282 (1997).

Harvey, N., Harries, C. and Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes, 81,* 252-73.

Heath, C. and Gonzales, R. (1995). Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision-making. *Organizational Behavior and Human Decision Processes, 61,* 305-326.

Henrion, M., Fischer, G.W. and Mullin, T. (1993). Divide and conquer? Effects of decomposition on the accuarcy and calibration of subjective probability distributions. *Organizational Behavior and Human Decision Processes, 55,* 207-227.

Hirschheim R. (1985). Information Systems Epistemology: An Historical Perspective in E. Mumford, R. Hirschheim, G. Fitzgerald, T. Wood-Harper (eds.). *Research Methods in Information Systems* (IFIP 8.2 Proceedings). Amsterdam.

Hirschheim R, and Klein, H.K. (1989). Four Paradigms of Information Systems Development. *Communications of the ACM (32:10),* 1199 – 1216.

Holyoak, K. J. and Morrison, R.G. (2005). *The Cambridge Handbook of Thinking and Reasoning.* Cambridge University Press, UK.

Hotelling, H. (1931). The generalisation of Student's ratio. *Annals of Mathematical Statistics, 2,* 360-378.

Johnson, J., and Bruce, A. (2008). *Decisions: Risk and reward*. Routledge. NY.

Kahneman, D., Tversky, A., and Slovic P. (1982). *Judgement under Uncertainty: Heuristics and Biases.* Cambrdige University Press. Cambridge.

Kameda, T., and Nakanishi, D. (2003). Does social/cultural learning increase human adaptability? Rogers's question revisited. *Evolution and Human Behavior, 24*, 242-260.

Kerr, N.L. and Tindale, R.S. Group performance and decision making. *Annual Review of Psychology, 55*, 623-655 (2004).

Kleinmuntz, D.N. (1990). Decomposition and the control of error in decision analytic models. In R. Hogarth (ed). *Insights in decision making: A tribute to Hillel J. Einhorn.* Chicago: University of Chicago Press.

Kleinmuntz, D.N., Fennema, M.G. and Peecher, M.E. (1996). Conditioned assessment of subjective probabilities: Identifying the benefits of decomposition. *Organizational Behavior and Human Decision Processes, 66,* 1-16.

Khlentzos, D. (2004). *Naturalistic Realism and the Antirealist Challenge*. Cambridge, Massachusetts: The MIT Press.

Knapp, H., & Kirk, S. A. (2003). Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter? *Computers in Human Behavior, 19,* 117–134.

Koehler, D. J., and Beauregard, T. A. (2006). Illusion of confirmation from exposure to another's hypothesis. *Journal of Behavioural Decision-Making, 19*(1), 61-78.

Koriat, A., Lichenstein, S. and Fischhoff, B. (1980). Reasons for Confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6, (2),* 107-118.

Kuhn, L.M. and Sniezek, J.A. Confidence and uncertainty in judgmental forecasting: differential effects of scenario presentation. *Journal of Behavioral Decision Making, 9,* 231-247 (1996).

Kunda, Z. and Oleson, K.C. When exceptions prove the rule: how extremity of deviance determines the impact of deviant examples on stereotypes. *Journal of Personality and Social Psychology, 72,* 965-979 (1997).

Lee, A.S. (1991). Integrating Positivist and Interpretive Approaches to Organizational Research. *Organization Science 2, (4),* 342-65.

Lee, P.J. (2007). Ordinal Judgments and factors impacting accuracy, cue selection, and one-reason decision-making. Unpublished manuscript.

Lichtenstein, S., Slovic, P., Fischoff, B., Layman, M. and Coombs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 551-578.

Lichtenstein, S, Fischoff, B. and Phillips, L.D. (1982). Calibration of probabilities. The state of the art to 1980. In D. Kahneman, P. Slovic, and A. Tversky (eds.), *Judgment under uncertainty*: *Heuristics and biases* (306-334). New York: Cambridge University Press.

Lichtenstein, S. (1990). Rerieval of knowledge through algorithmic decomposition. PERCEPTRONICS INC WOODLAND HILLS CA. ADA225667

Lichtenstein, S. and Weathers, A.G. (1998). Creating algorithms as an aid to judgment. Part 2. Defense Technical Information Center OAI-PMH Repository (United States)

Linstone, H.A. and Turoff, M. (1975). *The Delphi Method: Techniques and Applications*. London. Addison-Wesley.

Luan, S., Sorkin, R. D., and Itzkowitz, J. (2004). Weighting information from outside sources: A biased process. *Journal of Behavioral Decision Making, 17,* 95-116.

Martino, J.P. (2002). A review of selected recent advances in technological forecasting. *Technological Forecasting and Social Change, 70,* 719-733.

MacGregor, D.G. (2001). 'Decomposition for judgmental forecasting and estimation' in J.S. Armstrong (ed) *Principles of Forecasting: A Handbook for researchers and practitioners.* Boston MA: Kluver Academic publishers.

MacGregor, D.G. and Armstrong, J.S. (1994). Judgmental decomposition: When does it work? *International of Forecasting, 10,* 495-506.

MacGregor, D.G. and Lichtenstein, S. (1991). Problem structuring aids for quantitative estimation. *Journal of Behavioral Decision Making, 4,* 101-116.

MacGregor, D.G., Lichtenstein, S. and Slovic, P. (1988). Structuring knowledge retrieval: An analysis of decomposed quantitative judgments. *Organizational Behavior and Human Decision Processes, 42,* 303-323.

McClelland, A. and Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-1993. In G. Wright and P. Ayton (Eds.), *Subjective probability* (pp. 453-481). Chichester: Wiley.

McCroskey, J.C. and Combs, W.H. (1969). The effects of the use of analogy on attitude change and source credibility. *Journal of Communication 19,* 333-39.

McCroskey, J.C. (1966). Toward an understanding of the importance of 'evidence' in persuasive communication. *The Pennsylvania Speech Annual 23,* 65-71.

Meehl, P. (1954). Clinical vs. Statistical prediction: A theoretical analysis and a review of the evidence. Minneapolis. University of Minnesota Press.

Meehl, P. (1993). When shall we use our heads instead of the formula? *Journal of Counselling Psychology, 4,* 81-89.

Morera, O.F. and Budescu, D.V. (1998). A psychometric analysis of the 'divide and conquer' principle in multicriteria decision making. *Organizational Behavior and Human Decision Processes,75, (3),* 187-206.

Morgan, G., & Smircich, L. (1980). The case for qualitative research. *Academy of Management Review, 5,* 491-500.

Myers, M.D, Avison D. (2002). An Introduction to Qualitative Research in Information Systems.in M.D. Myers and D. Avison, (eds) *Qualitative Research in Information Systems: a Reader*. Sage . London.

Newell, B.R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences, 9,* 11-15.

Orlikowski, W.J. & Baroudi, J.J. (1991). Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research (2),* 1-28.

Oskamp, Stuart (1965). Overconfidence in case-study judgments. *The Journal of Consulting Psychology* **2**: 261–265. reprinted in Kahneman, Daniel; Paul Slovic, Amos Tversky (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press, 287-293.

Parente, F.J., Anderson- Parente, J.K., Myers, P., O'Brien, T. (1984). An examination of factors contributing to Delphi accuracy. *Journal of Forecasting, 3,* 173-182.

Parente, F.J. and Anderson-Parente, J.K. (1987). 'Delphi Inquiry Systems' in G.Wright and P.Ayton (eds.), *Judgmental Forecasting*. Chichester, UK; Wiley.

Payne, J.W, Bettman, J.R. and Johnson, E.J. (1993). *The Adaptive Decision Maker*, Cambridge University Press. Cambridge, UK.

Parente, R.J., Hiob, T.L., Silver, R.A., Jenkins, C., Poe, M.P., Mullins, J.R. (2003). The Delphi Method, impeachment and terrorism: Accuracy of short range forecasts for volatile world events. *Technological Forecasting and Social Change, (5562),* 1-12.

Perelman, C. and Olbrecht-Tyteca, L. (1969). *The new rhetoric: a treatise on argumentation* (J. Wilkinson and P. Weaver, Trans). Notre Dame, IN: University of Notre Dame Press.

Pettit, F.A. (2002). A comparison of World-Wide Web and paper-and-pencil personality questionnaires. *Behavior Research Methods, Instruments & Computers, 34,* 50-54.

Petty, R.E., Cacioppo, J.T. and Goldman R. (1981). Personal involvement as a determinant of argument based persuasion. *Journal of Personality and Social Psychology, 41,* 847-855.

Petty, R.E., Cacioppo, J.T. and Schumann, D. (1983). Cengtral and peripheral routes to advertising effectiveness: the moderating role of involvement. *Journal of Consumer Research, 10,* 135-146.

Petty, R. E., and Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

Popper, K. (1959). *The logic of scientific discovery*. New York: Basic Books.

Psillos, S. (1999). *Scientific Realism: How Science Tracks the Truth.* London: Routledge.

Pronin, E. (2006). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences, 11, (1),* 37-43.

Reason, P., (ed). (1988), *Human Inquiry in Action: Developments in New Paradigm Research*, Sage, London.

Reinard, J.C. (1988). The empirical study of the persuasive effects of evidence: the status after fifty years of research. *Human Communication Research 15 (1),* 3-59.

Ross, M. and Fletcher, G.J.O. (1985). Attribution and Social perception. In G. Lindsay and E Aronson (Eds.), *Handbook of social psychology: Special fields and applications* (Vol. 2, 73-122). New York: Random House.

Rohrbaugh, C., and Shanteau, J. (1999). Context, process and experience: research on applied judgment and decision making in F. Durso (ed), *Handbook of Applied Cognition.* John Wiley. NY. 115-119.

Rowe, G., Wright, G. and Bolger, F. (1991). Delphi: A reevaluation of research and theory. *Technological Forecasting and Social Change, 39,* 235-251.

Rowe, G. and Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting, 12,* 73-89.

Rowe, G. and Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15,* 353-375.

Rowe, G and Wright, G. (2001). Expert judgments in Forecasting: The role of the Delphi technique' in J.S. Armstrong (ed.) *Principles of Forecasting: A Handbook for Researchers and Practitioners.* Kluver. London.

Rowe, G, Wright, G and McColl, A. (2005). Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence. *Technological Forecasting and Social Change, 72,(4),* 377-399.

Ryland, E.H. (1973). Information input and performance in small decision making groups (Doctoral dissertation, Louisiana State University, Agricultural and Mechanical College, Baton Rouge). *Dissertation Abstracts International, 33,* 4572A.

Ruscio, J. (2003). Holistic judgment in clinical practice: Utility or futility? *The Scientific Review of Mental Health Practice, 2, (1)*

Salvadori, L., Van Swol, L.M. and Sniezek, J.A. (2001). Information sampling and confidence within groups and judge advisor systems. *Communication Research, 28,* 737-771.

Savage, L.J. (1954). *The Foundations of Statistics*. New York, NY: Wiley.

Schotter, A. (2003). Decision-making with naïve advice. *American Economic Review, 93,* 196-201.

Schrah, G.E., Reeshad S. D. and Sniezek, J.A. (2006). No decision-maker is an Island: integrating expert advice with information acquisition. *Journal of Behavioral Decision Making 19, (1),* 43-60.

Scriven, M. (2000). *Logic & Methodology of Checklists*, Western Michigan University.

Simon, H. A. (1982). *Models of bounded rationality* (three volumes). Cambridge, Massachusetts: MIT Press.

Slaugher, J., and Highhouse, S. (2003). Does matching up features mess up job choice? Boundary conditions on attribute-salience effects, *Journal of Behavioral Decision Making 16,* 1–15.

Smith, T.J. (1972). The effects of truth and desirability evidence on judgments of truth and desirability of a proposition. Unpublished Master's thesis. Michigan State University. East Lansing.

Sniezek, J.A. An examination of group processes in judgmental forecasting. *International Journal of Forecasting, 5,* 171-178 (1989).

Sniezek, J.A. A comparison of techniques for judgmental forecasting by groups with common information. *Group and Organization Management, 15, (1),* 5-19 (1990).

Sniezek, J.A. (1992). Groups under uncertainty: an examination of confidence in group decision making. *Organizational Behavior and Human Decision Processes, 52,* 124-155.

Sniezek, J.A. and Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making, 4,* 263-272.

Sniezek, J.A. and Henry, R.A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43 (1),* 1-28.

Sniezek, J.A. and Henry, R.A. Revision, weighting, and commitment in consensus group judgment. *Organizational Behavior and Human Decision Processes, 45,* 66-84 (1990).

Sniezek, J.A., and Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision-making. *Organizational Behavior and Human Decision Processes, 62 (2),* 159-174.

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

Soll, B. and Larrick, R. (2009). Strategies for Revising Judgment: How (and How Well) People Use Others' Judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, (3),* 780-805.

Stanchi, K.M. (2006). The science of persuasion: an initial exploration. *Michigan State Law Review I,* 1-45.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245-251.

Steiner, I.D. (1972). *Group Process and Productivity.* New York: Academic.

Teigen, K, H. and Joergensen, M. (2005). When 90% confidence intervals are 50% certain: on the credibility of credible intervals. *Applied Cognitive Psychology*, *19* 455–475.

Toulmin, S. (1958). *The Uses of Argument.* Cambridge. Cambridge University Press.

Trafimow, D. and Sniezek, J.A. Perceived expertise and its effect on confidence. *Organizational Behavior and Human Decision Processes, 57,* 290-302 (1994).

Truell, A. D., Bartlett, J. E., II, & Alexander, M. W. (2002). Response rate, speed, and completeness: A comparison of Internet-based and mail surveys. *Behavior Research Methods, Instruments & Computers*, *34*(1), 46-49.

Turner, C.F., Ku, L., Rogers, S.M., Lindberg, L.D., Pleck, J.H., Sonenstein, F.L. (1998). Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science 280 (5365),* 867–73.

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

Van Swol, L.M. and Sniezek, J.A. (2001). Trust, confidence, and expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes, 84, (2),* 288-307.

Von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ. Princeton University Press.

Warfield, J.N. (1990). *A science of generic design:. Managing complexity through systems design.* Intersystems Publications, Salinas.

Weber, E., and Johnson E. (2006). Mindful judgment and decision making. *Annual Review of Psychology, 60,* 53-85.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Williams, E.J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society (Series B), 21,* 396-399.

Wright G. and Ayton, P. (1987). The Psychology of Forecasting' in G. Wright and P. Ayton (eds.) *Judgmental Forecasting*. John Wiley and Sons Ltd.

Yalch, R. F., and Elmore-Yalch, R. (1984). The effect of numbers on the route to persuasion. *Journal of Consumer Research, 11,* 522-527.

Yaniv, I. (2004b). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93,* 1-13.

Yaniv, I. (2004b). The benefit of additional judgments. *Current Directions in Psychological Science, 13,* 76-79.

Yaniv, I., and Foster, D.P. (1995). Graininess of judgment under uncertainty: An accuracyinformativeness tradeoff. *Journal of Experimental Psychology*: *General*, 124, 424-432.

Yaniv, I., and Foster, D.P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, 10, 21-32.

Yaniv, I. and Kleinberger, E. (2000). Advice taking in decision-making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83 (2),* 280-281.

.