2009

Probability Elicitation Methods for Avoiding Biases: An Exposition

Bethany Mihajlovits
Virginia Commonwealth University

College of Humanities and Sciences
Virginia Commonwealth University

This is to certify that the thesis prepared by Bethany H. Mihajlovits entitled PROBABILITY ELICITATION METHODS FOR AVOIDING BIASES: AN EXPOSITION has been approved by his or her committee as satisfactory completion of the thesis requirement for the degree of Masters of Science in Operations Research

Dr. Jason R. W. Merrick, Department of Statistical Sciences and Operations Research

Dr. Edward L. Boone, Department of Statistical Sciences and Operations Research

Dr. Linda E. Zyzniewski, Department of Psychology

Dr. D'Arcy P. Mays, Chair of Department of Statistical Sciences and Operations Research

Dr. Fred M. Hawkridge, Dean of the College of Humanities and Sciences

Dr. F. Douglas Boudinot, Dean of the Graduate School

30 April 2009

PROBABILITY ELICITATION METHODS FOR AVOIDING BIASES: AN EXPOSITION

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science at Virginia Commonwealth University.

by

BETHANY H. MIHAJLOVITS
Bachelor of Science, Virginia Commonwealth University, 2003

Director: DR. JASON R.W. MERRICK
GRADUATE PROGRAM DIRECTOR, DEPARTMENT OF STATISTICAL SCIENCES AND OPERATIONS RESEARCH

Virginia Commonwealth University
Richmond, Virginia
May 2009

Acknowledgement

Table of Contents

List of Tables

Page

List of Figures

Page

Abstract

PROBABILITY ELICITATION METHODS FOR AVOIDING BIASES: AN EXPOSITION

By Bethany H. Mihajlovits, B.S.

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2009

Major Director: Dr. Jason R.W. Merrick
Director of Graduate Planning, Department of Statistical Sciences and Operations Research

A large portion of decision analysis lies in a decision maker's uncertainty about an outcome and what they perceive is the chance (probability) of that outcome occurring (in other words, an individual's "degree of belief" that an outcome will occur). However, thinking probabilistically can be difficult and we rely on "rather primitive cognitive techniques to make" such assessments (these techniques are termed heuristics) (Clemen & Reilly, 2001 p.311). Heuristics are simple and intuitive but tend to result in probabilities that are biased. This thesis will connect the literature available from both the psychology

behind the biases and the mathematical problems associated with the probability elicitation itself. Additionally, this thesis will present a better understanding of the biases that distort the probability elicitation for the decision maker along with suggestions for improving such assessments.

CHAPTER 1 Introduction

The overall process of arriving at an informed decision, that is at present a best value, is often a long and tedious process that can consist of multiple quantitative and qualitative values, varying degrees of uncertainty associated with the likelihood of events, and differing opinions from various people concerning what is most important about the decision at hand. Decision analysis exists to assist a decision maker in choosing the best option available with decisions that exhibit such difficulty. This thesis will focus on the uncertainty aspect of decision analysis and specifically the biases that are associated with probability assessment, methods used for eliciting uncertainty from an individual with knowledge of a specific field, and ways to improve the assessments provided.

Probability elicitation is not a strictly academic situation. Many documented instances exist where decision makers have had to use probabilities assessed by experts to make a proper decision. Chapter two reviews some of these real life decisions that were made with decision analysis that incorporate the use of probability assessments from subject matter experts.

There are a wide variety of biases that may be brought about with the probability assessments provided by a subject matter expert. These biases can be produced purposefully for motivational reasons or as a simple cognitive misjudgment. In general, humans have developed two separate means of drawing conclusions from data: a reasoning

ability for processing mental representations, and a judgmental ability for characterizing stimuli along key dimensions (Benson *et al*, 1995). Wallsten and Budescu (1983) postulate that when a subject is asked to provide a probability of an event's occurrence, the individual will search their memory for applicable knowledge, combine that knowledge with the information at hand, and attempt to provide the best feasible judgment. Therefore, the judgment provided depends on the subject's memory, the current information aspects, and also the order in which the subject generally processes information. Gigerenzer *et al* (1991) presents literature suggesting that memory is "excellent in storing frequency information from various environments and the registering of event occurrences for frequency judgments is fairly automatic cognitive process requiring very little attention or conscious effort" (p.510). This series of cognitive processes that occur without effort in the subject leaves much room for error of interpretation of uncertainty and that error can be presented as a biased assessment. This thesis will review several biases including some that are associated with the cognitive processes termed heuristics. "Heuristics are essentially informal arguments employed by subjects in bringing evidence to bear upon the propositions in question" (Benson *et al*, 1995 p.1641).

A main goal of probability elicitation is to avoid any biases (or adjust those that surface) that will present themselves due either to heuristics or motivational purposes. Thus, proper probability elicitation methods are reviewed in chapter four that will hopefully minimize any biases that could present themselves and create havoc on the uncertainty aspect of the problem itself. Training the experts to recognize biases associated with cognitive processes is a preferred method for avoiding them entirely. For the few

biases not avoided with a proper elicitation procedure, chapter five reviews methodology for improving expert probability assessments. These methods may also be used during the elicitation process and they work not only to eliminate biases but also to improve the probability assessments so that a more accurate and better decision can be made. Often times, a motivational bias can work its way into the subject's psyche presenting inaccurate assessments that can be damaging to the integrity of the decision. The final chapter, Scoring Rules, reviews how scoring rules can be used to encourage honesty in forecasting.

CHAPTER 2 Applications of Probability Elicitation

There are many major decisions that are made based on probability assessments which are supplied from subject matter experts. There are a variety of ways to elicit these expert judgments and many fields of study have already performed these procedures and analyzed subsequent expert probability assessments to arrive at better decisions. Five studies will be covered that had to perform an elicitation procedure to extract probabilities or probability distributions from experts in a given field. North and Stengel (1982) brought to the surface a methodology for major decisions in the development of magnetic fusion energy. Keeney, Sarin, and Winkler (1984) developed a risk assessment model to relate adverse health effects to alternative carbon monoxide standards. Winkler and Poses (1993) presented a probability analysis on the survivability of intensive care unit patients. DeWispelare, Herren, and Clemen (1995) were concerned with acquiring accurate probabilistic judgments for future conditions at a high-level nuclear waste repository. Lastly, a decision analysis model, incorporating risk aversion and probabilistic dependence, for bidding oil and gas lease sales was provided by Donald Keefer (1991).

The U.S. magnetic fusion program was looking for a way to provide an inexhaustible energy supply within the next century (North & Stengel, 1982). North and Stengel (1982) present the problem of deciding what facility to fund for the U.S. magnetic fusion energy program. This problem is plagued by uncertain strategic outcomes, vague

available alternatives, and successes that are not measured in simple terms. Because this is a long-term program, it naturally involves multiple decisions and to keep the analysis cost to a minimum, only a critical decision (and a subset of crucial decisions) were processed. The set of critical and crucial decisions encompass the building timeframe and the nature of the major test facilities. The test facility decision comes with a considerable number of uncertainties to include commercial reactor performance, environmental concerns, time considerations, and commercial market acceptance. To maintain a workable model some assumptions involving performance, cost and feasibility were made.

In the North and Stengel (1982) study, the following probabilities needed to be assessed in order to complete the decision tree: probability of commercialization of the leading fusion concept; probability of distribution over five-year increments; probability of distribution on the status of competing electric energy technologies; and the probability distribution on magnetic fusion reactor performance for each concept. The type of analysis that was being performed required a probability value, any and all of the above mentioned, be assigned to each path option in their influence diagram structure. North and Stengel were aware of the conscious and unconscious biases that can surface with the probability elicitation process. They referenced the work of Tversky and Kahneman (1974) when developing their interview process. The probability assessments were created in two interviews with staff members knowledgeable in fusion research and program planning. Each interview, meant to develop a complete set of probability inputs for the decision tree analysis, was approximately three hours in length and incorporated a process that helped avoid biases.

A final optimal result was not reached due to the usage of illustrative data that was not based on a thorough consultation with energy experts. However, several tendencies in the outcome suggested that nominal development of the systems was slightly better than the accelerated option and the delay option. North and Stengel (1982) conclude that the systematic assessment of performance, cost, and time of commercial availability, built up from engineering judgment, is the strongpoint of the methodology.

Keeney, Sarin, and Winkler (1984) offered an analysis that addressed the "complexities inherent to ambient air quality standard-setting problems" by developing a risk assessment model that relates adverse health effects to alternative carbon monoxide standards (p.518). Keeney *et al* implement an individual interviewing process for multiple scientists to avoid any cognitive biases involving dependence.

Three major and five minor health effects were identified as resultants of CO exposure in the Keeney *et al* study. The overall problem of determining the number of people that would suffer from the health effects (and the magnitude of suffering) based on each alternative CO standard is too complex a problem to solve without decomposition. Because experimentation was not a viable approach due to ethical consideration, professional judgments on estimates of health risks from CO exposure were obtained by interviewing numerous medical and health experts. A representative sampling of experts were individually interviewed on multiple occasions, either in person or over the phone, with time lengths ranging from 30 minutes to two hours. Each expert was interviewed individually to attempt to preserve independence of opinion. In each interview, the experts were asked for a qualitative identification of health effects and its corresponding probability. The same information was elicited from different perspectives to allow for any inconsistencies and redundancy of information. The fractile

method and direct probability assessments were used to quantify the probability distributions. The fractile method is an assessment procedure where a value (e.g. '*x*') is adjusted until an indifference point between lotteries is reached by the subject for a particular percentile (e.g. 0.35). Thus the value (*x*) at that indifference point is the fractile of the distribution (i.e. *x* is the 0.35 fractile of the distribution) (Clemen & Reilly, 2001). It was noted that like most assessment procedures, the experts were reluctant at first to quantify their judgments but became more willing throughout the course of the process. Overall, the responses revealed substantial stability and consistency.

In an attempt to estimate health risks associated with variable CO levels, the methodology in the study by Keeney *et al* was applied to four cities with varying magnitudes of risk, response estimates, and exposure rates. The relative values across the standards did not vary as greatly suggesting that "the relative ordering of standards in terms of risk estimates may not change with the refinements in exposure data" (Keeney *et al*, 1984 p.527). Lastly, although the individual interviews provided good estimates, an expert group meeting could have proven beneficial.

Winkler and Poses (1993) surmise that a physician assessed probability of survival can be of interest to the patient or another physician making a prognosis or decision regarding treatment. Their article analyzes "data from an observational study in which physicians with different levels of experience in critical care medicine and different familiarity with the patient assess the probability that a patient admitted to an intensive care unit in a teaching hospital would survive until discharge from the hospital" (Winkler & Poses, 1993). The intent of the paper is to present an extensive analysis of both assessed probabilities and combination

probabilities by evaluating the ability of the physicians to assess probabilities of survival of ICU patients.

Quantitative probability assessments of the patient's survival were obtained as soon as possible (usually within an hour) after admittance from four physicians: an intern (the least experienced), the critical care fellow (three years post MD training), the critical care attending physician (five years post MD training), and the primary physician. No training or feedback was provided to the physicians except the critical care attending physicians. The critical care attendings were the only group to know of the study's purpose and thus maintain awareness in probabilistic judgment. Each physician saw a differing number of patients because the study involved actual patients admitted to an ICU, and ICU staffing patterns are variable. An overall evaluation was done in addition to looking at a number of individual quality measures including the distributions of probabilities and outcomes, summary measure of such distributions, and decompositions of scoring rules. A quadratic scoring rule called the Brier score was used for the overall evaluation as it integrates discrimination and calibration and is influenced by the base rate (scoring rules will be defined in chapter 6). As for the distributions of probability assessments, the primaries gave extremely high probabilities much more often than the others (with a relatively low number in the middle sections), the critical care attendings seldom used extremely high probabilities but used the middle sections much more often, and the interns and critical care fellows tended to be in between the two extremes. Lastly, the primaries were particularly good at identifying patients who would survive. A regression was run to determine if a model using the probabilities could improve upon the physician's performance. The outcome provided hardly any change in the Brier scores,

suggesting that any patterns in the data that might be capitalized on to improve assessment were very minor.

The general concept of getting additional information is to get to a better decision. Additionally, obtaining multiple forecasts and combining them leads to improved forecasting performance. The best combination was the primary physician and the critical care attending. In general, "the worst combination of two probabilities does better than the average of the four individual Brier scores, and the combination of all four probabilities does better than the best individual Brier score. Thus, obtaining additional opinions not only leads to improved performance on average, but results in reductions in the risk of poor performance" (Winkler & Poses, 1993 p.1534). These results were based on simple averages of probabilities.

The analysis showed that the physicians most familiar with the patient and most familiar with intensive care were the best performers, individually and in combination. Some biases could have affected the assessments including ego bias which could lead to higher assessments for a desirable outcome (like survival), hanging crepe bias where lower assessments are given so not to raise hopes, and uncertainty bias where the physicians are inclined to express their uncertainty by using less extreme probabilities. Winkler and Poses recommend that some improvements in assessment might be reached by providing training on probability assessment, providing feedback on performance, having the physicians assess probabilities on a regular basis for experience, and having a group probability be presented after a collaboration of physicians.

The performances of high-level nuclear waste geologic repositories have to be assessed for a 10000 year regulatory period as stipulated by the Environmental Protection Agency (EPA) (DeWispelare *et al*, 1995). Specifically, the study presented by

DeWispelare, Herren, and Clemen (1995) was concerned with acquiring probabilistic judgment through expert elicitation of future climate parameters in the Yucca Mountain Nevada vicinity (YMNV).

The formal elicitation procedure used for this DeWispelare *et al* (1995) study consisted of eleven steps: determine the objectives and goals of the study; recruit the experts; identify issues and information needs; provide initial data to the experts; conduct the elicitation training session; discuss and refine the issues; provide a multi-week study period; conduct the elicitations; provide post-elicitation feedback to the experts; aggregate the experts' judgments (if required); and document the process. The eleven steps were then broken into five strictly sequential phases: pre-training activities (steps one through four); training (step five); study period (steps six through seven); elicitation (step eight); and follow-up (steps nine through eleven). The intent of the training process was to provide the experts with knowledge of potential biases that can be encountered during the elicitation process. The discussion process should be used by the experts to refine the initial conceptual scenarios to arrive at an unambiguous definition of the events to be elicited and result in a convenient model that they can proceed with into individual research efforts. Individual elicitations were conducted by having each expert prepare a document that provided the basis of their judgments, bringing them together into a common meeting and presenting each paper to the group for clarification and reasoning of assessments, and then interviewing each individual afterwards to obtain probabilities. The experts then reviewed the results to ensure that they accurately reflected the information provided in the elicitation.

As for the aggregation of expert judgments, "aggregating judgments across a group of experts provides a distribution that attempts to summarize the state of knowledge across the group" (DeWispelare *et al*, 1995 p.18). The two ways of aggregating the judgments of experts are behavioral, when the experts themselves produce a combined or consensus view or mechanical, when a central analyst is responsible for the aggregation by applying an aggregation formula.

DeWispelare *et al* make some concluding remarks of the documentation needs and general comments on the study including information from the experts themselves. "Expert judgments, elicited through a formal expert elicitation process, will be scrutable and the rationales used to reach the judgments will be documented" (DeWispelare *et al*, 1995 p.21). The experts felt that the training session to make aware the potential biases involved was essential, and that they had little difficulty in representing their judgments as probability distributions. The consensus distributions that were generated behaviorally were not necessarily similar to the ones generated from mechanical combination techniques. It is the responsibility of the decision maker to determine if and how to combine the individual distributions. Finally, the mechanical aggregation of the individual expert judgments is an efficient, feasible, and appropriate way to satisfy the needs of modelers (DeWispelare *et al*, 1995).

"Bidding for offshore U.S. oil and gas leases is a major corporate resource allocation problem involving enormous uncertainties and very high stakes" (Keefer, 1991 p.377). Donald Keefer (1991) developed operationally useful decision analysis models that incorporate risk aversion and probabilistic dependence among the lease values for bidding at oil and gas

lease sales in his article. This paper considers three expected utility models that use exponential utility functions and include both bids and partnership shares as potential decision variables. Probability elicitation is needed for this study to determine the certainty equivalents of the increasing levels of risk aversion for each of the three models presented.

Thanks to the documentation of several decision studies that were performed, we can see that probability elicitation from experts is indeed a widely used method for combating uncertainty in a problem. Unfortunately, there are many biases that are usually associated with the probability elicitation process. The main biases that are present in most elicitation procedures are discussed in the following chapter.

CHAPTER 3 The Problem: Heuristics and Biases

As previously noted, people have a tendency to incorporate both conscious and unconscious biases into their predictions of uncertainty. A valid reason for this is that we, as humans, tend to use rather "primitive cognitive techniques" termed heuristics to assess uncertainty (Clemen & Reilly, 2001). Clemen and Reilly (2001) stated that "heuristics are easy and intuitive ways to deal with uncertain situations, but they tend to result in probability assessments that are biased in different ways depending on the heuristic used" (p.311). Tversky and Kahneman (1974) define three heuristics that are used to assess probabilities and their associated biases. The three heuristics are representativeness, availability, and anchoring and adjusting. There are additional biases that have been observed and are not specific to the proposed heuristics. There are a lot of instances where these heuristics and biases have been observed.

The representative heuristic is assessing a probability that someone or something belongs to a specific category (Clemen & Reilly, 2001). More generally it is the method in which probabilities are evaluated "by the degree to which A is representative of B, that is, the degree to which A resembles B" (Tversky & Kahneman, 1974 p.1124). This heuristic involves making the judgment by comparing the information provided on the subject with a stereotypical member of the category. The better the resemblance between the two, the higher the probability is that the subject will be placed into the category. There are a

multitude of biases that result from the representative heuristic: base rate bias, prior probability bias, insensitivity to sample size, incorrect interpretation of randomness, insensitivity to predictability, illusion of validity, and failure to regress.

The prior probability, or base-rate frequency, of an outcome is often neglected when people evaluate probability with representativeness. The base-rate should have an effect on the probability assessment but is often overlooked. To illustrate this concept consider the example of Tom W from Kahneman and Tversky (1973), whereby a description of Tom is provided that implies he is introverted, mechanically minded, lacking in creativity, and well organized. This description is given to a group of people so that they may assess the likelihood that Tom is a graduate student in various academic fields. Consequently, the assessors allowed the stereotyped properties of Tom W's description to command their decision versus incorporating the population statistics for each field of study. Many such studies were performed by Tversky and Kahneman (1974); some with stereotyping descriptions, some with no descriptions, and still some with non-specific descriptions. The overall response from these studies was that when no descriptions were given, prior probabilities were utilized correctly, when descriptions were introduced, prior probabilities were ignored.

The next bias caused by the representative heuristic is insensitivity to sample size, where people assess the likelihood of a sample result by the resemblance of the result to the corresponding population parameter. Tversky and Kahneman (1974) perform a study where a small hospital and a large hospital each recorded the number of times that the daily percentage of boys born reached 60% or greater (versus the norm of 50%). Students were

then asked which hospital would have recorded more days of 60% male born population or if the two hospitals were approximately equal in results. Most assessors believed the two hospitals should have been approximately equal because they are representative of the general population. These assessments were obviously made without taking into account the theory that the smaller hospital would have a greater chance at the 60% because a large sample is less likely to stray from the mean (50% male population born). Lastly, the judgment of posterior probabilities, the probability that a sample was chosen from one population versus another, has also accounted for such insensitivity. In this case the assessments are dominated by the sample proportion and thus unaffected by the sample size.

An incorrect interpretation of randomness, where one expects that the condensed sequence of events of a random process is representative of the actual process, is another bias produced by the representative heuristic (Tversky & Kahneman, 1974). A good example is when people view the sequence H-T-H-H-T-H of a flip of a coin as more likely than H-H-H-T-T-T because it appears to be more random, even though both are just as likely. This also includes the gambler's fallacy; when a gambler believes that a black is due after a long run of red on the roulette wheel because a black would start to restore the equilibrium of the process. Tversky and Kahneman (1974) note that, "deviations are not 'corrected' as a chance process unfolds, they are merely diluted" (p.1125).

Insensitivity to predictability occurs when assessors make predictions (e.g., the outcome of game or future value of stock) based on descriptions and their favorability instead of evidence and the expected accuracy of prediction (Tversky & Kahneman, 1974).

It should be that if no description is available or relevant, predictions should mirror one another and furthermore, the higher the predictability, the wider the range of values. However, throughout the course of many studies, subjects have shown little regard for the considerations of predictability.

"The unwarranted confidence which is produced by a good fit between the predicted outcome and the input information may be called the illusion of validity" (Tversky & Kahneman, 1974 p.1126). It will endure even when the assessor is aware of his/her own prediction accuracy limiting factors. An example provided by Tversky and Kahneman (1974) states that assessors are more confident in predicting the final grade point average of a student with first year grades consisting solely of B's versus that of a student with A's and C's because people have greater confidence in predictions based on redundancy. Nevertheless, a prediction based on several inputs is able to achieve higher accuracy when independence is prevalent instead of correlation or redundancy. Thus, even though redundancy increases confidence, it decreases accuracy, and the resulting judgments are potentially confidently bad.

The last bias associated with the representative heuristic is the concept of failure to regress. There is a general phenomenon called regression toward the mean which is the notion that for random events, extreme cases will tend to be followed by less extreme cases (Clemen & Reilly, 2001). Despite the fact that this notion has been around for more than 100 years, people generally do not develop accurate intuitions concerning it. People do not expect to see it in many situations where it is likely to happen and even when the occurrence is recognized, they come up with explanations for it (Tversky & Kahneman,

1974). Tversky and Kahneman (1974) reference a study concerning a flight training instructor in which he praised a student pilot for a smooth landing, whose subsequent landing was considerably worse, and criticized a different student pilot for a rough landing, whose subsequent landing was considerably better. This led the instructor to conclude that verbal praise was detrimental and verbal criticism was beneficial. The instructor failed to recognize that the student pilot who performed badly was probably performing below his average level and would likely perform closer to his mean next time around, no matter the consequences involved. The same goes for the student pilot who executed a tremendous landing; he would probably perform closer to his average (which would be a worse landing) for the subsequent flight without regard to consequence or praise. Thus, the instructor failed to recognize that the students were regressing to the mean and actually overestimated the effect of punishment and underestimated the effect of praise.

The availability heuristic is when a probability of an event's occurrence is assessed based on the ease with which we can retrieve similar events from memory (Clemen & Reilly, 2001). It is often easier to recall instances of large classes than those of less frequent classes. External events and influences can have a substantial impact the availability of similar incidents. The biases associated with the availability heuristic are: overestimation when events are more easily remembered, underestimation when events are harder to recall, effectiveness of a search set bias, imaginability bias, illusory correlation, and hindsight bias.

A bias is created when the size of class is assessed by the ability to retrieve its instances thus; a class with easily retrievable instances will appear more numerous than

that of class whose instances are less retrievable (Tversky & Kahneman, 1974). Familiarity is not the only factor for this bias, salience also affects retrievability. Tversky and Kahneman (1974) note this with the idea that seeing a burning house is a greater impact for assessment than just reading about a house fire. Also recent activity is more retrievable than earlier activity, therefore the probability judgment concerning automobile accidents will rise temporarily when a person witnesses an overturned vehicle.

The availability heuristic also provides us with biases due to the effectiveness of a search set and different tasks elicit different search sets (Tversky & Kahneman, 1974). The example provided by Tversky and Kahneman (1974) is the supposition that an individual samples a random three letter word from the dictionary. The ensuing question is "is it more likely that the word starts with r or that r is the third letter?" People judge the word to be more likely to begin with r because it is easier to search for words by their first letter despite the fact that consonants, such as r, are more frequent in the third position versus the first.

When a frequency assessment is necessary for a class that is not stored in memory, but can be generated by a rule, a bias of imaginability can occur (Tversky & Kahneman, 1974). Typically for this situation the assessor will think up multiple instances and assess the frequency based on the ease of the instance construction. The ease of construction does not always reflect the accurate frequency. Imaginability has an important role in the probability assessment of real life situations. The example provided by Tversky and Kahneman (1974) associates the risk involved with an adventurous expedition is evaluated by imagining the things that can go wrong and can't be dealt with. If a whole lot of these

instances can be thought up, the trip can seem overly dangerous even if their actual likelihood of occurrence is very small. On the other hand, the risk involved may be underestimated if some dangers are not imagined.

The illusory correlation bias is a disconnect in the judgment of the frequency of the co-occurrence of two events (Tversky & Kahneman, 1974). Tversky and Kahneman (1974) presented the draw-a-person test. It was a situation where subjects were given a diagnosis of a mental patient and a portrait drawing made by the patient. The subjects then estimated the frequency of a particular diagnosis accompanied by distinguishing facial features on the drawing. The subjects overestimated the co-occurrence frequency of natural associates, such as peculiar eyes and suspiciousness, because suspiciousness is more readily associated with the eyes than any other body part. When an associative bond is strong between two events, they are assumed to occur together frequently.

The hindsight bias is when a subject misremembers the degree to which he previously forecasted an event's occurrence, and generally the forecast is remembered to be more accurate than it was (Mellers & Locke, 2007). One reason for this bias is that most people are better able to recall information that was consistent with the outcome versus information that was inconsistent. Mellers and Locke (2007) review a study performed by Fischoff and Beyth (1975) involved students that were initially asked to predict what would happen in 1972 when Nixon visited China and then later asked to recall their probability assessments. Majority of the students recalled themselves being more accurate at the two week interval (67%) and most all students recalled better accuracy at the three to sixth month interval (84%).

An individual will often assess a probability by selecting an initial anchor point (suggested either by formulation of the problem or by result of partial computation) and then adjust the anchor based on their knowledge of the specific event (Clemen & Reilly, 2001). Tversky and Kahneman (1974) term this heuristic anchoring and adjusting. The adjustments are usually inadequate since different estimates are obtained because they yield to the different starting positions. The anchoring and adjusting heuristic also has multiple biases: insufficient adjustment, biases in the evaluation of conjunctive and disjunctive events, anchoring in the assessment of subjective probability distributions, overconfidence, and even conservatism.

The insufficient adjustment bias typifies the anchoring heuristic in general (Tversky & Kahneman, 1974). Anchoring occurs regardless of whether the initial value was provided to the subject or they arrived at it from an incomplete computation. Tversky and Kahneman (1974) illustrate this with a study done with high school students were they were provided a numerical expression (one ascending (1x2x3x4x5x6x7x8) or one descending (8x7x6x5x4x3x2x1)) and only five seconds to estimate the product. The median estimate for the ascending sequence was considerably smaller than the estimate for the descending sequence because the first few steps of multiplication for the descending sequence was higher than the ascending sequence.

Studies of probability judgments suggest that people are inclined to overestimate the probabilities of conjunctive events and underestimate the probability of disjunctive events (Tversky & Kahneman, 1974). This is the bias in the evaluation of conjunctive and disjunctive events. Tversky and Kahneman (1974) state that planning events prove a

significant bias since they are typically conjunctive because they are comprised of a series of events that must occur to accomplish completion. Even with each event having a high likelihood, the probability of success can actually be very small if there are many events to consider. Thus, there is general tendency of unjustifiable optimism due to the overestimation of the likelihood of a conjunctive event. Conversely, people tend to underestimate the likelihood of failure in complex systems.

Another bias of anchoring and adjusting is anchoring in the assessment of subjective probability distributions (Tversky & Kahneman, 1974). A probability distribution is often constructed by asking the assessor to evaluate specific percentiles then producing the distribution curve with the points provided. Once several different distributions are collected, the judge can then be tested for proper calibration. Liberman and Tversky (1993) define proper calibration as when a decision maker has his/her judgments match the corresponding frequencies. This is different than being resolute; when decision maker can discriminate between events that do and do not occur. Over the course of many probability elicitations the distributions provided by the experts indicated "large and systematic departures from proper calibration" (Tversky & Kahneman, 1974). This means that the subjects provided overly narrow confidence intervals that suggest their certainty does not coincide with their knowledge base of the assessed quantities. The elicitation procedure also affects the degree of calibration. Calibration will be further reviewed in chapter 5.

The overconfidence bias is associated with the hard easy effect and is created when people make the mistake of overestimating their knowledge (Clemen & Reilly, 2001).

Gigerenzer, Hoffrage and Kleinbölting (1991) provided the description that the overconfidence effect occurs when the confidence judgments are larger than the relative frequencies of the correct answers. The hard-easy effect occurs when the degree of overconfidence increases with the difficulty of the questions, where the difficulty is measure by the percentage of correct answers. Continuously, confidence in one's immediate and spontaneous knowledge (rather than long term reflection) is largely determined by the structure of the task and the structure of a corresponding, known environment in a person's long-term memory. Gigerenzer *et al* (1991) provide a framework, probabilistic mental model, which explains the overconfidence effect and predicts the conditions under which the effect appears. This framework is detailed below. Brenner, Koehler, and Liberman (1996) note that the selection of difficult questions leads to "spurious" overconfidence, conversely the deliberate exclusion of difficult questions is likely to produce underconfidence. Through studies and research Brenner *et al* determine that the major (not sole) reason for overconfidence is question difficulty and not how the questions are selected (i.e. random vs. nonrandom selection). Liberman and Tversky (1993) distinguish between two types of overconfidence; specific, referring to the overestimation of the probability of a specific event, and generic, referring to the overestimation of the probability of the event that the subject considers most likely. Lastly, Mellers and Locke (2007) correlate overconfidence to the above average effect, "closely related is the above average effect when people perceive themselves as being better than others on most desirable attributes, including honesty, cooperativeness, etc."

Keren (1997) suggests the idea that overconfidence and the hard-easy effect may be due to regression towards the mean. Keren (1997) notes that even though regression towards the mean may induce overconfidence, the concept does not rule out genuine overconfidence. Additionally, even unbiased estimates are subject to random error and the larger the random error the larger the regression effect. As for genuine overconfidence, Keren (1997) reviews a study that he previously performed with bridge players that suggests that at least part of the exhibited overconfidence "was genuine and caused by cognitive or motivational factors" (p.273). Some of these factors may include a subject who tends to think more in terms of positive reasons and consequently affirmative evidence appears larger than disconfirming evidence; that overestimation of our intellectual abilities is actually deeply rooted in humans; a social norm that encourages overconfidence because knowledge is prized and ignorance is appalling; and/or a self motivation mechanism (e.g. a soccer fan believes his favorite team will win even if losing at the midpoint of the game).

Another bias from the anchoring and adjusting heuristic is the conservatism bias suggested by Wallsten and Budescu (1983). Conservatism is a bias towards the prior probabilities when a subject is in a position to revise their posterior probabilities, thus high probabilities are underestimated and low values are overestimated. There is confusion however, concerning whether this bias is formed from subjects who are accurately reporting their misguided opinions or if it is from subjects who are biased in translating their accurate judgments to probabilities. In addition, conservatism appears to be situation

dependent and is not necessarily specific to the subject's behavior (Wallsten & Budescu, 1983).

Mellers and Locke (2007) also provided some insight into framing effects and their bias towards judgment. When people consider different choices they often accept and utilize the information as it was received leading to systematic differences in preference known as framing effects. Behavioral decision researchers discovered that when the same option is offered but with different frames, people will repeatedly reverse their preferences. A good example is providing decision makers with a decision concerning an unknown Asian disease expected to kill 600 people presented by Tversky and Kahneman (1981). The two frames present the same situation differently: the gain frame has a program A in which 200 people will be saved and a program B where there is a 1/3 chance that 600 people will be saved and 2/3 chance that no-one will, and the loss frame has a program A in which 400 people will die and a program B where there is a 1/3 that no one will die and a 2/3 chance that 600 people will die. Despite that fact that the situation for each program was the same in both frames, the majority of people preferred the gain frame. Mellers and Locke (2007) stated it best with "frames are subtle, yet consequential" (p.9).

Similar to framing effects is the idea that stimuli contexts, either local or global, can alter a response from a subject (Mellers & Locke, 2007). The explanation of this idea is that responses to the same event will differ based on the surrounding stimuli. The example provided by Mellers and Locke is that a ten pound sack of apples will appear light in weight after carrying a fifty pound child but will appear heavy in weight after carrying a can of soda.

Another noteworthy bias, discussed by Spetzler and Von Holstein (1975), is the motivational bias. A motivational bias is when a subject adjusts an assessment based on the perception of a personal reward for various responses. Thus, the subject may be influencing the decision in his/her favor by giving a specific set of assessments. The motivational bias can show up either consciously, a salesman predicting low sales so that his numbers are greater than forecasted, or subconsciously, a doctor who gives a higher likelihood of survival because she is hoping that a preferred patient will survive. Related to the motivational bias is the ego bias, suggested by Winkler and Poses (1993), which is an illusion of control that can lead to higher assessed probabilities of desired outcomes like patient survival. Winkler and Poses (1993) also defined the "hanging crepe" bias, whereby a physician may provide a lower probability of survival so as to not raise any hopes and furthermore attribute any survival to the skill of the physician.

Gigerenzer *et al* (1991) claim that even though the cognitive bias research supports that people naturally make mistakes in reasoning and memory, including overestimating their knowledge, that people are in fact decent judges of the reliability of their knowledge provided that it is representative of a specific reference class. Basically, the overconfidence bias results as a consequence of selection instead of some deficient cognitive heuristic. They present a framework that helps to explain a broad range of the experimental findings concerning confidence, including inconsistencies. In addition, the theory for which this framework is based deals with spontaneous confidence or an immediate reaction instead of a product of long reflection.

The framework presented by Gigerenzer *et al* (1991), exhibited by a flow chart in figure 3.1, supports two strategies when presented with a two-alternative confidence task. The first is for the subject to attempt to construct a local mental model (local MM) which is a solution created by memory and elementary logic. The second strategy is to create a probabilistic mental model (PMM) using probabilistic information from a natural environment if the local MM fails. Figure 3.1, provided by Gigerenzer *et al* (1991, p.508), are the cognitive processes in solving a two-alternative general-knowledge task with MM representing a mental model and PMM representing a probabilistic mental model.

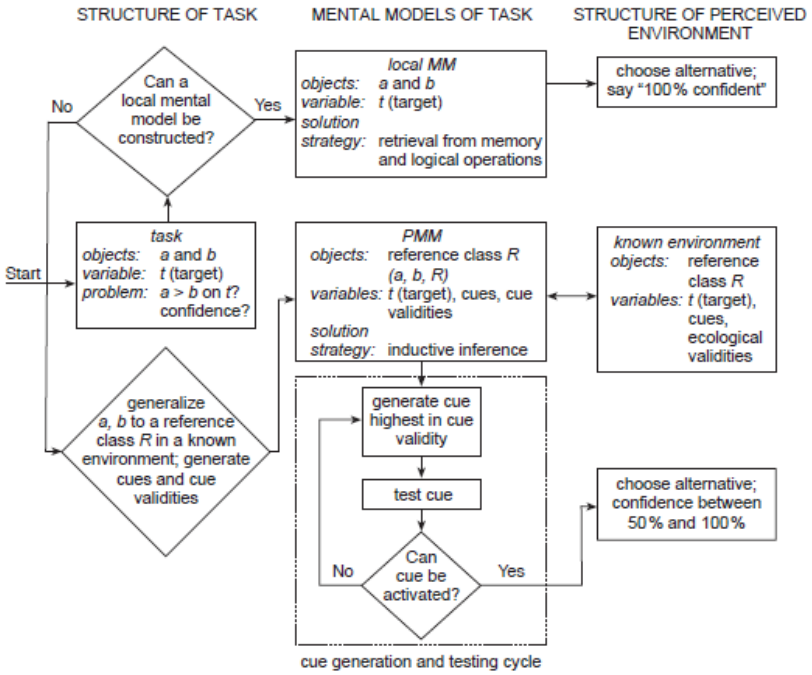

Figure 3.1 Cognitive processes in solving two alternative task, Gigerenzer (1991)

A local MM can be created successfully if the subject can retrieve precise figures from memory for both alternatives, retrieve non-overlapping intervals, and have elementary logical operations compensate for absent knowledge. The structure is as

follows: first, the task is local, meaning only two alternatives are taken into account and there is no generation of a reference class; second, it is direct, meaning no probabilities cures are utilized and it only contains the target variable; third, only inferences of elementary operations of deductive logic occur; and fourth, upon a successful search the confidence in the produced knowledge is evaluated as certain (Gigerenzer *et al*,1991).

If a local MM cannot be created by the subject, a PMM is created and solves the task with inductive inference by connecting the structure of the task with a corresponding probability structure. The PMM differs from the local MM because it contains a reference class, it uses a network of variables in addition to the target variable, hence it is neither local nor direct, the cognitive process includes probabilistic inference, and uncertainty is included in the outcome. The purpose of the reference class is to "determine which cues can function as probability cues for the target variable and what their cue validities are," where cue validities are thought of as conditional probabilities (Gigerenzer *et al*,1991). The probability cues are then generated, tested, and when possible, activated with the assumption that the order in which cues are generated is based on the hierarchy of cue validities. The end of the cue generation and testing cycle for typical general knowledge questions comes if there is a time constraint or the number of problems is large and after the first cue is found that can be activated if the activation rate of cues is small. The cue validity determines the choice of answer and confidence judgment, with choice following the rule; choose a if $p(a \mid aC_ib;R) > p(b \mid aC_ib;R)$, and the confidence, if a is chosen, that a is correct is given by the cue validity; $p(a \mid aC_ib;R)$ where a and b are objects that are

included in reference class R and C_i is a probability cue for a target variable in R (Gigerenzer *et al*,1991).

Gigerenzer *et al* (1991) also argue that the probabilities of single events (confidences) and the relative frequencies are not evaluated by the same cognitive processes. They state that the general knowledge tasks that are frequency tasks, i.e. involve a judgment of the frequency of correct answers, can rarely be answered by creating a local MM because the structure of the task contains one sequence of N questions and answers with the number of correct answers being the target variable. In addition, the PMM's of frequency tasks and confidence tasks differ because each supports a different reference class, the target variables are different, and the PMM of the frequency task will contain different cues and cue validities. Overall, the PMM is what connects the task structure with a known structure the subject's environment and confidence and frequency judgments refer to different kinds of reference classes (Gigerenzer *et al*,1991).

Five predictions are made from Gigerenzer *et al* (1991) concerning overconfidence with relation to the PMM framework provided above. The first prediction is that "typical general knowledge tasks elicit both overconfidence and accurate frequency judgments" with the term typical referring to a set of questions that is "representative for the reference class "sets of general knowledge questions"" (p.10). This condition is brought forth from the notion that "if PMMs for confidence task are well adapted to an environment containing a reference class R and the actual set of questions is not representative of R, then confidence judgments exhibit overconfidence" (Gigerenzer *et al*, 1991 p.511). In addition, if correct answer frequency question PMMs are well adapted with respect to a

reference class *R*' environment and the actual set of questions is representative of *R*', then frequency judgments are expected to be accurate. Combined, the statements predict that the same subject will exhibit overconfidence in a particular answer and accurate estimates of a frequency of correct answers. The two points on the left side of figure 3.2, provided by Gigerenzer *et al* (1991, p.511), are representative of this prediction.

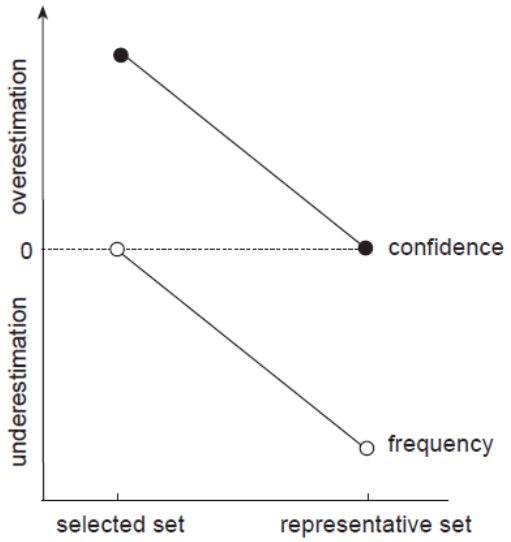

Figure 3.2 Predicted differences between confidence and frequency judgments, Gigerenzer (1991)

The second prediction is that "if the set of general knowledge tasks is randomly sampled from a natural environment, where natural environment denotes a knowledge domain familiar to the subjects, we expect overconfidence to be zero, but frequency judgments to exhibit underestimation" (Gigerenzer *et al*, 1991 p.512). This condition is brought forth from the notion that "if PMMs for confidence tasks are well adapted with respect to *R* and the actual set of questions is representative sample from *R*, then overconfidence is expected to disappear" (Gigerenzer *et al*, 1991 p.512). In addition, if correct answer frequency question PMMs are well adapted with respect to *R*', and the actual question set is not representative for *R*' then frequency judgments are expected to be underestimations of true frequency. The others side of figure 3.2 is representative of this prediction.

The third prediction is "comparing estimated relative frequencies with true relative frequencies of correct answers makes overestimation disappear" (Gigerenzer *et al*, 1991 p.512). Basically, overestimation or underestimation is expected to be null if the set of questions is randomly sampled. Thus, psychologically speaking there is a real distinction between confidence and relative frequency wherein subjects do not believe a confidence judgment of *X*% implies a relative frequency of *X*%.

Prediction four states, "if two sets, hard and easy, are generated by the same sampling process, the hard easy effect is expected to be zero" (Gigerenzer *et al*, 1991 p.512). If both the hard and easy set deviate equally from representative sampling, points will lie on a zero confidence parallel line.

The fifth and final prediction is "if there are two sets, one is a representative sample from a reference class in a natural environment, the other is selected from another reference class for being difficult, but the representative set is harder than the selected set; then the hard easy is reversed" (Gigerenzer *et al*, 1991 p.512).

Gigerenzer *et al* performed two studies to test the first, second, and third predictions. The results of the studies suggested that the first prediction is viable because there were quite accurate frequency judgments that coexisted with overconfidence. The second prediction was deemed viable because in the representative set, overestimation in confidence judgments disappeared and zero overconfidence coexisted with frequency judgments showing large underestimation. In the second study, overconfidence in single answer coexisted with mostly accurate frequency judgments that showed slight underestimation. For the third prediction, the subjects of the first study distinguished

between confidence in single answers and relative frequency of correct answers which was implied by the PMM theory where different reference classes are cued by confidence and frequency tasks. The only finding that didn't conform to prediction three was that the magnitude of underestimation was not as pronounced as expected. The second study resulted in the subject's average estimated percentage correct was different from confidence and similar to the actual percent correct.

Gigerenzer *et al* (1991) summarize that overconfidence results from one of two causes or both jointly; "a PMM for a task is not properly adapted to a corresponding environment, or the set of objects used is not a representative sample from the corresponding reference class in the environment but is selected for difficulty" (p.25). Therefore confidence should not be eliminated if the first cause is true and should be eliminated if the second is true.

Brenner, Koehler, and Liberman (1996) disagree with the final results of Gigerenzer *et al* (1991) because of new studies in the recent past. Brenner *et al* (1996) state that the predictions of the PMM, that a random sampling of questions from a natural domain will eliminate confidence and random sampling will eliminate the hard easy effect, were both wrong. This is due to recent studies that show for questions of relatively high difficulty, overconfidence is prevalent with random sampling of the questions from a natural domain of knowledge. Brenner *et al* review a study done by Griffin and Tversky (1992) whereby random pairs of states were selected (e.g. Virginia, Iowa) and subjects were asked which state had a higher value for some attribute (e.g. population) with an assessment of their confidence in each answer. The results showed significant

overconfidence for all attributes (contrary to first prediction) and also overconfidence for the more difficult attributes (contrary to the second prediction).

The majority of biases that can be brought forth by the use of uncertainty in decision making were presented in this chapter. The subsequent chapters will review some proper methods for eliciting unbiased probabilities and other methods for improving the assessments provided.

CHAPTER 4 How To Elicit Probabilities Properly

Even though people can be irrational when left to their own devices, they are not typically irrational in all situations, therefore an analyst should aim to structure a situation where biases are minimized and the assessments determined actually symbolize the subject's opinions (Wallsten & Budescu, 1983). This chapter is aimed at providing the methods to accomplish such a task. The methods presented by Spetzler and Von Holstein (1975) and Keeney and Von Winterfeldt (1991) are aimed at eliciting accurate probabilities while minimizing the effects of the biases presented in the previous chapter.

Spetzler and Von Holstein (1975) start with a set of five principles to use as a guideline for encoding uncertain quantities, and violating any one of these principles can lead to problems in the encoding process. The first principle is to choose only those uncertainties that are important to the decision itself. Principle two is to define the quantity as an unambiguous state variable; a state variable being one that has values beyond the control of the decision maker. The third principle states that the quantity should be structured carefully with thoughtful consideration to conditionalities. Principle four is to define the quantity precisely by performing the clairvoyance test. The fifth and final principle is to describe the quantity using an appropriate scale that the subject finds meaningful, so that the subject doesn't spend time trying to fit his/her answers to the scale versus evaluating the uncertainty.

When it comes to encoding methods, Spetzler and Von Holstein (1975) review the P (probabilities) and V (values) methods for answers that can be represented as points on a cumulative distribution function. There are three basic types of these methods: the P-method requires the subject to specify probability points with fixed values, the V-method requires the subject to specify value points with fixed probabilities, and the PV-method requires the subject to answer questions based on both scales simultaneously. These methods are explained in more detail later by Clemen and Reilly (2001).

The above mentioned questions can either be answered directly, by providing numbers, or indirectly, by choosing between alternatives until an indifference point is determined. Direct assessment, or structured questioning, supports data generation with questioning schemes that are designed to trigger elements in the subject's associate memory (Benson *et al*, 1995). There are multiple indirect response techniques to choose from. One option is the probability wheel which is a disk with two adjustable sections and a fixed pointer in which the subject determines which event is more likely until equality is reached. Another option is to use fixed probability events whereby the subject is asked to judge values that correspond to fixed probabilities. One can use the interval technique whereby an interval is split and the subject is asked which bet is preferred then subintervals are introduced until quartiles are determined. An additional option is the use of relative likelihoods, which is when the subject is requested to assign odds to two well defined events. There are also multiple direct response techniques to choose from. One option involves cumulative probability and fractiles, whereby the subject is asked to assign the cumulative probability at a given value or vise versa. A visual option involves the use of

graphs in which the subject either draws or must choose one graph, out of several provided, with a density function. Another option uses verbal encoding whereby verbal descriptions are used to characterize events and then quantitative interpretations are added. Benson, Curley, and Smith (1995) suggest that graphic representation is one of the most used and best developed aids for assessment, with decision trees being the dominant form of problem representation. However, graphic methods have a limitation in being a weak account of relevance because they construe relevance as probabilistic dependence. Thus, graphic representation, like other existing aids, has a limiting account of relevance and therefore has a tendency to constrain rather than exploit the variety of reasoning strategies that humans are capable of (Benson *et al*, 1995). The indirect response mode is generally the better way to begin encoding because subjects tend to experience difficulty in giving direct numerical probability (Spetzler & Von Holstein, 1975). Additionally, indirect assessment avoids the availability heuristic because subjects will not be assessing probabilities directly based on memory.

Spetzler and Von Holstein (1975) provide an interview process procedure that is divided into five phases: motivating, structuring, conditioning, encoding, and verifying. The purpose of the motivating phase is twofold; introduce the subject to the encoding task at hand and explore the possibility of any motivational biases. The structuring phase is designed to define the uncertain quantity to be assessed. Wallsten and Budescu (1983) suggest that when encoding an expert's (or non-expert) opinion, it is imperative that the analyst carefully specifies the class of events in question, the sources of information to be considered, and the potential causes of unreliability in the information (e.g. sampling

error). The anchoring and adjusting heuristic can play a key role when/if the analyst accidentally gives undue weight to the earliest information considered. The third phase, conditioning, is meant to review how the subject assesses probabilities so that any potential biases can be avoided. The analyst should review the heuristics with the subject so as to introduce them to how the majority of people process uncertainty and thus attempt to eliminate the majority of the biases associated with the heuristics. The encoding phase is exactly that, encoding the uncertain quantity. The suggested route to take is to begin by asking for the extreme values, use the probability wheel with a new set of values to encode the corresponding probability levels, plot each response provided on a cumulative distribution, and use the interval technique to generate the median and quartiles. This route is suggested to ensure that the biases associated with anchoring and adjusting are eliminated.

In the final verifying phase the assessment is reviewed and tested for subject approval. If the subject is not comfortable with the assignments he has provided, the process may need to be repeated as many times necessary until the subject is confident with his judgments. A typical interview should last thirty to ninety minutes depending on the complexity of the situation. Spetzler and Von Holstein (1975) make a final note that the pre-encoding steps generally take longer than the encoding steps but are the most essential to good probability assignments. In addition, interactive techniques have proven to facilitate judgment by minimizing randomness and eliminate unwanted inconsistencies (Wallsten & Budescu, 1983).

Keeney and Von Winterfeldt (1991) provided some lessons learned from the NUREG 1150 study (a study performed by the Nuclear Regulatory Commission to estimate the uncertainties of accidents in nuclear power plants) and also a recommended process for eliciting probabilities from experts. The recommended process for eliciting probabilities from experts consists of seven components: identification and selection of the issues; identification and selection of the experts; discussion and refinement of the issues; training for elicitation; elicitation; analysis, aggregation, and resolution of disagreements; and documentation and communication.

The identification and selection of issues component emphasizes that even before the study begins, it is important to recognize all uncertainties that need expert probability elicitation and select the ones that need to go through a formal elicitation procedure. The second component, identification and selection of experts, is crucial and should involve both specialists and analysts. Specialists should be the most knowledgeable in their field of study and the analysts who perform the elicitation should have expertise in probability theory, statistics and decision analysis. The third component, the discussion and refinement of the issues, should be conducted in a meeting of the specialists and analysts. The intent of this first meeting is to ensure that the specialists understand the project and what is expected of them and to arrive at unambiguous definitions of the events and the uncertainties that will be elicited. The in depth definition discussion amongst experts should help to eliminate biases associated with the representative heuristic with the elimination of any ambiguity.

After the first meeting is completed, the issues are clarified, and the specialists are given some time to think about the issues at hand, the training for elicitation should be conducted. It is suggested that the training be performed by an analyst with intent to familiarize the specialist with the techniques used in elicitation, provide them with an assessment practice session, inform them of the potential biases that can occur, and motivate them on their judging task at hand. Wallsten and Budescu (1983) determined that the accuracy of the assessments determined by either direct or indirect methods can be increased with just simple and short training procedures.

Keeney and Von Winterfeldt (1991) proclaim that the elicitation process, the fourth component, is an interview between the specialist and analyst with a possible generalist assistant who provides project inputs and recording services. The interview should start with an informal discussion of the specialist's approach to the issue and review of the event definitions in order to map out the decomposition to identify the necessary judgments. Next, a series of easy questions should be asked to determine the probabilities or probability distributions of the decomposition, followed by harder questions. Consistency checks should be administered once the judged probabilities are obtained. The endpoints or boundaries are next defined, followed by the fractiles with assessing the median judgment first. The determination of endpoints first helps to eliminate any anchoring that might occur on behalf of the specialist. A graphical depiction should be maintained throughout the entire process with continuous consistency checks.

The next step in Keeney and Von Winterfeldt's overall process is the analysis, aggregation, and resolution of disagreements component. It is not always feasible to

combine the specialist's judgments during the elicitation process, thus the analyses should be provided to the specialist as soon as possible after the interview so that any necessary changes can be made while the interview is still fresh in mind. Aggregation across the experts is then performed by an averaging process with the addition of a sensitivity analysis. A final group meeting can be held for all specialists and analysts involved to establish any agreements or disagreements on the probabilities elicited or events in question.

The final component, documentation and communication, is actually performed throughout the entire elicitation procedure. At the very minimum, documentation of elicitation results and reasoning, event definitions, expert identification and selection, the training session, the actual elicitation session, and the aggregation and analysis should be maintained.

Mellers and Locke (2007) have additional suggestions for the elicitation procedures to eliminate biases. The first concerns disclosure, which is a common procedure for constraining self interest and is based on the assumption that when a subject reveals of a conflict of interest he will provide a less biased assessment or the analyst will discount the assessment and thus result in an unbiased decision. A study performed by Cain *et al* (2005), however suggested that disclosure actually increased the subject's bias but also increased the analyst's discount, but not enough to correct the initial increase in bias. Mellers and Locke (2007) believe the reasoning behind this to be that subjects may feel that if their self interests are disclosed they are less liable to recipients and decision makers should rely more heavily on their own information without knowing how much the conflict

of interest has affected the provided assessment. In real life, if the decision maker trusts the subject providing the assessment and the decision maker is unsure of his own information, having a provided disclosure may allow the interpretation that the subject's information is more credible.

Another suggestion from Mellers and Locke (2007) for the elimination of biases during the elicitation procedure is to apply accountability to the situation, because pre-decision accountability to an unknown audience can reduce cognitive biases especially those biases resulting from lack of effort of understating one's own judgmental processes. This suggestion can however have a null or detrimental effect if the decision maker hedges his choice by sticking with the status quo. For the last suggestion, Mellers and Locke (2007) state that some tasks, such as recall of items and the effects of anchoring and adjusting, showed improved performance with the use of incentives. Incentives can however be harmful if the task is too complicated and most studies actually showed little effect on performance with the use of incentives.

Clemen and Reilly (2001) identify three basic methods for assessing discrete probabilities. The first method is to simply ask the decision maker what the probability of the event's occurrence is. The problem with this method is the lack of confidence that the assessor will undoubtedly have in their answer. The second method is to question the decision maker about any bets they would be willing to place on the event's occurrence. The notion behind this method is to find a specific value to win or lose such that the decision maker is indifferent about which bet to take. The example provided for this method is centered on an NBA championship game between the Lakers and Pistons. The

value of interest is the decision maker's probability that the Lakers will win the game. The decision maker is willing to take either of the following two bets: Bet 1, Win $X if the Lakers win or Lose $Y if the Lakers lose; and Bet 2, Lose $X if the Lakers win or Win $Y if the Lakers lose. Figure 4.1, provided by Clemen and Reilly (2001, p.300), is the decision tree that represents what the decision maker faces. As you can see by the decision tree this is a V-method because the value is being modified until an indifference point between the bets is determined. With some simple reductions, the subjective probability for the decision maker that the Lakers will win can be represented by the equation: $P(La\ker sWin) = \frac{Y}{X+Y}$. Thus, if a friend was willing to take either side of the bet: Win $2.50 if the Lakers win or Lose $3.80 if the Lakers lose, his subjective probability would be 3.80/(2.50+3.80) = 0.063 (Clemen & Reilly, 2001).

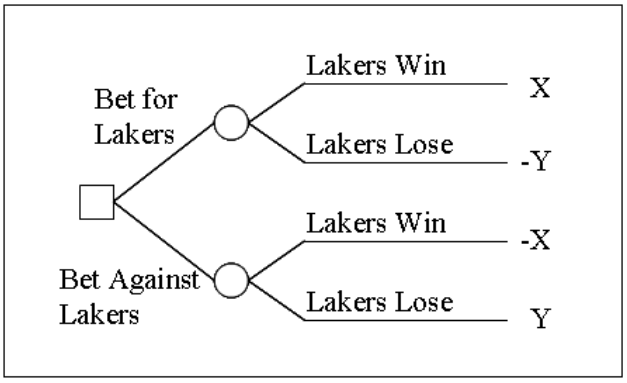

Figure 4.1 NBA Bet decision tree, Clemen & Reilly (2001)

Clemen and Reilly (2001) explain that the method to finding a bet that a decision maker is willing to take either side of is fairly basic. The first step is to offer a bet that is highly favorable to one side or another and make note of which side the decision maker chooses. Next, offer a bet that heavily favors the other side and ask which side is now preferable. Continue this process of offering bets back and forth that favor each side and adjust the payoffs gradually as you persist. By adjusting the bet to be either more or less attractive based on the previous bet each time, the indifference point can be reached. This

method of starting with extremes and working inward avoids anchoring, assuming that the starting points are extreme enough to bound the possible values. This process nonetheless may not avoid the framing bias because the prizes are variable and one frame could look better than the other even though they are the same situation based on the same outcome. The betting approach also has related problems; most people are risk averse and don't like the idea of betting and this approach presumes that the individual cannot protect himself from losses by "hedging" one bet with another. Yet, because this approach does not permit protection by "hedging" the conservatism bias should be avoided.

The third method for assessing discrete probabilities is to have the decision maker compare two lotterylike games, either of which can result in a Prize (A or B) (Clemen & Reilly, 2001). This method is easiest if the prize structure is set up so that the decision maker prefers Prize A (e.g. a free vacation in Hawaii) to Prize B (e.g. a free beer). This method is representative of the P-method because the values (prizes) are fixed and the point is to obtain the associated probabilities. Clemen and Reilly continue their prior example concerning the NBA championship game between the Lakers and the Pistons with this method. For this method, the analyst would ask the decision maker to compare the lottery, Win Prize A if the Lakers win or Win Prize B if the Lakers lose, with the second 'reference lottery', Win Prize A with known probability *p* or Win Prize B with probability 1-*p*. Figure 4.2, provided by Clemen

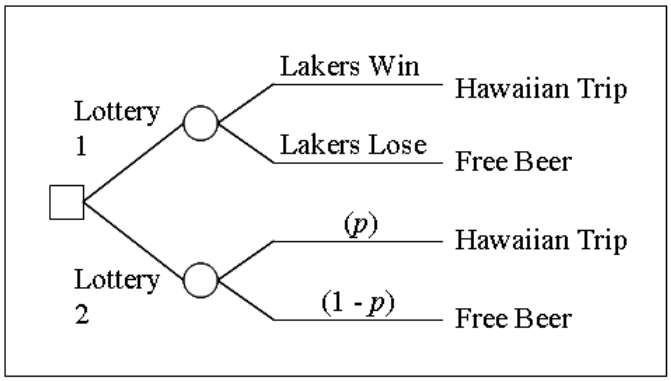

Figure 4.2 NBA Lottery decision tree, Clemen & Reilly (2001)

and Reilly (2001, p.302), is the decision tree representation for this decision. The probability mechanism must be well specified for the reference lottery and can be as simple as drawing a colored ball from an urn with proportion p of colored balls or using a "wheel of fortune," with a win section and a pointer that provides the decision maker with Prize A when it lands in the win section. Once an understood mechanism is in place, the analyst adjusts the probability of winning in the reference lottery until the decision maker reaches indifference between the two lotteries. The subjective probability that the Lakers win must be the p that makes the decision maker indifferent, meaning that the decision maker has no preference between the two lotteries and slightly changing the probability p makes one or the other clearly preferable.

Clemen and Reilly define a method for obtaining the p that makes the decision maker indifferent between the two lotteries. Start with some probability p_1 and question which lottery is preferred. If the reference lottery is preferred, then p_1 must be too high, so choose a p_2 less than p_1 and ask for the preference again. Continue this process of adjusting the probability in the reference lottery and questioning preferences until the indifference point is found. An analyst should start with exceptionally wide brackets and converge on the indifference point slowly, allowing time for the decision maker to deliberate about the assessment and thus avoiding anchoring. Additionally, because the prizes are fixed in this method, framing effects can also be avoided. There are still problems even with this approach to probability assessment. Some people don't like the idea of lottery game and others have difficulty understanding a hypothetical game and thus have problems providing realistic assessments.

The last step in the discrete probability assessing process for all three methods is to check for consistency. It is important that the probabilities be consistent amongst themselves and the assessed probabilities should also obey all probability laws; probabilities must lie between zero and one, probabilities must add up, and total probability must equal one (Clemen & Reilly, 2001). If the set of assessed probabilities is found to be inconsistent, the decision maker should reconsider and make modifications where necessary.

Clemen and Reilly (2001) also discuss two methods for assessing a subjective continuous probability. The first strategy is to assess the probability for a few values, plot them, and then draw a smooth curve through the points. Thus, one must select a few values from the horizontal axis and then assess the corresponding cumulative probabilities. The example provided for consideration is deriving a probability distribution to represent the uncertainty of a movie star's age. A typical cumulative assessment would be represented by P(Age ≤ *a*), where *a* is the particular value. So if the subject were to consider P(Age ≤ 46), the decision maker could use any of the three methods previously provided for assessing discrete probabilities to find the probability *p* that would make them indifferent between the lotteries. Figure 4.3, provided by Clemen and Reilly (2001, p.304), is the decision tree representation of this problem.

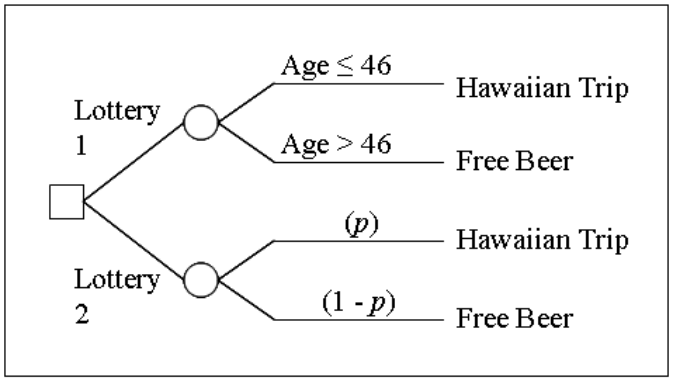

Figure 4.3 Age cumulative distribution decision tree, Clemen & Reilly (2001)

The second strategy for assessing the cumulative distribution function is to reverse the method and pick a few cumulative probabilities from the vertical axis and assess the corresponding ages. An example for this method is the supposition that the decision maker wants to assess a corresponding value for the probability 0.35. Consequently, $a_{0.35}$ is called the 0.35 distribution *fractile*. By using what was assessed; P(Age ≤ 29) = 0.00, P(Age ≤ 40) = 0.05, P(Age ≤ 44) = 0.50, P(Age ≤ 50) = 0.85, and P(Age ≤ 65) = 1.00, it can determined that the 0.35 fractile of the distribution is approximately 42 years. The general term *fractile* includes the median (the 0.50 fractile), quartiles (the 0.25 and 0.75 fractiles), and even the percentiles (the 0.10 fractile is the 10^{th} percentile) of the distribution. After eliciting the values for each of these fractiles, the points should be plotted on a graph and a smooth line should be drawn through the points (Clemen & Reilly, 2001).

A situation of discontent is the argument of using verbal assessment (i.e. using descriptive words such as most likely, fairly, least likely, etc.) versus numerical and the notion that most people would prefer verbal estimates because they convey their opinions more accurately. Wallsten, Budescu, and Zwick (1993) remind us that people, generally speaking when making a decision, prefer to communicate verbally and receive information numerically. In addition, Wallsten *et al* determined that the overall quality of decisions made from verbal probability assessment was not different than those made with numerical probability assessment aside from differences in their patterns. Wallsten *et al* (1993) propose that if a verbal probability elicitation is to be performed, the analyst should provide a fixed list of approximately eleven to fifteen words and those words should be

assigned numerical values by the subject. A main difference between the two methods (verbal vs. numerical), as determined by Wallsten *et al* (1993) in their experiment, is that while overconfidence was prevalent in both models, the magnitude was greater in the verbal response method. They recognize that the honesty encouraged by the verbal method allows the analyst to see that overconfidence is even greater than was apparent when using the numerical method. The main advantage of verbal encoding is that the forecaster is more comfortable with their assessments, however numerical encoding provides a higher level of resolution, thus the trade-off is that the 50% category is more widely used in the numerical method and overconfidence is greater with the verbal method. Wallsten *et al* (1993) suggest that the reasoning behind using the 50% category in the numerical method is because it is equally defensible regardless of the final outcome and furthermore, the verbal method, because it allows vagueness, relieves the stress of assessment defense.

The basic steps of proper probability elicitation from an expert – explaining the problem in depth, training the subject on the cognitive heuristics and their associated biases, eliciting the probability through an in depth interview process, and verifying the assessments – will eliminate most biases presented in chapter three. There are however further steps that can be taken to improve the probabilities because of overconfidence and other biases that slipped through the cracks during the elicitation process. These improvement methods are presented in chapter five.

CHAPTER 5 Improvement of Assessments

This chapter brings to the forefront ways to improve the probabilities elicited, eliminate some of the biases defined in chapter three that weren't eliminated by the elicitation procedures defined in chapter four, and in addition, some potential assessment practices to avoid them from the very start. This chapter will review the concept of calibration and its role in the process of probability elicitation. Concerning calibration and refinement, Wallsten and Budescu (1983) and Liberman and Tversky (1993) provide some general thoughts on the subject, Harrison (1977) discusses issues with independence and calibration, Degroot and Fienberg (1983) provide calibration and refinement methods for evaluating and comparing forecasters, and Keren (1997) reviews some of the problems associated with calibration. The issue of coherence is discussed by Lindley, Tversky, and Brown (1979) who tackle the problem on incoherence with two reconciliation approaches and Lindley (1982) who evaluates whether assessment can be improved by having a better understanding of the assessment mechanism. Finally, the concept of decomposition is reviewed by Ravinder, Kleinmuntz, and Dyer (1988) who suggest the decomposition method for obtaining more reliable subjective probabilities, Howard (1989) who shows us how knowledge maps that represent decomposition are a convenient process for representing knowledge and the involved uncertainty, and finally, Moskowitz and Sarin (1983) who test whether using a judgmental aid will reduce in the number of inconsistent probability assessments.

5.1. Calibration and Refinement

A method for attempting to establish the validity of a probability elicitation method or of a subject is deemed calibration (Wallsten & Budescu, 1983). Calibration is determined by whether for all events assigned probability *p*, the proportion that occurs is actually *p* and it is therefore an indirect measure of validity. Wallsten and Budescu (1983) also note that there are training methods that can improve calibration and one method in particular suggested by Koriat *et al.* is to have the subjects list reasoning for and against their assessments which yields improvement because subjects generally fail to consider the counter-indicants for their decisions. Lastly, Wallsten and Budescu (1983) determined that experts are generally well calibrated when making assessments in their field of expertise while non-experts are limited but do show great improvement with training.

Harrison (1977) provides a discussion on the operational issue brought forth by calibration error and he argues that, "the existence of calibration error does not affect the theoretical justifications for the subjective view of probability, but does severely restrict what we can do with subjective assessments" (p.320). Thus, the issue is whether we can justify independence assumptions with subjectively assessed probabilities. This means that if a subject learns about the occurrence of an event it will potentially inform him of his own assessing characteristics and thus might affect the assessment he makes for the other event. Harrison uses the calibration exercise, which is to select random variables of unknown quantities, assess the subjective distribution for each, observe the actual value for each variable, and then compute the calibration curve which plots assessed probability against observed frequency. The calibration curve should mirror a 45 degree line and if it

doesn't, the subject is deemed miscalibrated as an assessor for that particular group of distributions. Figure 5.1 shows three separate calibration curves: a well calibrated forecaster, an overconfident forecaster, and an underconfident forecaster. Note the overconfident forecaster's curve is representative of his tendency to use values close to the boundaries instead of values approximate to the median range. The underconfident forecaster exhibits the exact opposite trend with over-usage of the range of values close to 0.5 and under-usage the extreme values.

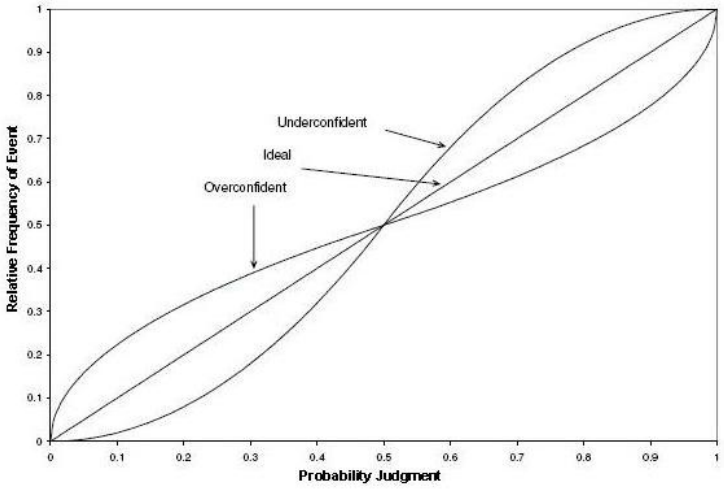

Figure 5.1 Representative Calibration Plots

Liberman and Tversky (1993) designate two forms of calibration plots, the designated form, where a target event is pre-selected for each problem and the data are displayed based on the assigned probabilities to the events with disregard to their complements, and the inclusive form, where the probabilities assigned to the two complementary events are incorporated for each problem. The usage for each form differs in the designated form is generally used when all judgments refer to a common hypothesis

(e.g. rain vs. no rain, etc.) and the inclusive form is generally used with general knowledge problems that have no common hypothesis, although the form is not dictated by the hypotheses under consideration. Liberman and Tversky (1993) note that both the designated and the inclusive indices can be useful for describing and evaluating probability assessments but the inclusive analysis is deemed appropriate when there is interest in the subject's use of the probability scale, not necessarily the specific outcome, and the designated analysis is more relevant when there is interest in the specific outcome prediction.

Harrison (1977) provides an example where a subject is asked for the subjective probability distributions (F_k) for a series of 1000 numerical almanac quantities (X_k). The analyst uses the probability wheel elicitation procedure to aid in the probability assessments by determining the setting of the wheel that makes the subject indifferent between the lottery L, where a payoff depends on the percentage of the random variables that actually have values less than or equal to the 30^{th} percentile of the assessed distribution, and the reference lottery. The analyst then looks up the actual numerical values for each of the quantities and plots the calibration curve. Now the subject has an opportunity to view how calibrated he is and determine if it's a consequence to how reasonable his axioms are. Harrison notes that the current literature suggests that people are generally quite badly miscalibrated as probability assessors and the risk of being miscalibrated averts the subject from treating events as independent. Miscalibration can occur when a subject anchors to probabilities that are too high or too low, thus the anchoring and adjusting heuristic is in effect.

Harrison then provides a statement (*) that the subject might subscribe to concerning the nature of the perceived dependence among events that would enable some concrete calculations, since the dependence issue makes computation nearly impossible. Statement (*) declares that there exists a random variable α, such that $E_1,\ldots, E_{1000}$ (where events $E_k = \{X_k \leq x_k\}$) are conditionally independent given α, with $P(E_k|\ \alpha) = \alpha$ for all k. Statement (*) might aid in providing a reasonable approximation to the type of dependence that the subject feels is in existence, and the degree of dependence is expressed by the subject's prior distribution for α. Thus it follows that mean of the distribution is $E(\alpha) = P(E_k) = 0.30$. Using this statement, Harrison reviews three problems and in each case $P(\cdot)$ denotes a probability computed from Statement (*) and the indicated prior distribution for α, while $Q(\cdot)$ denotes a probability based on the assumption that $E_1,\ldots, E_{1000}$ are independent.

Harrison's problem 1 considers a lottery where P gives a wider spread distribution for K than does Q, and the effect is more pronounced when the outcomes from a larger number of trials are considered. The problem also provides the option of buying perfect information as to the value of α before deciding whether or not to play. The option to pay for information shows whether the subject views himself as miscalibrated, and thus validates the suggestion that when a subject worries over his potential miscalibration his assessments are modified. Problem 2 considers a lottery where a severe loss is suffered if $E_1,\ldots, E_8$ all occur and a moderate gain is enjoyed otherwise. The probabilities are calculated ($Q(E_1$ and … and $E_8) = (0.30)^8 = 6.56\times10^{-5}$ and $P(E_1$ and … and $E_8) = E(\alpha^8) = 29!16!/8!371 = 3.33\times10^{-4}$) and the lottery is determined to be unattractive because the

negative outcome probability is larger by a factor of about five. Harrison notes that the percentage error caused by an assumption of independence grows rapidly when one increases the number of events being intersected. This is a situation however, where the assessor's uncertainty about his calibration must be accounted for. The third problem lets $K = \theta_1 + \ldots + \theta_{10}$ (where θ_k is an indicator variable that equals 1 if E_k occurs and 0 otherwise) and a lottery L that pays an attractive prize if $K = 2$, 3, or 4, which are the three most likely single values for K, but pays nothing otherwise. The probabilities P(win) and Q(win) are computed to be 0.63 and 0.70 respectively suggesting that once again the assumption of independence introduces significant error.

Harrison presents De Finetti's theorem on how a subject must subscribe to Statement (*) if they believe that the original collection of random variables X_k could be augmented so that the sequence of events $E_1, E_2,\ldots$ could be viewed as exchangeable. The theorem consists of two statements and the first is that there exists a random variable α such that $P\{\alpha_n \rightarrow \alpha \text{ as } n \rightarrow \infty\} = 1$, where α_n denotes the fraction of the first n events that occur. Therefore, even though the subject may not actually know the exact value, they are positive (probability of one) that a limiting frequency will be achieved. The second statement of De Finettis' theorem is that the events E_k are conditionally independent given α, with $P(E_k \mid \alpha) = \alpha$. Thus, α is defined as a limiting calibration factor. De Finetti's theorem shows that a subject can account for his uncertainty about his calibration through simple calculations. Overall, Harrison believes that the current professional practice of probability elicitation using the lottery procedure to reduce a sensitive outcome is what opens up the possibility of serious error through unwarranted independence assumptions.

In addition, the analyst should have some trepidation about the potential effect of a subject's uncertainty of his own calibration.

DeGroot and Fienberg (1983) present an article that provides methods for evaluating and comparing forecasters. They make sure to first differentiate between calibration and refinement, as these two concepts are used for comparing forecasters. They define calibration as the "agreement between a forecaster's predictions and the actual observed relative frequency of rain" and refinement is specified as "how spread out or how sharp a forecaster's predictions are" (DeGroot & Fienberg, 1983 p.13). Thus a forecaster is well calibrated if $\rho(x) = x$, where $\rho(x)$ denotes the relative frequency of rain on all the days that the forecaster's prediction was x, for every value of x such that $v(x) > 0$, where $v(x)$ is the frequency function of the forecaster's predictions over the days. There are, however, a few reasons why a forecaster may not be as well calibrated as he would prefer. A forecaster's predictions are only observed for a finite number of days and there really is no reason why his predictions should realistically coincide with the actual occurrence of rain. Because being well calibrated is a desirable characteristic of forecasting, a forecaster could be tempted to actually lie and specify predictions that do not represent his true subjective probabilities in order to attain certain values, thus introducing a motivational bias. Even so, a forecaster can be well calibrated with his own predictions in the long run, but if they aren't accurate on a specific day, they are of no use. While calibration is a desired characteristic of forecasters, it does not ensure informativeness (Liberman & Tversky, 1993).

DeGroot and Fienberg's concept of refinement is used to compare the well calibrated forecasters. The logical concept for two forecasters who are well calibrated is as follows: "A is at least as refined as B if we can artificially generate a well-calibrated forecaster with the same probability function $v_B(x)$ as B simply by passing A's predictions through a noisy channel" (DeGroot & Fienberg, 1983 p.14). A forecaster who makes the same prediction each day is *least-refined* because any other well calibrated forecaster is at least as refined as he is. And on the other end, the forecaster whose predictions are binary and always correct is *most-refined* because he is at least as refined as any other well calibrated forecaster. It is possible that a comparison exists and it can be possible that two forecasters are just not comparable in terms of refinement. If this be the case, DeGroot and Fienberg present a theorem, supplied from their previous work (1982), which eliminates the need to construct an appropriate stochastic transformation. Theorem 1 considers two well-calibrated forecasters A and B and states that forecaster A is at least as refined as B if and only the following inequalities are satisfied: $\sum_{i=0}^{j-1}(x_j - x_i)\{v_A(x_i) - v_B(x_i)\} \geq 0$ for j=1, …, k-1. An example provided by DeGroot and Fienberg (1982) is, if A and B are well-calibrated with the probability functions; $v_A(x)$ = {0.1 for x=0, 0.8 for x=0.5, 0.1 for x=1 and $v_B(x)$ = {0.5 for x=0.1 and 0.5 for x=0.9, then neither A nor B is at least as refined as the other. The condition provides the reasoning that if "one well calibrated forecaster A is at least as refined as another well-calibrated forecaster B, and if we must choose between learning the prediction of A or learning the prediction of B, then we should choose to learn

the prediction of A regardless of the purposes for which we will use the prediction" (DeGroot & Fienberg, 1983 p.15).

It is also possible to compare forecasters who are not necessarily well calibrated with the concept of sufficiency provided by DeGroot and Fienberg (1983). They define sufficiency as, "A is sufficient for B if we can artificially generate a forecaster with the same functions $v_B(x)$ and $\rho_B(x)$ as B simply by passing A's predictions through a noisy channel" (DeGroot & Fienberg, 1983 p.16). However, when the forecasters are well calibrated, the sufficiency relationship reduces to the refinement relationship. Just like in refinement, all forecasters might not be comparable in terms of sufficiency and even if sufficiency exists it does not automatically assume that one forecaster is as good as the other. A good example provided is if forecaster A only makes binary predictions and is always wrong, then he is sufficient for every other forecaster even though he is the worst at forecasting. However, because he is always wrong, his predictions are just as useful as a forecaster who is always right. The concept of sufficiency is based on whether we should choose to learn the prediction of one forecaster versus another. Thus, even if A is sufficient for B (i.e. we should learn A's predictions), more information might be obtained if we were to learn both sets of predictions instead of just A's. DeGroot and Fienbert (1983) continue their analysis by using scoring rules to evaluate forecasters with relation to calibration and refinement. The topic of scoring rules will be covered in chapter 6.

Keren (1997) presents some of the problems associated with the research on calibration. His first problem is that studies in calibration assume that probabilities are subjective yet they are evaluated by frequentistic criterion which provides controversy on

the validity of this procedure. The second problem he brings forth concerns the trade-off between calibration and resolution, as calibration and resolution are not completely independent of one another and there may be incompatibility between the requirements for maximizing calibration and achieving high resolution.

Keren (1997) reviews the framework provided by Juslin *et al* that states there are three necessary requirements for being well calibrated: first is cognitive adjustment, requiring internal probabilities to equal the corresponding ecological probabilities; second is error free translation, assuming that translation from internal uncertainty to a response should be free from error; and third is representative design, requiring the "selection of tasks such that the probabilities of the events in the task sample should coincide with ecological probabilities" (p.270). Should these assumptions be violated, random error will be affected and thus reduce the calibration quality. Keren notes that this approach does adopt a frequentistic interpretation, which in turn makes confidence or probability assessments of unique events meaningless, thus limiting the scope of studying how people deal with uncertainty and also overlooking the fact that judgment of uncertainty in everyday life cannot be justified on frequentistic grounds. Furthermore, Juslin *et al* clarifies that the hard-easy effect is inconsistent with the conjunction of the three assumptions and in continuation, the third assumption is doubted because of studies that were performed that made an attempt to employ a representative design, subsequently showed clear overconfidence. Additionally, Keren (1997) provides the error source applicable to assumption two; internal probabilities are unobservable and therefore cannot be directly matched to external responses but insight can be obtained from the examination

of different response scales (e.g. whether the probability scale is half or full range, a choice or no choice situation, and/or discrete or continuous). Additional research to determine a clearer notion of the relevant factors is needed. Overall, Keren (1997) suggests that random error can be affected by more than one factor and "the magnitude of the random error affects the degree of goodness of calibration indirectly by determining the extent to which regression toward the mean will be present" (p.272).

5.2. Coherence

The concept of coherence, when an individual's assessments obey the probability postulates, is a necessity for assessed subjective probabilities (Clemen & Reilly, 2001). Wallsten and Budescu (1983) state that there is one axiom for a rational belief structure that leads to an additive probability measure, representative of a coherent opinion, and one that doesn't. An axiom system that doesn't lead to an additive probability measure is due to the lack of perfect sensitivity, a feature of human judgment, which is basically the reality that a subject is not capable of making infinitely fine judgments and therefore they cannot place order to events that are close in relative likelihood. Lindley, Tversky, and Brown (1979) explore the reconciliation of assessments that are contradictory to the laws of probability, or incoherent probability assessments. Some examples of this incoherence are probability assessments that are not additive or have conditional probability ratios that do not coincide with the corresponding unconditional probability ratios. They developed two approaches, an internal and external, to solve the reconciliation problem. "In the internal approach, the observed probability assessments are related to internal coherent

probabilities in a manner analogous to the relation between the observed score and the true score in test theory" (Lindley *et al*, 1979 p.149). Thus the internal approach is concerned with attempting to approximate the "true" probabilities using the observed assessments. The external approach is only concerned with deriving coherent probabilities using the original set of incoherent assessments instead of addressing any "true" probabilities.

The basic model for both approaches contains: one subject that will be providing the judgments and is not necessarily coherent (S), an uncertain event under consideration (A), a probability distribution for A ($\pi(A)$), a vector that describes the assessed values ($q(A)$), and an investigator who is coherent (N) and is tasked with reconciling the subjects stated values of the observed measurements (q) and providing an assessment of the true measurements (π). The investigator is concerned with the true measurements in the internal approach and the world external to the subject in the external approach. The investigator N could assess his own probabilities, denoted $p(\cdot)$ and thus provide a joint probability distribution of the uncertain quantities. The joint probability is described in three stages with each corresponding to a different aspect of N's contemplation of the problem; first, there is a distribution A, $p(A)$, second, there is a conditional distribution of π given A, $p(\pi \mid A)$, and third, there is a conditional distribution of q given the other two elements, $p(q \mid \pi, A)$. Basically, $p(A)$ can be seen as N's appreciation of the world, $p(\pi \mid A)$ as N's opinion of S's knowledge of the world, and $p(q \mid \pi)$ as N's opinion of S as an assessor.

The internal approach provides a complete probabilistic description of q and π because N is only concerned with π and q: $p(\pi) = \sum_{\underline{A}} p(\pi \mid A)p(A)$, where $\sum_{\underline{A}}$ is the

summation over a partition of events in A. Using Bayes theorem provides, $p(\pi \mid q) \propto p(q \mid \pi)p(\pi)$, which is N's appraisal of S after S has reported his assessments q and can be used when the subject has made several incoherent probability judgments and wishes to reconcile them to coherent values. With the direct form for the internal approach, i.e. using $p(\pi)$ and $p(q \mid \pi)$ to obtain $p(\pi \mid q)$, π is the set of parameters, q is the data, and the "prior" $p(\pi)$ is updated by the likelihood $p(q \mid \pi)$ to give a posterior probability $p(\pi \mid q)$. The technique of sending the whole argument through with log-odds instead of probabilities, because the q's cannot be normally distributed, becomes the most usable method for reconciliation. This is partly because log-odds are more convenient and sensible to work with than probabilities. In addition, log-odds are necessary to handle bivariate distributions and it's more reasonable to assume that the measurement error has constant variance when expressed in log-odds. The technique only then requires the addition of the variances and covariances of q, because they describe the normal likelihood, and an assumption that the prior is "flat."

Wallsten and Budescu (1983) have a different belief concerning the true measure of a subject and thus a different approach for the use of the concept of a true score. Wallsten and Budescu (1983) explain that their version of a true score incorporates the subject providing an assessment based on current knowledge and information at hand a large number of times (with no memory of each previous judgment) and the provided distribution would define the true score of that subject. Thus the true score t, is the expected value of the distribution obtained by a series of statistically independent

assessments made by a subject. So instead of assuming a set of internal subjective probabilities, they assume that independent probability assessments will provide distributions with well-defined expected values termed the true subjective probabilities for that subject, the specific task, and particular situation.

The external approach provides a probabilistic description of q and A because here N is only concerned with q and A: $p(q \mid A) = \sum_{\pi} p(q \mid \pi) p(\pi \mid A)$. Again, using Bayes theorem we obtain $p(A \mid q) \propto p(q \mid A) p(A)$, which can be used when the investigator is a decision maker in need of a probability for A for a decision to be made. The internal approach is deemed simpler than the external approach because $p(\pi)$ can be assessed directly and it does not need to use $p(\pi \mid A)$, nor the derived probability $p(q \mid A)$.

Lindley *et al*. (1979) also provide a least squares procedure with the internal approach in which they are only concerned with the subject's stated and coherent values for some events, q and π. With this, the reconciled values $\hat{\pi}_i = E(\pi_i \mid q)$ are approximated by the values of the π's that minimize $\sum_i (q_i - \pi_i)^2$. And in other situations, the q's will be correlated with different variances, when the quadratic form $\sum w_{ij} (q_i - \pi_i)(q_j - \pi_j)$ is minimized subject to the w's being weights. Minimization is performed by equating the first derivatives to zero, thus providing the approximate variances and covariances of the reconciled values from the inverted minimized matrix of second derivatives. The one difficulty associated with the procedure is that the coherence constraints are typically non-linear resulting in non-linear equations. There also comes a consideration of a metric (e.g.

probability metric, log-odds metric, etc.) when using the least squares application because the variance of the *q*'s may be more constant in one metric versus another. An alternative to changing the metric for precision purposes is for either *S* or *N* to provide the precision, "the degree to which it varies upon further reflection", associated with the values (Lindley *et al*, 1979 p.156). For example, the likelihood that that a coin will come up heads is very precise as it will not likely differ from 0.5.

Lindley (1982) explores the option of better understanding the assessment mechanism to improve the probability assessments. Lindley first notes that an investigator (*N*) is not essential to process the subject's (*S*) evaluations as in his previous work, though having it is not unreasonable. The question asked by Lindley is whether *q*, the log-odds for uncertain event *A*, can be transformed to a better log-odds assessment *q**. The subject's probability appraisal can be described as, if *A*($\bar{}A$) is true, let $f_i(q)[f_0(q)]$ be the probability that *S* will give log-odds *q*. Thus, for a subject who is excellent at assessing, *q* will be positive and large under *A*, and negative and large under $\bar{}A$. The only case however where *N* will believe *S* is when *q** = *q*: which from the log-odds assessment for $A, q^* = \log[\gamma f_1(q)/(1-\gamma)f_0(q)]$, is when $\gamma f_1(q)/(1-\gamma)f_0(q) = e^q$, or when log-odds $(A \mid q) = q$. Thus, *S* is probability calibrated, when the frequency and assessed probability agree, if this occurs. More generally, *S* is probability calibrated if there is a change from *q* to *q**.

The example provided by Lindley (1982) for this is to let $\gamma = ½$ and $f_1(q) = f_0(-q)$, so that *S* is just as likely to say *p* for the probability of *A* when *A* is true, as 1-*p* for *A* when $\bar{}A$ is true, furthermore suppose $f_1(q)$ is $N(\mu,\sigma^2)$, so that $f_0(-q)$ is $N(-\mu,\sigma^2)$. Then the

improved assessment will be the original assessment multiplied by $2\mu/\sigma^2$,

$q^* = -\frac{1}{2}(q-\mu)^2/\sigma^2 + \frac{1}{2}(q+\mu)^2/\sigma^2 = 2\mu q/\sigma^2$. Now, q^* will be $N(\mu_1,\sigma_1^2)$ with $\mu_1 = 2\mu^2/\sigma^2$ and $\sigma_1^2 = 4\mu^2/\sigma^2$ under A, and since $2\mu_1/\sigma_1^2 = 1$, so that S is probability calibrated, the original assessment is only probability calibrated when $2\mu = \sigma^2$.

Basically, the subject may believe that he has biases ($f_i(q)$) that might affect his judgments. However, the distributions can be updated once the subject learns the truth about such events and therefore the basic assessments are improved because the subject learns about his own prejudices. An advantage of having an investigator, and identifying him with the subject, is that the idea of an infinite regress is avoided. However, a difficulty that arises when the subject and investigator are identified is that as soon as the subject learns about $f_i(q)$ he will most likely adjust his assessment procedure. As for changing the assessment procedure, the subject might continue to assess as before and then transform the assessments based on the noted bias or he might adjust the actual procedure. An example provided by Lindley (1982) of these two options is that a surveyor might recognize a bias in his theodolite (a measuring instrument) but continue to use it and make corrections to the readings accordingly, just like the previously discussed example where S transformed to q^*, or he might make an adjustment to the actual theodolite.

A connected difficulty in having the investigator separate from the subject is that if the subject is calibrated for one value of γ, N's probability for the event, he will not be calibrated for a different value. This difficulty can be overcome with French's (1980) idea that the distributions $f_i(q)$ might depend on γ. Lindley's example for this is the supposition

that the subject was informed that for a series of almanac type questions the suggested answers were at random with the upper answer being correct three quarters of the time. Reasonably, this information should affect the *f*'s. Therefore, the original log-odds equation, $\gamma f_1(q)/(1-\gamma) f_0(q) = e^q$, would become $\gamma f_1(q \mid \gamma)/(1-\gamma) f_0(q \mid \gamma) = e^q$, and calibration for all γ would be possible. Lindley supplies an example where $f_1(q \mid \gamma) \sim N(\mu + \delta, \sigma^2)$ and $f_0(q \mid \gamma) \sim N(-\mu + \delta, \sigma^2)$, with $\delta = \log[\gamma/(1-\gamma)]$, the initial log-odds for *A*. Verifying *S* to be calibrated for all γ with $2\mu = \sigma^2$, thus under the same conditions as before, is fairly easy. Normally, when the subject is asked to assess the probability for an unknown event, like the almanac questions, it is not described in a way that would favor its truth or falsity, and $\gamma = ½$. However, if the subject feels that the questioner phrased the event in a way that even though it appears to be true, it is more likely to be false, a catch is implicated and $\gamma < ½$. Lindley proves this suggestion with an example involving the relative latitude coordinates of places in Europe and North America; since subjects generally do not realize how far north Europe really is.

5.3. Decomposition

Ravinder, Kleinmuntz, and Dyer (1988) targeted their article around how an analyst can reduce inconsistency by relying on decomposition rather than direct elicitation by establishing the impact of the decomposition procedure on the reliability of the assessed probabilities. Decomposition is the process of breaking down a probability assessment into smaller or manageable chunks (Clemen & Reilly, 2001). Clemen and Reilly recognize three different scenarios where decomposition is appropriate: thinking about how the event

in question is related to other events, thinking about what uncertain outcomes could lead to the event in question occurring, and/or thinking about what uncertain outcomes must happen for the event in question to occur. Clemen and Reilly (2001) also state that the best decomposition to use is the one that is easiest to contemplate and that gives the most transparent view of the uncertainty in question. Ravinder *et al* (1988) regard decomposition as a useful technique for reducing the complexity of challenging judgment problems. The decomposition approach is able to capture the relationship between the event in question, the target event, and its background events (e.g. the survivability of a patient may depend on the disease they had, the strength of the patient, the complexity of the medical procedure performed, etc.) with conditional probabilities of the target event. For example, the different distributions of the survivability of the patient just mentioned could be assessed conditionally upon the outcome of the disease they had, or conditional upon the occurrence of multiple events (e.g. the disease the patient had was non-significant, the complexity of the surgery was minuscule, and the patient's strength was favorable). The broken down judgments are later combined using an aggregation rule based upon statistical considerations, i.e. the law of total probability provides a convenient method. The example provided by Ravinder *et al* (1988) is to denote the probability of the target event as A and assume it is assessed conditional upon its background events $B_1,\ldots,$ B_n. The event B_i can be a single event or a scenario and the set must form a mutually exclusive and exhaustive partition of the relevant event space, meaning it must encompass all possible background events that could/would lead to the target event, and thus: $\sum_{i=1}^{n} \Pr(B_i) = 1$. So the probability of the target event would

be: $\Pr(A) = \sum_{i=1}^{n} \Pr(A \mid B_i) \Pr(B_i)$. Obviously, elicitation by decomposition is more laborious than direct assessment, because a marginal and conditional probability must be assessed for each background event.

Ravinder *et al* (1988) use the psychometric framework developed by Wallsten and Budescu (1983) for evaluating the accuracy of probability encoding. The proposed model is as follows: each encoded probability is considered a random variable, x, composed of a fixed true measure, t, and a variable error, e, such that $x = t + e$. There is importance of noting and maintaining the distinction between random and systematic sources of error when working with this model, because an analysis of reliability is only concerned with random errors and does not address errors that are observed to be consistent, predictable differences in the true scores.

In order to compare decomposition to direct assessment, the probability of the target event can be rewritten as $a = \sum_{i=1}^{n} c_i b_i$ with the direct assessment, denoted a, the components of the decomposition estimate, denoted a', the conditional probabilities of the background events, denoted $\Pr(A|B_i) = c_i$, and marginal probabilities of the background events, denoted $\Pr(B_i) = b_i$. Now the measurement model can be applied both to the direct assessment and the decomposition estimate components, such that for the direct assessment, $a' = \alpha' + \delta'$, where α' is the true score and δ' is the random error with its standard error of measurement (SEM) denoted by σ_0. In addition, marginal probabilities of

the background events are $b_i = \beta_i + \delta_i$, with each SEM denoted by σ_i, and conditional probabilities of the background events are $c_i = \gamma_i + \varepsilon_i$, with each SEM denoted by τ_i. Ravinder *et al* (1988) do not accept the additional assumption provided by Wallsten and Budescu, that random errors are uncorrelated, because they argue that marginal probabilities cannot be completely independent since they must sum to one to be coherent and errors in the elicitation of the conditional probabilities could be correlated due to the anchoring and adjusting heuristic discussed in chapter three. Thus, correlations for the errors in the marginal elicitations, denoted φ_{ij}, are accepted in lieu of independence and, in addition, the errors in the condition elicitations may be correlated, denoted by ρ_{ij}.

The set of equations previously noted can be used to obtain the expected value and variance of the decomposition estimate (Ravinder *et al*, 1988). The expected value of the decomposition estimate, $E(a)$, can be derived from $E(a) = \sum_{i=1}^{n} E(c_i b_i)$, and with the assumption of independence between the errors in the c_i and b_i terms, $E(a) = \sum_{i=1}^{n} E(\gamma_i \beta_i)$. If the estimates of the various probabilities are unbiased, because of the assumption that the random error terms have expected values of 0, the expected value of the decomposition estimate (previous equation) should be equal to the expected value of the direct assessment, earlier equated *a'*. The decomposition estimate error variance is denoted by, $Var(a) = \sigma_d^2$, and with the help of the adjusted target event equation, the variance can be rewritten as the sum of a set of covariances, $\sigma_d^2 = Var(\sum_{i=1}^{n} c_i b_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(c_i b_i, c_j b_j)$.

Using the idea that c_i and b_i terms have independent errors, a general formula for the decomposition error is produced:

$$\sigma_d^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(\gamma_i\sigma_i)(\gamma_j\sigma_j)\varphi_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n}(\beta_i\tau_i)(\beta_j\tau_j)\rho_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n}(\sigma_i\tau_i)(\sigma_j\tau_j)\varphi_{ij}\rho_{ij}.$$

Ravinder *et al* address the question of the relative sizes of the errors associated with direct elicitation and decomposition with a comparison of the SEM of direct elicitation to the SEM of decomposition. The percent change due to decomposition, denoted: $\pi = \frac{\sigma_d - \sigma_0}{\sigma_0} = \frac{\sigma_d}{\sigma_0} - 1$, can express the relative size of the measurement errors. If $\pi = 0$, then there are identical levels of measurement error for both techniques. If a negative value is produced, there is an improvement due to decomposition, and a positive value indicates that "decomposition is inferior to direct assessment" (Ravinder *et al*, 1988 p.195). Deriving upper and lower bounds for π can be extremely informative when comparing the amount of error associated with each technique. In general, decomposition can considerably reduce the potential error, especially if the component probabilities can be more precisely assessed than the target event (direct assessment).

Ravinder *et al* (1988) conclude with some recommendations about structuring probability encoding tasks that will minimize inconsistency: decomposition can reduce random errors associated with probability encodings; error reduction has an upward limit and assessing additional probabilities will not be necessary to obtain further reduction; the conditioning events marginal probabilities should be selected to be equal; conditional probabilities that cover an extremely wide range of values should be avoided; the

component probability assessment errors should be as small as possible; and the analyst should try to maintain independence of the component assessments. Finally, a decomposition approach is simple (conceptually), is implemented easily, and apparently has great potential for error reduction.

Howard (1989) continues with the idea of decomposition in his work. He notes that information is generally received fragmentally by an analyst when it comes from an expert or a group of experts, and the daunting task of gathering and coordinating these fragments can be lessened with the use of knowledge maps. A knowledge map is centered on the concept of influence diagrams and captures the relevant knowledge of an event. An influence diagram represents the actions a decision maker may take and the information possessed at the time of the decision by way of decision nodes (boxes representing an action to be taken), chance nodes (circles representing an uncertain event), and arrows (showing connectivity between the events). A knowledge map is also another method for decomposing a complex uncertain event into workable and more easily understood units.

Howard (1989) uses the Tversky and Kahneman (1973) study of Tom W. to prove his point on how knowledge maps (decomposition) can aid in correcting the representative heuristic. As previously reviewed in chapter two, the Tom W. study presents a high school description of Tom implying that he is introverted, mechanically minded, lacking in creativity, and well organized to a group of students who are supposed to assess the likelihood that Tom would be enrolled in a specific field of graduate study. Generally, the assessors relate Tom to being a "nerd" and subsequently place him into the field of computer science or engineering without regard to the relative enrollment populations of

each field of study. Howard (1989) first suggests decomposing this problem into assessing the likelihood that a student would be in each field of study and the likelihood of having a report such as Tom's given he is in each field of study. Next the assessors are to multiply the two likelihoods together and then normalize to yield a posterior distribution on the different fields. This first step provides an improvement step in understanding but still does not capture all that is relevant to the second probability, that a report like Tom's would be in a given field. Howard believes that the remaining relevant information can be captured with three considerations: is Tom's personality now different than in high school (when the report was completed), are there any validity issues with the report itself, and is personality a good indicator of field of study. Figure 5.2, provided by Howard (1989, p.912), is the corresponding knowledge map that captures these three considerations.

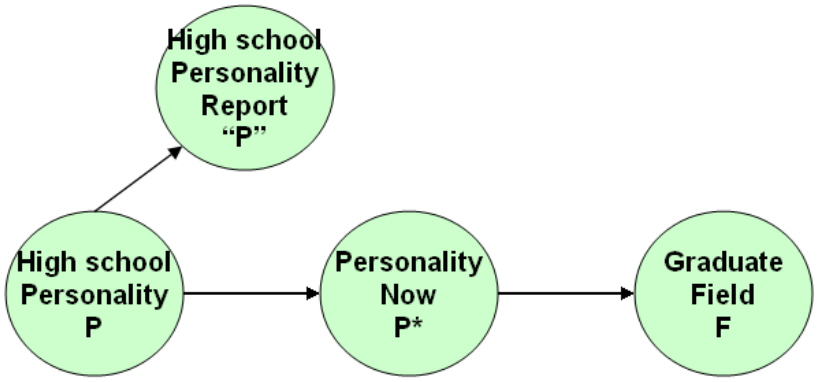

Figure 5.2 Tom W. knowledge map, Howard (1989)

Despite the fact that this knowledge map encompasses all the necessary information for determining the likelihood, it does not reflect the notion that an individual might not feel comfortable trying to assess the relevance of present personality to graduate field because they do not think about characterizing the personalities of graduate students.

Yet assessing the probability distribution on field of study and what is known about the personalities of graduate students in each field is probably more comfortable for most assessors, because people do think of graduate students by field. Thus, the redundancy map in Figure 5.3, provided by Howard (1989, p.913), is created to represent this more

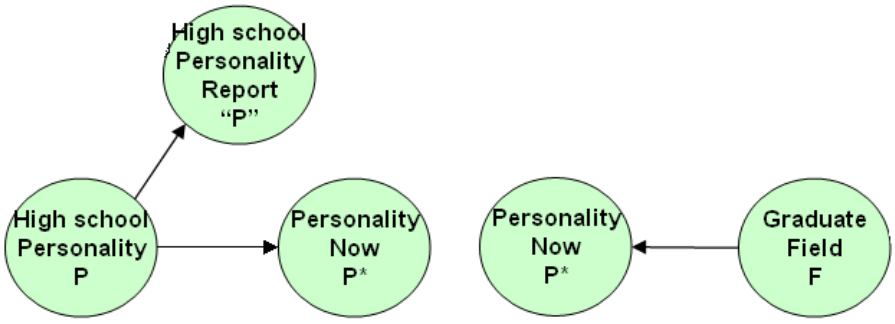

Figure 5.3 Tom W. redundancy map, Howard (1989)

comfortable way of thinking. The assessments for the original knowledge map (figure 5.2) can now be determined by using the new assessments from figure 5.3 iteratively until the Personality Now distributions agree and then performing the Bayesian reversal of the arrow from Graduate Field to Personality Now.

"A knowledge map represents a possible assessment order for the joint distribution on the uncertain quantities (events) it contains", and in some instances the decision maker is only interested in the joint distribution of a few of the events (Howard, 1989 p.918). For these such instances constructing a disjoint knowledge map can simplify the assessment. The corresponding example provided by Howard (1989) is for a subject to be interested in assessing a probability distribution on the income of the next person to enter the room. To aid in thinking about this particular assessment, the subject might want to consider the amount of education the person has had and even the age of the individual. With these additional events however comes additional assessments and the subject is only interested

in the distribution on income, thus the subject would want to draw the corresponding disjoint knowledge map represented in figure 5.4, provided by Howard (1989, p.920). The subject can now assess a distribution on age and another on education given age, multiply them, and then sum age to obtain a distribution on education. Finally, the desired distribution on income can be obtained by substituting this distribution into the equation for deriving education (point of large arrow), $\{e \mid \&\} = \int_a \{e \mid a\&\}\{a \mid \&\}$, multiplying by an assessed distribution on income given education, and summing over education. As Howard suggests, creating disjoint knowledge maps can reduce the difficult task of assessment to manageable pieces. One problem with disjoint knowledge maps however, is that they only represent one point in the spectrum and they do not provide all the consistency checks and inferences of a complete knowledge map. Yet overall, knowledge maps permit assessing beyond direct assessment, and the use of redundant and disjoint knowledge maps "can often reconcile the needs of consistency and convenience" (Howard, 1989 p.921).

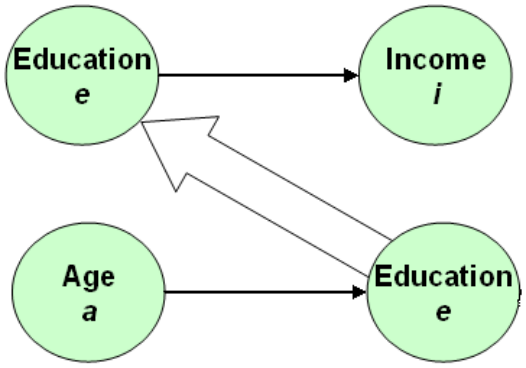

Figure 5.4 Income disjoint knowledge map, Howard (1989)

Moskowitz and Sarin (1983) are more focused on the elicitation and additional problems associated with the elicitation of conditional probabilities instead of marginal probabilities because the elicitation of conditional probabilities presents extra problems that are not encountered in marginal probability assessment. In their article, they present these difficulties (produced due to the heuristics described in chapter 3) and provide

suggestions for improvement in the assessment procedure such as the introduction of a judgmental aid. They suggest that these additional assessment difficulties are affected by the causal versus diagnostic and positive versus negative relationships of the assessed events. It was determined that the probability judgments are larger if the relationship is perceived to be causal, in assessing $p(B \mid A)$, A is perceived as the cause of the occurrence or nonoccurrence of event B, rather than diagnostic, if B is perceived as the cause of A. This is because individuals perceive a causal relation as more informative. In addition, a positive relationship, the knowledge of A occurring should increase the probability of B occurring, between the events led to larger conditional probability assignments than a negative relationship because subjects also felt that a positive relation was more informative. Normatively speaking however, neither the causality/diagnostic relationship nor the positive/negative relationship should influence the conditional probability assignments. These effects are noted in the study reviewed below.

Moskowitz and Sarin (1983) performed a case study in which each subject, an undergraduate enrolled in managerial statistics with a knowledge of the concept probabilities, assumed the role of a bank lending officer who was given the base rate probabilities of delinquency, $p(D)$, good or bad credit probabilities $p(G)$ and $p(B)$ respectively, and was to assess the conditional probabilities $p(D|B)$, $p(B|D)$, and $p(D|G)$. The subjects first specified their perceptions of the relationship strengths between delinquency and credit rating on a scale, they then assessed conditional probabilities for 15 questions partitioned into 5 groupings with reasoning as to how they arrived at their assessments, the subjects were then given a Joint Probability Table (JPT) with the marginal

event probabilities included as an assessment aid, with which the subjects reassessed several of the previous questions, and finally they rated the JPT as an assessment aid on a scale with descriptions as to how it helped.

Each group's questions were designed to test the consistency with respect to the probability axioms that must be satisfied for a conditional probability, say $p(B \mid A)$: $p(B \mid A) \le 1$, $p(B \mid A) \le \frac{p(B)}{p(A)}$, $p(B \mid A) \ge \frac{p(B) + p(A) - 1}{p(A)}$, and $p(B \mid A) \ge 0$. The results showed that a number of violations were observed with possible causes being that the subjects ignored the based rate frequency of the events (the base rate bias) and the causal versus diagnostic and positive versus negative relationships. For this study, where the event of interest is delinquency *D*, if the event *B* is bad credit rating then the event *B* is a causal datum, and conversely, the event *B* is a diagnostic datum if *D* is perceived as the cause of *B*. Thus, if $B \rightarrow D$ and $p(D) = Cp(B)$, then $p(D|B) > Cp(B|D)$, where "→" denotes the direction of causality, *C* is a constant, and $p(D|B)$ and $p(B|D)$ are assessed directly. Once again, the causal relationship led to more revisions and higher conditional assessments than did the diagnostic and there shouldn't be a difference. For this study, a positive relationship between events *B* and *D* implies that knowledge of *B* should increase the probability of *D* occurring and therefore $p(D \mid B) > \frac{p(D) - p(D \mid G)p(G)}{p(B)}$, where *G* denotes a good credit rating, and $p(D|B)$ and $p(D|G)$ are directly assessed probabilities.

With the addition of the JPT, the subjects tended to have more confidence in the marginal probability assessments and were more likely to revise their conditional probability assessments. The results from the study suggested that the use of the JPT

drastically improved the consistency, reduced the frequency of the violations of the conditions, and also lessened the effects of the causality/diagnostic and positive/negative relationships. However, if the JPT is not used properly, the conditional probabilities can be inconsistent. Generally though, the subjects indicated that the JPT was in fact a useful tool.

In addition, Moskowitz and Sarin (1983) prove that the task of assessing higher order probabilities can be thwarted by appropriate bounding. They suggest that the assessments of marginal and pairwise conditional probabilities are sufficient because they alone place a very tight bound on the scenario probabilities. However, if they do not provide a tight enough bound and further precision on the estimation of the scenario probabilities is needed, higher order probabilities must be assessed. The example provided by Moskowitz and Sarin is the consideration of n binary events, e_1, e_2,…, e_n, thus there exists $N = 2^n$ combinations of occurrence or nonoccurrence, and the assessment of probabilities would require $N - 1$ judgments from an expert. If only the marginal and pairwise conditional probabilities are known, then the lower and upper bounds of the probabilities can be induced by the linear programs: Min x_i s.t. $x \in X$ and Max x_i s.t. $x \in X$ respectively where only p_i and $p_{ij} = p(e_i / e_j) \cdot p_i$ for $i = 1$ to $n, j > i,$ are specified. One basic possible alternative for this formula is that with simple modification, the bounds on a subset of scenarios can be easily determined. The results of one simulation study indicated that only having only the marginal and pairwise conditional probabilities may be sufficient in many applications. However, in some situations the marginal and pairwise conditional probabilities may not provide tight enough bounds and higher order probabilities will need to be assessed, meaning, in the constraint set X the p_{ijk}'s will need to

be assessed. The example provided is the consideration of a 3 event case, p_{123}, where the bounds are obtained by solving eight equations in the constraint set X and by restricting each $x_i \geq 0$. The bounds that facilitate the assessment of higher order probabilities are:

$p_{123} \leq p_{12}, p_{13}, p_{23}, 1-(p_1+p_2+p_3)+(p_{12}+p_{13}+p_{23})$,

$p_{123} \geq (p_{12}+p_{13}-p_1),(p_{12}+p_{23}-p_2),(p_{13}+p_{23}-p_3),0$, so that if $p_1 = 0.5, p_2 = 0.2, p_3 = 0.9, p_{12} = 0.1, p_{13} = 0.45$, and $p_{23} = 0.18$, then $0.08 \leq p_{123} \leq 0.1$. "The mean values of the bounds in the simulation study were considerably smaller because the bounds decrease rapidly with the variations in p_i's and p_{ij}'s" (Moskowitz & Sarin, 1983 p.747).

These were only some of the methods used to improve assessments provided by experts during probability elicitation for decision analysis. The subsequent chapter will review scoring rules, which is a method used to encourage honesty from subjects who are providing assessments.

CHAPTER 6 Scoring Rules

This chapter will review the concepts of scoring rules and how they can be applied to encourage more accurate and honest forecasting. Most of the literature uses weather forecasting as their research for measuring improvement with these methods. Matheson and Winkler (1976) will provide the methods used for scoring discrete and continuous probability distributions. A continuation of the subject involving utilities is presented by Kadane and Winkler (1988). The use of asymmetric scoring rules (versus symmetric) is described in detail by Winkler (1994). The use of scoring rules in coordination with incentive plans for big business is provided by Sarin and Winkler (1980). And finally, Lichtendahl and Winkler (2007) show how competition between forecasters provides different uses of scoring rules.

Scoring rules are put in place to encourage a subject to reveal his true opinion and thus make his probability assessments correspond with his judgment (Matheson & Winkler, 1976). Scoring rules can be used to either evaluate the assessor by measuring the goodness of the assessments (an ex post sense), or aid in the elicitation process by encouraging the assessor to be careful in making assessments (an ex ante sense). Scoring rules typically are set to overcome motivational biases, anchoring, and overconfidence.

Matheson and Winkler (1976) explain how to perform the generation of a scoring rule for a binary probability assessment. Consider the event E and a subject who assigns probability p to the occurrence of the event, but when asked to reveal his assessment provides a probability r which may not be equal to p. A scoring rule $S(r)$ provides the

subject a payoff $S(r) = S_1(r)$ if the event occurs and $S(r) = S_2(r)$ if the event does not occur. Thus the subject's expected payoff for the binary situation is

$E(S(r)) = pS_1(r) + (1-p)S_2(r)$ and the scoring rule is strictly proper if $E(S(p)) > E(S(r))$ for $r \neq p$. So for a subject to maximize the expected score, the reported forecast r should be equal to the probability p, meaning the subject should be honest in his reporting (Winkler, 1994).

Averages can be determined when using scoring rules to perform ex post evaluations (versus using them during the elicitation process to promote honesty) (Winkler, 1994). Winkler explains that over a series of assessments procedures with the same value r given as the forecast, the average score is $\bar{S}_f(r) = fS_1(r) + (1-f)S_2(r)$ where f is the relative frequency of E on those occasions and is strictly proper if $\bar{S}_f(f) > \bar{S}_f(r)$ for $r \neq f$. "That is, over the set of occasions with forecast value r, $\bar{S}_f(r)$ is maximized when the probability forecast r equals the relative frequency f" and when $f = r$, the forecaster is perfectly calibrated (Winkler, 1994 p.1396). As a result for forecasters who are well calibrated the equation can be simplified to $\bar{S}(r) = rS_1(r) + (1-r)S_2(r)$. Sharpness in probabilities is rewarded with strictly proper scoring rules and almost all scoring rules used in practice are symmetric in the sense that $S_1(r) = S_2(1-r)$ for all $r \in [0,1]$. This implies that $\bar{S}(r)$ is symmetric with a minimum value when r = ½ and a maximum value when r = 0 or 1, and for that reason the average scores are smallest for forecasts around ½ and biggest for forecasts around 0 or 1, and thereby gratifying greater sharpness through higher average scores. For the instance when

the same value of r is not given over different occasions, $g(r)$ denotes the proportion of occasions that the value r is used in a series of forecasts and $f(r)$ denotes the relative frequency of occurrence of the event on the occasions for which a forecast of r is given and thus the overall average score is now

$\bar{S} = \sum_r \bar{S}_{f(r)}(r)g(r) = \sum_r \{f(r)S_1(r) + [1 - f(r)]S_2(r)\}g(r)$. Now for $f(r) = r$ for all r, or perfect calibration, $\bar{S} = \sum_r [1 - r(1-r)]g(r)$ and sharpness is again rewarded via $\bar{S}$ because a distribution g of forecasts r with values of r close to 0 or 1 given more often and values close to ½ not given as much will yield a higher average score.

Scoring rules can also be generalized to any discrete probability distribution (Matheson & Winkler, 1976). To do so, let E_i represent the ith event (or ith value of a random variable), where $i \in I$ and I is finite or countably infinite. Also, let p_i and r_i correspond to previous p and r in the binary situation, and lastly suppose $S(r_1, r_2, \ldots)$ provides the subject's payoff $S_j(r_1, r_2, \ldots)$ if E_j occurs.

Then, $E(S(r_1, r_2, ...)) = \sum_{j \in I} p_j S_j(r_1, r_2, ...)$, and S is strictly proper if

$E(S(p_1, p_2, ...)) > E(S(r_1, r_2, ...))$ when $r_i \neq p_i$ for any $i \in I$.

There are several forms of strictly proper scoring rules that have been developed and the three most encountered examples are the quadratic, $S_j(r_1, r_2, ...) = 2r_j - \sum_{i \in I} r_i^2$, the logarithmic, $S_j(r_1, r_2, ...) = \log r_j$, and the spherical, $S_j(r_1, r_2, ...) = r_j / (\sum_{i \in I} r_i^2)^{1/2}$ scoring rules. Matheson and Winkler (1976) note that scoring rules can easily be extended to a

continuous state by limiting arguments. Thus, if x is the value of the variable of interest and $r(x)$ is the density function assessed by the subject, the continuous quadratic function is $S(r(x)) = 2r(x) - \int_{-\infty}^{\infty} r^2(x)dx$, the logarithmic functions is $S(r(x)) = \log r(x)$, and the spherical function is $S(r(x)) = r(x)/(\int_{-\infty}^{\infty} r^2(x)dx)^{1/2}$.

The Brier score is an example of a quadratic scoring rule that is used to evaluate probabilistic weather forecasts (Winkler, 1994). If considering the quadratic rule $S_1(r) = 1-(1-r)^2$ and $S_2(r) = 1-r^2$, the average score is expressed as $\overline{S}_f(r) = 1 - f(1-f) - (r-f)^2$ with the last term rewarding calibration by penalizing deviations from perfect calibration. For perfect calibrated forecasters the average score would then be $\overline{S}(r) = 1 - r(1-r)$, with the last term penalizing assessments that move from 0 or 1 towards ½ and thereby rewarding sharpness. For the instance when the same value of r is not given over different occasions, $g(r)$ denotes the proportion of occasions that the value r is used in a series of forecasts and $f(r)$ denotes the relative frequency of occurrence of the event on the occasions for which a forecast of r is given and thus the overall quadratic average score is now $\overline{S} = \sum_r \{1 - f(r)[1-f(r)] - [r-f(r)]^2\} g(r)$. Now for $f(r) = r$ for all r, or perfect calibration, $\overline{S} = \sum_r [1 - r(1-r)]g(r)$.

Kadane and Winkler (1988) are focused on the separation of probability elicitation from utilities and they review three elicitation procedures, one of which is the use of scoring rules. Kadane and Winkler focus on an event A, with $g(f \mid A)$ and

$g(f \mid \overline{A})$ representing the probability distributions of the expert's fortune given *A* and it's complement $\overline{A}$, with *U* denoting the expert's utility function for his or her fortune. Kadane and Winkler (1988) also present the no-stakes condition which is built upon the instance when no part of the expert's fortune is contingent on whether *A* occurs or not, thus a condition where $c = 1$ meaning, $g(f \mid A) = g(f \mid \overline{A})$ for all *f*. "Independence of *f* and the events or variables for which probabilities are being elicited can be thought of in terms of the expert having no stakes in these events or variables other than stakes that might be created as part of the elicitation process through lotteries or other devices" (Kadane & Winkler, 1988 p.358).

The scoring rule elicitation method is the second method to be reviewed by Kadane and Winkler, who tell us that the expert will receive a payoff equal to a score *S* from a scoring rule that is strictly proper if and only if the reported value is exactly the expert's probability, denoted π. They use the quadratic strictly proper scoring rule, $S = -r(x - p)^2$, where *p* is the expert's stated probability of *A*, *r* is a positive constant, and *x* is an indicator variable corresponding to the occurrence (*x*=1) or nonoccurrence (*x*=0) of *A*. The expert's expected utility as a function of *p* would therefore be:

$EU(p) = \pi \int U[f - r(1-p)^2]g(f \mid A)df + (1-\pi)\int U(f - rp^2)g(f \mid \overline{A})df$. The expert would then set the derivative of *EU*(*p*), with respect to *p,* equal to 0 to maximize the expected utility. Thus,

$\pi(1-p)\int U'[f - r(1-p)^2]g(f \mid A)df + (1-\pi)p\int U'(f - rp^2)g(f \mid \overline{A})df = 0$ would

simplify to $p/(1-p) = c\pi/(1-\pi)$ with $c = \dfrac{\int U'[f - r(1-p)^2]g(f \mid A)df}{\int U'(f - rp^2)g(f \mid \overline{A})df}$. Now if $c = 1$, meaning U is linear, then the expert's stated probability p will equal π. However, without linearity, such as for a risk averse individual whose U' is strictly decreasing, the no stakes condition will not be sufficient for $c = 1$ with scoring rules. Consequently, a risk averse subject who satisfies $g(f \mid A) = g(f \mid \overline{A})$ will have $c < 1$ when $\pi > ½$ and $c > 1$ when $\pi < ½$ resulting in p moving away from π toward ½. And for a risk seeking subject, the movement would be in the opposite direction, moving toward 0 when $\pi < ½$ and toward 1 when $\pi > ½$ with the possibility of moving all the way to $p = 0$ or 1 for extremists.

When the payoff from the scoring rule is of little consequence or when r approaches 0, the no-stakes condition is sufficient for $c = 1$ to hold approximately. The limit, $\lim_{r\to 0} c = c_0 = \dfrac{\int U'(f)g(f \mid A)df}{\int U'(f)g(f \mid \overline{A})df}$, will equal 1 if the no-stakes condition holds.

Expanding c in a Taylor series in r around 0, Kadane and Winkler (1988) find that $c = c_0 + rc_1 + O(r^2)$, where

$c_1 = \left[c_0 - (1-p)^2 \dfrac{\int U''(f)g(f \mid A)df}{\int U'(f)g(f \mid \overline{A})df} + p^2 \dfrac{\int U''(f)g(f \mid A)df}{\int U'(f)g(f \mid \overline{A})df} \right]$. Therefore from these two equations, $c = c_0\{1 + r[(1-p)^2 E_{h_A}(w) - p^2 E_{h_A}(w)]\} + O(r^2)$, with

$w(f) = -U''(f)/U'(f)$, $h_A(f) = U'(f)g(f \mid A)/\int U'(f)g(f \mid A)df$, and

$h_{\overline{A}}(f) = U'(f)g(f \mid \overline{A})/\int U'(f)g(f \mid \overline{A})df$. When U is exponential with $w(f) = w > 0$,

$c = c_0[1 + r(1-2p)w] + O(r^2)$ and when r is small, $c > c_0$ if $\pi < ½$ and $c < c_0$ if $\pi > ½$. Thus, the no-stakes condition by itself is not sufficient for $c = 1$ making it necessary to have stronger assumptions for obtainment. In addition, c can differ substantially from 1 if the no-stakes condition is not satisfied.

Winkler (1994) presents an argument that symmetric rules aren't always an appropriate reward structure and he then presents a set of tailor-able asymmetric rules. When comparing forecasters the scores obtained are meant to compare the skill of the forecaster but instead they also depend on the current situation the forecaster is in. For example, forecasters who are located in very dry climates can be very refined because they can give precipitation predictions close to zero on multiple days, whereas a forecaster located in an area with a higher annual rainfall or unpredictable weather patterns may not be able to predict precipitation likelihood close to zero or one on very many days. Winkler (1994) develops a set of "skill scores" to help produce an average score that reflects the ability of the forecaster instead of a combination of the forecaster's skill and current situation. These work to neutralize the situation's contribution by comparing a forecaster's average score to the average score that a forecasting scheme, i.e. climatology, would obtain. The most common weather forecasting skill score is the percentage improvement over climatology in the average score, thus corresponding to a scoring rule S, $Skill = (\overline{S} - \overline{S}_{Cl}) / \overline{S}_{Cl}$ where $\overline{S}_{Cl}$ is the average score for climatology. The skill score does have two shortcomings; even if S is strictly proper, the skill score is not and will therefore not necessarily reward calibration and sharpness as would a strictly proper rule, and a linear transformation of S can alter the values of the skill score which limits the skill score

from being viewed as a standardized measure. However, if differences in average scores are considered instead of percentage improvement, the strictly proper nature of *S* will be maintained by the skill score, e.g. $Skill = \overline{S} - \overline{S}_{Cl}$.

The main goal of Winkler (1994) is to develop strictly proper scoring rules that are standardized, that will differ from previous rules because they are not symmetric (i.e. they will not satisfy $S_1(r) = S_2(1-r)$), and will have the capability to represent an evaluator's judgments about the associated difficulty of each forecasting situation. The benchmark to measure by, denoted *c*, could be a base rate such as climatology or based on a utility of forecasts in a decision problem, and when the forecast of *c* is given, unless its *c* = 0.5, the scoring rule must be asymmetric to achieve a minimum average score. To generate a family of strictly proper asymmetric scoring rules, define for any symmetric rule *S* and $c \in (0,1)$, $S^*(r) = [S(r) - S(c)]/T(c)$ where $T(c) = \{S_1(1) - S_1(c)$ if $r \geq c$, and $S_2(0) - S_2(c)$ if $r \leq c$. Therefore *S** is strictly proper because $E_p[S^*(p)] > E_p[S^*(r)]$ for $r \neq p$ since *T* is positive and independent of *r* and *E*. As before with any strictly proper scoring rule, *S** rewards calibration, and in addition, $\overline{S}^*(r) = [\overline{S}(r) - rS_1(c) - (1-r)S_2(c)]/T(c)$. So for well calibrated forecasters, $\overline{S}^*(1) = \overline{S}^*(0) = 1$ and additionally $\overline{S}^*(c) = 0$ because $\overline{S}(1) = S_1(1)$ and $\overline{S}(0) = S_2(0)$. The overall average score $\overline{S}^*$ is $\overline{S}^* = \sum g(r)[\overline{S}(r) - rS_1(c) - (1-r)S_2(c)]/T(c)$. A well calibrated forecaster with a distribution *g*(*r*) who provides *r* values approaching zero or one more frequently than values approaching *c* will obtain a greater average score. The asymmetric scoring rule

rewards skill whilst providing a measure of the skill when *c* is considered a benchmark probability and skill is considered a capability of being more discriminatory than *c*.

An example provided by Winkler (1994) is the consideration of the quadratic rule defined by $S_1(r) = 1-(1-r)^2$ and $S_2(r) = 1-r^2$. For this, $S_1^*(r) = [(1-c)^2 - (1-r)^2]/T(c)$ and $S_2^*(r) = (c^2 - r^2)/T(c)$ where $T(c) = c^2$ if $r \le c$ and $(1-c)^2$ if $r \ge c$, thus using one quadratic function when $r \le c$ and a different when $r \ge c$. For well calibrated forecasters, $\bar{S}^*(r) = (r-c)^2/T(c)$. Figure 6.1 shows $S_1^*(r)$, $S_2^*(r)$, and $\bar{S}^*(r)$ for the only value in which *S** is symmetric, *c* = 0.5, and figure 6.2, provided by Winkler (1994, p.1399,1400), shows $S_1^*(r)$, $S_2^*(r)$, and $\bar{S}^*(r)$ for *c* = 0.2. Overall, the reasoning for having an asymmetric scoring rule is to have the ability to measure forecasting skill efficiently and so it is logical to have the smallest value for $\bar{S}(r)$ occur when *r* is the least skilled forecast.

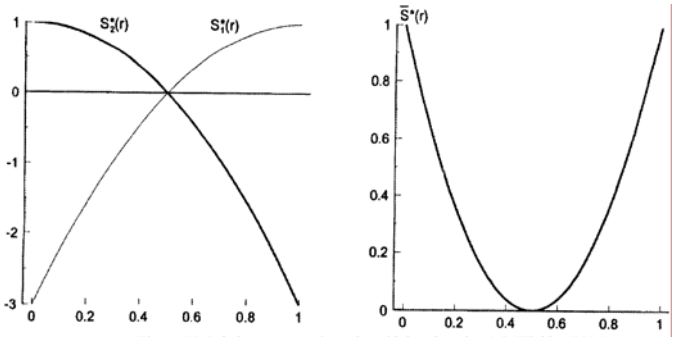

Figure 6.1 Strictly proper scoring rules with benchmark = 0.5, Winkler (1994)

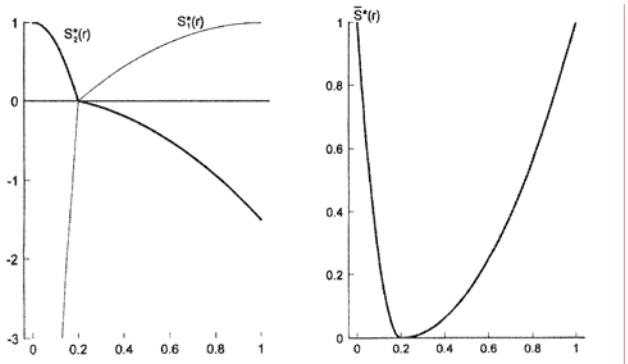

Figure 6.2 Strictly proper scoring rules with benchmark = 0.2, Winkler (1994)

Sarin and Winkler (1980) are focused on incentive plans and how they can be used to better motivate managers to forecast accurately and honestly, work harder to achieve the goals set forth, and to act on behalf of the company's best interests. Sarin and Winkler (1980) review two types of goal based incentive plans, deterministic and probabilistic. Deterministic goal based incentive plans involve rewards that are determined by performance in relation to the goals set forth by the managers in the planning process, as opposed to rewards based solely on the performance of the managers. Probabilistic goal based incentive plans involve decision making and planning when uncertainty is involved and that uncertainty is presented in the model with probabilistic terms. Sarin and Winkler show how to design an incentive plan encouraging managers to report probabilistic goals honestly. Here is where scoring rules become a topic of discussion, because using scoring rules as incentive plans should make it easier to obtain honest reporting of uncertainty about performance.

For decision making with uncertainty, the manager should provide a probability distribution reflecting his uncertainty of the situation x, denoted $x = (x_1,\ldots, x_n)$ instead of a single point estimate g_i of x_i for $i = 1,\ldots,n$. The reported distribution is represented by $q(x)$ with Sarin and Winkler simplifying by assuming the distribution to be discrete rather than continuous and the marginal distribution of x_i is denoted by $q_i(x_i)$ which is viewed as a "probabilistic goal." Achievement, denoted a_i^*, corresponds to $[q_i(x_i), x_i]$. The reward for probability assessors that encourages honest probability reporting is expressed in terms of a scoring rule $s_i[q_i(x_i), x_i]$ and is strictly proper if "it encourages honesty in the sense that

an assessor's expected score is maximized by honest reporting" (Sarin & Winkler, 1980 p.1139). Basically, if $r_i(x_i)$ represents the managers judgments regarding x_i and $q_i(x_i)$ is the managers reported distribution for x_i and $q_i \equiv r_i$, then the reporting is honest. Therefore a scoring rule s_i is strictly proper given the managers expected score $\sum s_i[q_i(x_i), x_i]r_i(x_i)$ is maximized if $q_i \equiv r_i$ over all possible values of x_i.

Though scoring rules encourage honest reporting they do not motivate managers to work harder. To do so, the performance function should depend on $s_i[q_i(x_i), x_i]$ and $p_{1i}(x_i)$, a function that represents preferences for values of x_i. Therefore the performance function p_i is a function of $q_i(x_i)$ and x_i through s_i and p_{1i} and the following condition enables the decomposition of p_i (note the *i* subscript is only dropped for convenience). The *condition* states first; "if all managers have identical evaluations $s(q,x)$ on a strictly proper scoring rule s, then the decision maker's evaluation of performance difference between any two managers should not depend on $s(q,x)$." And second; "if all managers have identical values of x, then the decision maker's evaluation of performance difference between any two managers should not depend on x" (Sarin & Winkler, 1980 p.1139). The *theorem* declares that the *condition* is satisfied if and only if $p(q,x) = \lambda_1 s(q,x) + \lambda_2 p_1(x)$ where λ_1 and λ_2 are scaling constants with $\lambda_1, \lambda_2 > 0$ and $\lambda_1 + \lambda_2 = 1$ and s and p_1 are scaled between 0 and 1 for their worst and best values respectively. With respect to honesty reporting motivation, for this theorem, "any positive linear transformation of s is a strictly proper scoring rule, where the coefficient of the transformation may depend upon x but not upon q, thus $p(q,x)$ is a strictly proper scoring rule" (Sarin & Winkler, 1980 p.1139).

The example provided by Sarin and Winkler employs the supposition that x represents the subsequent month's sales in thousands of units, where it is felt that $7.5 \leq x \leq 10.5$, and the manager is asked to make a probability assessments for values of x rounded to nearest thousand, so that $q(x)$ consists of the manager's probabilities for $x = 8$, $x = 9$, and $x = 10$, respectively. The decision maker has chosen a quadratic scoring rule, $s(q,x) = \left[2q(x) - \sum_{y=8}^{10} q^2(y) + 1 \right] / 2$ and assesses $p_1(x) = (x-8)^2 / 4$. The best possible achievement corresponds to $s = 1$ and $p_1 = 1$, implying that $x = 10$ and $q = (0, 0, 1)$, and the worst possible achievement corresponds to $s = 0$ and $p_1 = 0$, implying that $x = 8$ and $q = (0, 0, 1)$ or $(0, 1, 0)$. For the assessment of λ_1 and λ_2, the manager should assign a performance score p to an intermediate achievement, for instance $q = (1, 0, 0)$ and $x = 8$ because the level of sales is low but predictably so by the manager, enabling the decision maker to plan accordingly. Assessing $p[q = (1, 0, 0), x = 8] = \lambda_1(1) + \lambda_2(0) = 0.2$, yields $\lambda_1 = 0.2$, implying that that a high sales level which is assigned a zero probability [e.g., $q = (1, 0, 0)$, $x = 10$] should have a performance score of $\lambda_1(0) + \lambda_2(1) = 0.8$. Table 6.1, provided by Sarin and Winkler (1980, p.1140), provides some select performance scores of the example. "Note that for a given q, p increases as x increases, while for a given x, p increases as q becomes 'more accurate'" (Sarin &

Performance Scores With Probabilistic Goals

q	x	s	p_1	p
(0.1, 0.8, 0.1)	8	0.27	0	0.054
(0.1, 0.8, 0.1)	9	0.97	0.25	0.394
(0.1, 0.8, 0.1)	10	0.27	1	0.854
(0.1, 0.4, 0.5)	8	0.39	0	0.078
(0.1, 0.4, 0.5)	9	0.69	0.25	0.338
(0.1, 0.4, 0.5)	10	0.79	1	0.958
(0.1, 0.1, 0.8)	8	0.27	0	0.054
(0.1, 0.1, 0.8)	9	0.27	0.25	0.254
(0.1, 0.1, 0.8)	10	0.97	1	0.994

Table 6.1 Performance score with probabilistic goal, Sarin & Winkler (1980)

Winkler, 1980 p.1140). In order for incentive plans based on probabilistic goals to work, the manager's performance and the relation of the performance to the reported probabilities must be taken into consideration.

The consideration of the action of competition affecting the forecasters reported probabilities is covered by Lichtendahl and Winkler (2007). Lichtendahl and Winkler define the situation as; there exists an event *E*, a consultation of two forecasters to elicit the probability of *E* from each, rewarding forecaster *i* (*i* = 1, 2) via a score S_i from a strictly proper scoring rule *S* that yields $S_E(\beta_i)$ if *E* occurs and $S_{E'}(\beta_i)$ if its complement *E'* occurs, where β_i is forecaster *i*'s reported probability of *E*. Forecaster *i*'s probability for event *E* is denoted by p_i and *i*'s reporting strategy is represented by $\beta_i = b_i(p_i)$ and the scoring rule is normalized so that $0 \le S_i \le 1$, where a higher score is better. They use the normalized quadratic score for their examples, $S_E(\beta_i) = 1 - (1 - \beta_i)^2$ and $S_{E'}(\beta_i) = 1 - \beta_i^2$.

There is an additional assumption that forecaster *i*'s preferences can be represented by a utility function that is additive in relative performance R_i and score S_i, $u_i(R_i, S_i) = w_i R_i + (1 - w_i)S_i$, where $0 \le w_i \le 1$ and $R_i = 0(0.5)1$ if S_i <(=)> $S_j, j \ne i$ (i.e. if *i*'s score is worse than, equal to, or better than the score of the other forecaster), and w_i represents forecaster *i*'s preference trade-off between relative performance and score. Because the relative performance term does not depend on scoring rule that is used, if *E* occurs, the forecaster with the higher β_i has the better performance, with $R_i = 1$ (a tie can occur at $R_i = 0.5$) and only the second term on the right side is specific scoring rule

dependent. There are discontinuities associated with the utility function, through the relative performance term, that are however justifiable considerations such as career concerns (i.e. promotion or maintaining a current position). Nonetheless, the absolute performance term is a continuous function of β_i, and the discontinuity impact is dependent on the size of w_i which is reflective of the importance of the competition to the forecaster.

There are also the forecaster's conditional beliefs to consider; forecaster i ($i = 1,2$) knows her own probability p_i and does not know the probability of the other forecaster, but has the associated distribution. The conditional beliefs of forecaster i about p_j ($i = 1, 2, j \neq i$) given p_i are represented by continuous cumulative distribution functions (cdfs) $F_{p_j|E,p_i}$ and $F_{p_j|E',p_i}$, with corresponding probability density functions (pdfs) $f_{p_j|E,p_i}$ and $f_{p_j|E',p_i}$. It is to be noted however that the provided distributions are distributions for the reported probabilities, not the beliefs, which was determined can be different.

A second forecaster who doesn't care about relative performance has a $w_2 = 0$ thus her utility equals her score, meaning that her beliefs about the first forecaster hold no relevance making her optimal strategy the same as if she were the only forecaster under evaluation, i.e. truthful reporting; $\beta_2 = p_2$ (Lichtendahl & Winkler, 2007). The following proposition shows that truthful reporting is not necessarily the best response to combat truthful reporting. The proposition states that "if $0 < w_i < 1$, $0 < p_i < 1$, and forecaster 1 believes that forecaster 2 will report truthfully, truthful reporting satisfies the first-order condition to be forecaster 1's best response if and only if

$f_{p_2|E',p_1}(p_1)/f_{p_2|E,p_1}(p_1) = p_1/(1-p_1)$" (Lichtendahl & Winkler, 2007 p.1747). The example provided by Lichtendahl & Winkler is to use the normalized quadratic scoring rule already presented and suppose that forecaster 1's conditional beliefs about p_2 are beta with (*c, d, c', d'*) = (2, 1, 1, 2), the first order condition for maximizing expected utility is $\beta_1 = (p_1 - w_1)/(1-2w_1)$ and if p_1 and w_1 are such that this $\beta_1 \notin [0,1]$ or the second order condition $(2w_i - 1 \le 0)$ does not hold, then the optimal β_i is zero or one. So when $w_1 < 0.5$, $\beta_1 = \{0 \text{ if } 0 \le p_1 < p_1^*$, $\alpha_1 p_1 + \alpha_0$ if $p_1^* \le p_1 \le 1 - p_1^*$, and 1 if $1 - p_1^* < p_1 \le 1$, where $\alpha_1 = 1/(1-2w_1)$, $\alpha_0 = (1-\alpha_1)/2 = -w_1/(1-2w_1)$, and $p_1^* = -\alpha_0/\alpha_1 = w_1$. When $w_1 \ge$ 0.5, reporting extreme values is the best response: $\beta_1 = \{0$ if $p_1 < 0.5$, 0 or 1 if $p_1 = 0.5$, and 1 if $p_1 > 0.5$. Intuitively, this strategy shows that forecaster 1 is willing to stray from her beliefs and report higher probabilities in order to stand a better chance of winning the competition and is thus making a trade-off between the two terms in her utility function with a larger w_1 corresponding to how much she is willing to trade on the score for the relative performance term.

Lichtendahl and Winkler (2007) continue this competition theme with game theory to analyze both forecasters' decisions simultaneously (not as before with only one forecaster's decision based on her beliefs of the second forecaster's probability and utility function). Their focus involves each forecaster having the same utility function and the same beliefs about the other forecaster's abilities. Therefore, each forecaster's utility function is $u_i(R_i, S_i) = w_i R_i + (1 - w_i) S_i$ with $w_i = w$, i =1,2, the scoring rule is strictly

proper, and the cdfs $F_{p_j|E,p_i}$ and $F_{p_j|E',p_i}$ of each forecaster for the other forecaster's private information are continuous in p_j for $0 \le p_j \le 1$, i =1,2, $j \ne i$ and symmetric in the sense that $F_{p_2|E,p_1}(p_2) = F_{p_1|E,p_2}(p_1)$ and $F_{p_2|E',p_1}(p_2) = F_{p_1|E',p_2}(p_1)$ for $0 \le p_1, p_2 \le 1$. The following proposition invokes symmetry of utilities and conditional beliefs. The proposition states that "in this forecasting game, any strictly increasing, differentiable piece of a nondecreasing Bayesian Nash equilibrium reporting function $b(p_i)$ must satisfy the following differential equation:

$$w[p_i f_{p_j|E,p_i}(p_i) - (1-p_i) f_{p_j|E',p_i}(p_i)] + (1-w)[p_i s'_E(b(p_i)) + (1-p_i) s'_{E'}(b(p_i))]b'(p_i) = 0$$

" (Lichtendahl & Winkler, 2007).

Lichtendahl and Winkler provide an example with a quadratic scoring rule and beta distributions representing the forecaster's beliefs about their opponent's beliefs. With the application of the proposition equation, an ordinary, nonlinear, first-order differential equation of $b'(p_i) = \dfrac{w[p_i f_{Be}(p_i \mid c,d) - (1-p_i) f_{Be}(p_i \mid c',d')]}{2(1-w)[b(p_i) - p_i]}$ is yielded. Now, if (*c, d, c', d'*) = (2, 1, 1, 2) the equilibrium reporting function, denoted $\hat{p}$ equilibrium, when $w >$ 2/3, is of the form $b(p_i) =$ {0 if $0 \le p_i < \hat{p}$, $\alpha_1 p_i + \alpha_0$ if $\hat{p} \le p_i \le 1 - \hat{p}$, and 1 if $1 - \hat{p} < p_i \le 1$ where $\alpha_1 = (1 + \sqrt{1 + 8[w/(1-w)]})/2$, $\alpha_2 = (1 - \alpha_1)/2$, and $\hat{p}$ =0.2893 is the root of a quadratic equation, and extreme reporting is noted when $w \ge 2/3$. This strategy also implies that $b(p_i)$ should be closer to the extremes than to p_i, with noted jumps at

$\hat{p}$ and $(1-\hat{p})$ due to comparison of the expected utilities along the solution to the first order differential equation and reporting at the extreme values.

Lichtendahl and Winkler also consider this problem using the asymmetric case which presents a scheme of two differential equations (one per forecaster) and different reporting functions for the two forecasters. The first example entails the forecasters differing in their approach to the importance of the competitive aspect of relative performance, therefore $w_1 \neq w_2$. If $w_1 = 0.25$ and $w_2 = 0.5$, the second forecaster will give more extreme reports for intermediate values of p because he put more weight on the relative performance term. The second example modifies the beliefs instead of the weights, as in the previous example. This time $w_1 = w_2 = 0.25$ and the beliefs about and p_2 are beta with (c, d, c', d') = (1.25, 0.25, 0.25, 1.25) and (2, 1, 1, 2) respectively, showing that the conditional distributions of p_1 nearly dominate those of p_2, almost suggesting that forecaster 1 has more expertise. Overall the two examples suggest that the asymmetric beliefs did not lead to as great a difference in reporting functions as did the asymmetric weights. Lastly, knowing the utility functions and conditional beliefs of the other forecaster gives the analyst an idea of the manipulation that can occur when forecaster's have an inkling of each others abilities and preferences.

CHAPTER 7 Conclusion

Biased probability assessments can have a significant effect on the prescriptions from a decision analysis. Many of the applications that have been discussed herein that use probability assessments include consequences such as loss of life, health effects, and environmental impacts. In such cases, minimizing the impact of biases on assessed probabilities is critical. The goal of the decision analyst should be to provide an analysis that is as error free as possible.

Biased judgments can be made both consciously and unconsciously by both experts and non-experts alike. Heuristics play a key role in how people inject information into the judged uncertainty and how that uncertainty is projected into usable probabilistic quantities. Biases can and often do arise because of cognitive heuristics which may create havoc on an analysis. To minimize such problems, elicitation methods have been developed that include defining the problem in depth, training the assessor, eliciting the probabilities via an interview process, and verifying the results. Such methodology can eliminate most biases associated with the heuristics. Once the assessments are provided an analyst can determine the reliability of the forecaster (i.e. how good the forecaster is), the degree of calibration, and if the provided probabilities are coherent. If the probability assessments are determined to be biased, several methods to reduce this bias were provided. The assessments can be modified by calibrating them or by internally or externally ensuring coherence. The original elicitation process may be redone, possibly including a decomposition of the uncertainty into more manageable components. Lastly,

scoring rules can be used to encourage more honest forecasting. Moreover, the entire elicitation process will provide a reference for the expert to aid in the improvement of his own assessment ability. The process will also provide a reference for the decision maker in the determination of the forecaster's assessment abilities, should the decision maker need to call on that expert again for another decision analysis.

Proper elicitation and improvement methodology has been discussed for extruding unbiased probability assessments, but there remains a great disconnect between the fields of psychology and decision analysis when it comes to the subject of probability elicitation. There is an abundance of literature available for the psychological reasoning behind why the heuristics are utilized and the associated biases. There is an abundance of decision analysis literature that encompasses probability assessments, calibration of forecasters, scoring rules, how to elicit probabilities, etc. However, there is a lack of literature that connects the two subjects together. Many decision analysis papers discuss the elimination of biases but do not specify which biases their methods are trying to overcome exactly. There are psychology-oriented papers that mention the link to decision analysis but do not specify what methodology is employed. There is serious lack of connectivity between these two areas addressing probability elicitation, the biases associated with it, and the methods to overcoming how the biases are introduced. It seems apparent that decision analysts and psychologists who are interested in the subject of probability elicitation should collaborate to promote research that connects the two areas. Additionally, Benson, Curley, and Smith (1995) outline some research opportunities for decision analysis in data generation, argument construction, and associated judgments.

There are other areas that could benefit from additional research as well. Each improvement scheme; calibration, coherence, and scoring rules, is provided in a theoretical design but with a lack of proof of improvement. It appears that no (or very limited) studies have been performed to determine the exact affect of the methodologies, especially from a psychological standpoint. In essence, the psychological literature stated the biases that can be associated with heuristics, the decision analysts determined that training the subjects will overcome the cognitive judgment associated biases, but the psychology research to determine if the methods worked has not been performed. Additionally, there seem to be methods specific to each improvement concept but not any overarching methods that can incorporate multiple problems. For example, there is one methodology for calibration and one methodology for dealing with incoherence, but there is not a method that can overlap and involve both problems at once, though both problems are similar in stature.

Finally, there are many instances that are specific to probability elicitation that can be improved upon and many independent methods for accomplishing the improvements. There is not however, one standard overall process from start to finish that should eliminate the biases that are customary to assessing uncertainty. A complete process might include using the elicitation methods described in chapter four complete with subject training and an all encompassing set up procedure, using decomposition to minimize complexity, eliciting the assessments via the P and V methods with the addition of scoring rules to encourage honesty, verifying the assessments and checking for miscalibration and incoherence, and finally either repeating the process or manipulating the assessments should any biased assessments be produced. Overall, there is quite a bit of research being

done independently on each subject specific to probability elicitation. However, it appears that little research is being conducted on the overlapping concepts between psychology and decision analysis and also amongst the associated problems and the connectivity involving the improvement methodologies.

Literature Cited

Literature Cited

Benson, P.G., S.P. Curley, G.F. Smith. 1995. Belief assessment: An underdeveloped phase of probability elicitation. *Management Science* 41(10) 1639-1653.

Brenner, L.A., D.J. Koehler, V. Liberman. 1996. Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes* 65(3) 212-219.

Cain, D.N., G. Loewenstein, D.A. Moore. 2005. The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *Journal of Legal Studies* 34, 1-25.

Clemen, R., T. Reilly. 2001. *Making Hard Decisions*. Duxbury Press, California, USA.

DeGroot, M.H., S.E. Feinberg. 1983. The comparison and evaluation of forecasters. *The Statistician* 32(1) 12-22.

DeWispelare, A.R., L.T. Herren, R.T. Clemen. 1995. The use of probability elicitation in the high-level nuclear waste regulation program. *International Journal of Forecasting* 11 5-24.

Fischoff, B., R. Beyth. 1975. "I knew it would happen": Remembered probabilities of once-future things. *Organizational Behavior and Human Decision Processes* 13, 1-16.

Gigerenzer, G., U. Hoffrage, H. Kleinbölting. 1991. Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review* 98(4) 506-528.

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

Harrison, J.M. 1977. Independence and calibration in decision analysis. *Management Science* 24(3) 320-328.

Hora, S. 2007. Eliciting probabilities from experts. In *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press, Cambridge, UK.

Howard, R.A. 1989. Knowledge maps. *Management Science* 35(8) 903-922.

Jose, V.R.R., R.F. Nau, R.L. Winkler. 2008. Scoring rules, generalized entropy, and utility maximization. *Operations Research* 56(5) 1146-1157.

Kadane, J.B., R.L. Winkler. 1988. Separating probability elicitation from utilities. *J. Amer. Stat. Assoc.* 83(402) 357-363.

Keefer, D.L. 1991. Resource allocation models with risk aversion and probabilistic dependence: Offshore oil and gas bidding. *Management Science* 37(4) 377-395.

Keeney, R.L., R.K. Sarin, R.L. Winkler. 1984. Analysis of alternative national ambient carbon monoxide standards. *Management Science* 30(4) 518-528.

Keeney, R.L., D. von Winterfeldt. 1991. Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management* 38(3) 151-173.

Keren, G. 1997. On the calibration of probability judgments: Some critical comments and alternative perspectives. *Journal of Behavioral Decision Making* 10 269-278.

Liberman, V., A. Tversky. 1993. On the evaluation of probability judgments. *Psychological Bulletin* 114(1) 162-173.

Lichtendahl, K.C., R.L. Winkler. 2007. Probability elicitation, scoring rules and competition. *Management Science* 53(11) 1745-1755.

Lindley, D.V., A. Tversky, R.V. Brown. 1979. On the reconciliation of probability assessments. *J. Royal. Stat. Soc. Ser. A* 142(2) 146-180.

Lindley, D.V. 1982. The improvement of probability judgments. *J. Royal. Stat. Soc. Ser. A* 145(1) 117-126.

Matheson, J.E., R.L. Winkler. 1976. Scoring rules for continuous probability distributions. *Management Science* 22(10) 1087-1096.

Mellers, B., C. Locke. 2007. What have we learned from our mistakes? In *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press, Cambridge, UK.

Moskowitz, H., R.K. Sarin. 1983. Improving the consistency of conditional probability assessments for forecasting and decision making. *Management Science* 29(6) 735-749.

North, D.W., D.N. Stengel. 1982. Decision analysis of program choices in magnetic fusion energy development. *Management Science* 37(4) 377-395.

Ravinder, H.V., D.N. Kleinmuntz, J. Dyer. 1988. The reliability of subjective probabilities obtained through decomposition. *Management Science* 34(2) 186-199.

Sarin, R.S., R.L. Winkler. 1980. Performance based incentive plans. *Management Science* 26(11) 1131-1144.

Smith, J.E., D. von Winterfeldt. 2004. Decision analysis in management science. *Management Science* 50(5) 561-574.

Spetzler, C.S., C.S.S. von Holstein. 1975. Probability encoding in decision analysis. *Management Science* 22(3) 340-358.

Tversky, A., D. Kahneman. 1973. On the psychology of prediction. *Psychology Review* 80(4) 237-251.

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

Tversky, A., D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211 453-458.

Wallsten, T.S., D.V.Budescu. 1983. Encoding subjective probabilities: A psychological and psychometric review. *Management Science* 29(2) 151-173.

Wallsten, T.S., D.V. Budescu, R. Zwick. 1993. Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science* 39(2) 176-190.

Winkler, R.L., R.M Poses. 1993. Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science* 39(12) 1526-1543.

Winkler, R.L. 1994. Evaluating probabilities: Asymmetric scoring rules. *Management Science* 40(11) 1395-1405.

VITA

Bethany Hasbrouck Mihajlovits was born on Saturday, 06 October 1979, in Parkersburg, West Virginia and is an American citizen. She graduated from Lee Davis High School in Mechanicsville, Virginia in 1997. She received her Bachelor of Science degree in Operations Research in 2003 from Virginia Commonwealth University. She is receiving her Master of Science degree in Operations Research in 2009 from Virginia Commonwealth University. She has been an employee of CACI (Premier Technology Group) since November 2000 and has held the position titles of Office Manager, Operations Research Systems Analyst III, and currently Program Control Analyst Lead.