7 Expertise and Decision under Uncertainty: Performance and Process

Eric J. Johnson
The Wharton School, University of Pennsylvania

INTRODUCTION

In everyday life, we encounter many individuals we consider expert decision makers. Experts, whether they be physicians, security analysts, commodity traders, or bookies, all appear to possess an important talent: They make more accurate decisions than do other people in environments that are characterized by uncertain information. The commonplace belief is that they possess superior ability resulting from extensive training, hard work, practical experience, and professional dedication.

This opinion of expert judgment seems to be shared by cognitive science and artificial intelligence. Chapters in this volume, for example, examine experts in many domains ranging from computer programming and software design to medicine. Modeling expert performance has emerged as an important and difficult challenge for cognitive science.

In contrast, empirical research in behavioral decision theory presents a very different view of expertise: In many studies, experts do not perform impressively at all. For example, many expert judges fail to do significantly better than novices who, at best, have slight familiarity with the task at hand. This result has been replicated in diverse domains such as clinical psychology (Goldberg, 1970), graduate admissions (Dawes, 1971), and economic forecasting (Armstrong, 1978). Not surprisingly, this has led to strong recommendations. Consider for example, Armstrong's advice about experts' forecasts:

> People are willing to pay heavily for expert advice about the future. . . .The evidence is that this money is poorly spent. Expertise beyond a minimal level

> in the subject area is of almost no value. . . .The implication is obvious and clear cut: Do not hire the best expert you can—or even close to the best. Hire the cheapest expert. (pp. 84–85)

Indeed the results in the area are so uniformly negative that one wonders why Armstrong suggests hiring the expert at all: As Goldberg (1968, p. 484) stated ". . . [the] surprising finding—that the amount of professional training and experience of the judge does not relate to his judgmental accuracy—has appeared in a number of studies." More recently a dissertation has announced ". . . the end of the mystique of expertise" (Camerer, 1980a, p. 5).

Clearly these two divergent views warrant attention. In this paper, I describe the characteristics of experts as portrayed in these two literatures, concentrating on the behavioral decision literature. Two subsequent sections present an overview of research which examines the decision processes of two different kinds of experts. A final section attempts to use this research to help resolve the paradox presented by the two views of expertise held by cognitive science and behavioral decision theory.

Expert Judgment: Evidence

What are the essential properties which define an expert? There are many characteristics we associate with expertise: quick, confident judgments made under pressure, a reassuring manner, and an eye for the unusual or rare variable. However, one consideration seems tantamount: that experts make better—that is, more accurate—judgments than do untrained novices. In other words, expertise should provide both superior decision processes and superior performance.

Cognitive science has documented clear differences in experts' and novices' behavior. The specifics naturally depend upon the task, but it is clear that experts often have better and more complete representations of the task domain (Chi, Feltovich, & Glaser, 1982). These representations, in turn, allow experts to encode new information more quickly and completely (Chase & Simon, 1973; Johnson & Russo, 1984; Spilich, Vesonder, Chiesi, & Voss, 1979; Voss, Vesonder & Spilich, 1980). Experts apparently also have a richer repertory of strategies, and appropriate mechanisms for accessing and applying these strategies (Larkin, McDermott, Simon, & Simon, 1980). These strategies, and the appropriate organization of knowledge, often allow experts to perform tasks more quickly than novices.

Experts in domains such as physics problem-solving produce more accurate solutions than novices. However this focus on experts' performance is secondary, and problem-solving research often assumes the superiority of expert performance; it usually evaluates performance, if at all, on a small number of problems, concentrating instead upon difference in process

Research in decision and judgment provides a marked contrast. These tasks are characterized by an uncertain relationship between inputs and outcomes, task domains which are described by the term *decision under uncertainty*. The results in this literature present a rather pessimistic appraisal of experts. Experts are often found to perform no better than novices in the tasks studied. Consider, for example, Goldberg's (1959) result, comparing the ability of psychiatrists and their secretaries to diagnose brain damage using a common test, the Bender-Gestalt. He found no difference. Other studies may appear to be only slightly more heartening. Goldberg (1968) also compared undergraduate students and experienced clinical psychologists and psychiatrists in their ability to diagnose psychosis using the Minnesota Multiphasic Personality Inventory (MMPI). He found that experts were more accurate, making the correct diagnosis 65% of the time, compared to undergraduates' performance of 58%. A random responder would be right 50% of the time.

Much more devastating, however, are comparisons to simple statistical models, provoked originally by the "Clinical–Statistical" controversy in clinical psychology (Meehl, 1954). Here, experts are compared to simple regression models. The independent variables are the attributes describing a particular case, such as scores on a clinical test—for example, the scores on the Minnesota Multiphasic Personality Inventory. The dependent variable is an outcome, such as the eventual diagnosis of the patient as psychotic or neurotic. Such a model has no explicit knowledge of the environment, but combines the available numeric variables, using weights estimated through Ordinary Least Squares.

How well do these models perform? In almost every case, the models' predictions are more accurate than those of the expert judges. The experts in Goldberg's MMPI study made accurate diagnoses in 65% of the cases, the regression model in 70% of the cases. This finding has been replicated in many other domains, far beyond its original setting in clinical psychology. Einhorn (1972), for example, compared the ability of physicians to predict the severity of cases of Hodgkin's disease to the ability of a simple linear model. While the correlation of physicians' judgments with the observed outcome was no better than chance, cross-validated linear regressions performed modestly well, $r = .24$. Similarly, linear regressions have been shown to be superior to human experts in judging bankruptcy (Libby, 1976), predicting success in graduate school (Dawes, 1971), and predicting security prices (Wright, 1979).

These disappointing results have led to the comparison of experts to an even more limited mechanical combination rule. Here, rather than estimate weights, all important variables are weighted equally. Because no statistical estimation is involved, such prediction schemes are termed *improper linear models*. These models have no prior experience in the domain

and possess an almost trivial form. Often these models are superior to experts (Dawes, 1979). For example, in Einhorn's study of radiologists, an equal-weighted model was superior to all four physicians; and in Libby's (1976) study of bankruptcy judgments, a simple one-variable model predicted business failure more accurately than 31 of the 43 experts (Dawes, 1979, p. 579).

In sum, the behavioral decision literature does not present a flattering view of expert judgment. The superiority of experts to novices is often surprisingly small, or, in some cases, nonexistent; more disturbing may be the superiority of trivial linear representations to the performance of carefully trained human judges. These effects appear to be both robust and large: The surprisingly poor performance of experts has been replicated across a broad range of seemingly unrelated task domains, and models are often twice as good (in terms of variance explained) as expert judges.

Whereas the decision-making literature has evaluated the performance of experts in many domains, it has much less to say about process. Thus we have, at best, an incomplete picture. While process differences have been well documented in cognitive science, we know little about performance differences. The behavioral decision literature presents the opposite imbalance, with an emphasis upon performance and not process.

Expert Judgment: Process Explanations

Why do experts in these domains do so badly? The behavioral literature, with its emphasis on performance, does not offer much in the way of answers. Why is the performance of these experts so remarkably poor in comparison to those in domains such as physics? First, note the obvious differences in the tasks. Most problem-solving research examines well-structured tasks. Expertise consists of identifying a correct procedure for obtaining a solution and applying it. The procedure, whether learned through experience or instruction, is usually known to provide a correct answer, and often provides some means of checking if the answer is correct. In decision under uncertainty no single correct procedure exists, and there is no definitive way of assessing the correctness of a rule based upon the outcome of a single case. There is no *optimally* correct rule, for example, to predict psychosis, bankruptcy, or the severity of Hodgkin's disease; there are only rules which are *relatively* more accurate. Thus, the tasks may make radically different demands upon experts' abilities.

While there may be no single way of being right in these tasks, there seems to be more than one way of being wrong. At least three different hypotheses suggest themselves:

1. *Experts are fallible linear models.* Experts might attempt to behave

like a linear model, but, because of their limited ability to process information, fail. Dawes (1979), for example, speculates that linear models perform well in these tasks "because people – especially experts in a field – are much better at selecting and coding information than they are at integrating it." This theme, that the human judges are deficient in combining information, appears in a number of studies. Here experts are seen as noisy regressions, who can identify important variables, but who fail to combine them accurately. "People are good at picking the right predictor variables. . . . People are bad at integrating information from diverse and incomparable sources (Dawes, 1979)." Thus even expert judges might use inappropriate weights, or apply those weights in an inconsistent or unreliable fashion. It is important to emphasize that even the most pessimistic researcher would not eliminate experts from these tasks. Experts' strength is in the selecting and coding of relevant variables; their weakness seems to be in combining them.

2. *Experts use nonlinear rules.* According to this view, experts may use the same variables as statistical models, but combine them differently. Experts' discussions of their judgment processes, whether gathered informally, as in much of this literature, or using concurrent verbal reports (Kleinmuntz, 1963), do not fit the picture of the expert as a noisy linear combination rule. Rather, experts report that they use complex *configural* rules: The impact of one variable depends upon the level of another, in a fashion that is analogous to an interaction in an analysis of variance. The impact of one scale of the MMPI depends, for example, upon whether or not the other scales also have high values. These rules often sound very similar to the if–then form used in production system models of experts.

However, arguments for configurality must address two rather distressing facts: First, the judgments of most experts are well predicted by linear representations, with interactions representing, at best, a minor part of their judgments. Second, these regression models of judges (called *bootstrap models*) do better at predicting the criterion than do the judges themselves. This first finding, that judges who claim to be configural are well modeled by simple regressions, may well be epiphenomenal: Linear models can predict nonlinear processes under a wide variety of well-defined conditions (Dawes & Corrigan, 1974; Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Johnson & Meyer, 1984). However, the second finding, that experts are less accurate than linear models based upon their judgments, suggest that whatever the experts *are* doing does not improve their performance.

3. *Experts attend to different variables than do models.* Meehl (1954), a central actor in the clinical–statistical debate, suggested the following example: Imagine that we had to predict whether or not a faculty member would attend the movies on a given night. We might construct a sophisticated linear model consisting of variables such as marital status, whether

or not our colleague was tenured, age, number and age of children, and so on. However, if we found out that the faculty member had just broken his or her leg, then we would confidently discard our model and make our prediction based on this fact.

Such events, which we term *broken leg cues*, are not included in the regression because they occur too infrequently to be estimated. Yet when they are present, such cues can be quite diagnostic. Experts, according to this different-variable hypothesis, might tend to use such cues rather than those included in the regression model. Their resulting poor performance occurs because these cues are less predictive than those in the model.

Currently, it would be difficult to ascribe experts' poor performance to one of these causes or another. Decision theory's emphasis upon performance has allowed it to assess the performance of judges, but, unfortunately, it has less to say about the causes of that performance. While regression models can provide a good account of the outcomes of a decision process, they are relatively uninformative about the psychological process producing these outcomes. The actual nature of these processes deserves closer examination, using the methods that reveal the information examined by the judges, and the methods judges use to combine this information. Emphasizing process may also help reconcile the two views of expertise we have encountered. Do experts in decision-making behave like experts in other domains? Do they possess superior information-processing skills?

The next two sections present an overview of a series of studies which examine expert performance and processes in two different domains: The first, evaluating applicants for medical internships, is similar to what has become a standard task in the decision literature: graduate admissions judgments (Dawes, 1971). The second task is the prediction of stock prices by expert security analysts and by novice MBA students. In both studies we have changed the standard methodology used in decision research in two ways: (a) We have presented the decision-makers with an environment that is richer in information than that normally used in studies which evaluate experts' performance; and (b) we also have collected verbal reports on some subsets of the trials, which allow us to examine these experts' decision processes. Additional details are available in Johnson (1980) and in Johnson and Sathi (1988), respectively.

EXPERT JUDGMENT: PHYSICIANS' SELECTION OF HOUSE OFFICERS

The Task

Each year, applicants for internships and residencies (collectively termed *house officers*) participate in a program which matches applicants to post-

graduate positions in teaching hospitals. The applicants express their preferences by ranking desirable positions. The hospitals, in turn, must submit a rank ordering of candidates to ensure the admission of desirable applicants. The rank-order preferences of both applicants and hospitals is processed through the National Internship and Residency Matching Program, which uses a complex algorithm to make the assignments. Because the hospitals are blind to the preferences of the applicants, their ranking of applicants represents their only means of maintaining the quality of their house staff. Consequently, the physicians expend considerable effort at this process; each of the twelve physicians on the admissions committee we studied examined the folders for each of 200 applicants, and attended two day-long meetings in which final ranks were computed. Although administrative staff might perform this task, the physicians believe that they have the inherent expertise which justifies their expenditure of about one person-week apiece.

Information about each applicant is contained in a folder, which consists, on average, of 13 pages of material. The contents of the folders include an application form supplied by the Department of Medicine, letters from the dean and faculty of the applicant's medical school, transcripts of course work, two summaries of interviews conducted with the applicant, and the results of a standardized exam, the National Boards. Tallying the number of separate statements contained in these folders shows that there are over 400 potentially relevant facts to be considered.

The goal for each judge was to simply rate the applicant on a five-point scale using information in the application. These ratings are then summed to form an initial rank ordering of all applicants, which is then slightly modified following discussion.

Our analysis of these experts involved two sources of data: (a) Concurrent verbal reports collected from two of the physicians and two undergraduate novices as they reviewed six of the applications; and (b) The overall ratings of all twelve physicians and a single novice on 156 applications. In both cases, subjects provided ratings of non-numeric items such as the letters of recommendation, and we coded for each applicant a set of objective variables, such as National Board scores, listed in the applicant's folder. More detail is available in Johnson (1980).

Process Differences

To examine differences in the processes employed by experts and novices, each protocol was segmented into a series of complete thoughts, and coded into one of six categories, each representing a category of cognitive process. The categories and definitions are presented in Table 7.1. We also identified, for each occurrence of a retrieval operator, the actual statement which

TABLE 7.1
Description of Statement Types

Retrievals	Nonevaluative statements consisting of verbatim or paraphrased quotation of information presented in the folders.
Recall	Statements of similar nonevaluative information obtained from memory.
Evaluation	Statements which result in the judgment of some aspect of the applicant, his or her medical school or other object. This excludes judgments made in response to a question on the response form, or evaluations read from the folder.
Scaling statements	Responses made in completing the form provided with each application.
Inferences	Nonevaluative statements, based on retrieved information, but which clearly go beyond the presented information.
Goal statements	Statements of intentions or actions to be performed. Search for a source of information, etc.
Miscomprehension	Statements reporting difficulty in understanding presented information.
Comment	An uncodable statement or one irrelevant to the task.

was read. This allowed us to compare the experts and novices, both in terms of the information search and in the cognitive processes used to evaluate information.

These protocols reveal several qualitative and quantitative differences. Most striking are the differences in the time required to perform the rating task: The two experts averaged about 7.8 minutes per applicant, while the novices took almost twice as long, about 15 minutes per applicant. There was almost no overlap in the two distributions of time. These differences were due, in part, to the smaller amount of information examined by the experts. The novices' protocols contained almost twice as many retrievals as the experts' (126.5 vs. 64.2 retrievals per protocol). While the novices examined over 43% of all the statements that were available in the folder, experts examined about 22%. The experts also examined *different* information than the novices. Transcripts, for example, were barely examined by the experts. Only 3% of the statements contained in the transcript were examined by the experts, while the novices examined about 13%. The experts also seemed to limit their examination of the letters, concentrating upon one or two key sentences in each letter. The only item examined more closely by the experts was the application form. Here expert subjects examined about 42% of the statements and the novices about 38%.

These differences in search appear to reflect experts' belief that certain items provided by the applicant are relatively uninformative: The transcript, and the grades provided, often reflect a pass/fail system. An expert seems to know that any graduate of these programs has passed these courses. The dean's letter, in contrast, often contains a series of key phrases describing the students' progress in each course. There are several phrases that, like grades, indicate differential performance in class work. The experts appear to use their knowledge of medical education to focus upon more diagnostic information.

Experts not only search for different information, they also have different patterns of search. By and large, the novices examine the information present in the folders the way it is presented, reading one item at a time, moving sequentially down the page. The experts search much more actively: They return to previously examined information much more often, and change the focus of their attention from one part of the folder to the other much more frequently. The tendency of experts to examine the information in a more active, flexible manner also manifests itself in the frequency in which they apply certain operators. Specifically, more goal statements appear in the experts' protocols. About 3% of all statements in experts' protocols are coded as goal statements, while novices' protocols contain less than 1%. Similarly, experts make more extensive use of their knowledge of medical education. About 10% of all the information used by experts appears to be recalled from memory, and by novices, about 3%.

In sum, the experts appear to examine this information in a top-down fashion, using their knowledge of medical education to structure their search. The increased use of goals in their protocols, along with the greater use of knowledge retrieved from memory and their more active search patterns, presents a picture consistent with the portrait of experts in other domains. Our expects appear to use their knowledge to examine only information that they consider diagnostic, limiting their search to a smaller subset of the available information. Thus, although admissions tasks such as this have often resulted in rather disappointing evaluations of expert performance, these results suggest that the processes used by these experts are much like those used by experts in other domains.

More importantly, these processes do not resemble a linear regression model. Rather than being fallible approximations to a linear model, these judges seem to use information in a quite different manner. While there are occasional references to configural effects among the cues, there seems to be marked use of information which applies only to the particular case at hand. Consider the following example taken from one of the experts' protocols:

At Hopkins the thing to look for is the Ossler Clerkship
It's really a sub-internship in medicine

He did very well there . . . Honors.
I'd think he would therefore do well here. . . .

However mundane, this inference would be impossible to model within a regression framework. While one could include the complex interaction, the clerkship in question is available to only a few students at one particular medical school. There would be far too few observations to estimate a coefficient. Interpreting the impact of this information does not require statistical estimates of covariation, but rather the realization that the duties of the clerkship are similar to those of a house officer. The process data often seems to more closely resemble the different-variable model of expert judgment: Experts seem to pay attention to relatively rare variables, applicable to only the case under consideration.

Experts and Novices in Admissions Judgment: Performance

While we have seen important differences in the decision processes used by experts and novices, an important question remains: Do experts perform more accurately? An important aspect of the matching process used to assign house officers is that it allows some evaluation of the individual decision-makers. The National Residency and Intern Matching Program provided the rankings of these applicants at 32 teaching hospitals, which represents a large majority ($> 80\%$) of the hospitals in which the applicants sought positions. We can then identify, for each applicant, the rank which maximized the probability of the applicant being assigned to these experts' hospital. Deviations from this rank may be suboptimal: Either a candidate is ranked more highly than necessary to insure admission, or a candidate is ranked too low, increasing the probability of his or her assignment to a competing hospital. Although other, more sophisticated definitions of accurate judgment in this task are possible (Roth, 1984), the simple comparison of a judge's rankings to those that optimized a candidate's chances of being assigned to the physician's hospital represents a useful first step. To further ensure the relevance of this criterion, we collected ratings of the applicants' performance for that subset who eventually joined the Department of Medicine. The chairperson of the admissions committee and the head of the training program both rated and ranked each intern at the end of the first year of training. The resulting correlation between this measure and the committee ranking, $r = .48$, suggests some relationship between the criterion we use and eventual success as an intern.

To evaluate the performance of these experts we can compare their rank orderings of the applicants to one which would have given the training program the best chance at matching with each applicant. The simple correlation of the experts with this criterion ranges from .50 to .29 with a mean

of .37. Although this appears to be a disturbingly low level of performance, it is typical to that found in similar tasks: Wiggins and Kohen (1971) reported an average correlation of .33 between predicted and actual freshman Grade Point Averages for incoming college students; and Dawes reported that faculty rankings at the time of admission correlated .19 with a measure of eventual success in graduate school.

One evaluation of the experts in this study is the comparison of their performance with that of an undergraduate novice who had rated the 151 applicants. This novice performed as well as two of the physicians, ranking eleventh out of the total of thirteen judges. His correlation with the criterion was .33. Thus, although we find marked differences in the processes that are used by novices and experts in this task, these differences appear to have a relatively small impact upon performance. We also compared the expert judgments to a cross-validated linear regression model of the outcome. Recall here that the results are a uniformly disappointing assessment of expert performance: Experts never perform as well as the model. In this study, the cross-validated model performs well, $r = .48$, but here a single expert performed as well as the model. The model is quite simple, and uses only three predictor variables selected by a stepwise regression: the quality of the applicant's medical school, an interviewer's rating of an on-campus visit, and a variable indicating membership in an honorary society.

In sum, the evaluations of these experts presents, at best, a slightly more optimistic view of expertise: Most experts are slightly better than an undergraduate novice, and a single expert actually performs about as well as a simple linear regression. These results are summarized in Figure 7.1 which

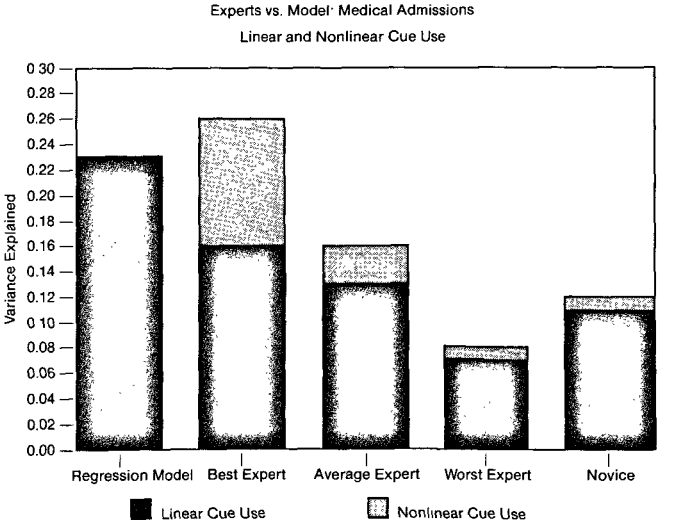

FIGURE 7.1 Experts, novices, and models' performance: Admissions task.

displays the performance of the best and worst expert judge, the novice, and various statistical models.

Yet these results present us with a quandary: How can a simple regression model be as accurate as the apparently complex processes revealed in the verbal reports? The regressions are more accurate than most experts, yet they consider only three variables. The regression and the judges seem to be doing different things. Unlike the experts, the models look at the same information for every case. In contrast, the verbal reports indicate that the physicians examine many more cues, and the pattern of their search seems to be contingent upon the information at hand.

This paradox suggests that the fallible regression hypothesis is wrong. The experts' strategies seem quite different than that portrayed by the regression. To examine this further, we explicitly tested the noisy regression hypothesis. Using stepwise regression, we constructed linear models of the judges (i.e., "bootstrap" models), which fit moderately well, explaining between 41% and 77% of the variance in judgments. If these models represent all the systematic variations in judgment, the residuals of these models should have no relationship with the outcome (Camerer, 1980b). If, on the other hand, judges are using cues in a valid configural fashion, or making use of broken-leg cues, we would expect a positive correlation between residuals (what the regression model cannot explain) and outcomes. In essence, this statistical technique divides the accuracy of the judgments into a linear component, modeled by a regression, and a nonlinear component, due to the valid use of configural and broken-leg cues. It also allows us to attribute the accuracy of each judge to two different sources: (a) a linear component, and (b) a nonlinear component, consisting of valid configural and broken-leg cues.

On average, the correlation between the residuals and the criterion was positive, $r = .124$, and statistically significant. The maximum, $r = .314$, belonged to the most experienced physician. This nonlinear cue usage was an important part of these experts' judgments, accounting on average for 15% of their predictive validity. The best judge also showed the most valid use of configural and broken-leg cues, accounting for 37% of his valid judgment. Figure 7.1 reflects the division of performance into linear and nonlinear cue usage by dividing total accuracy into two components. The top section represents nonlinear cue usage, while the bottom half represents linear cue usage captured by the regression.

In sum, the data from this study clearly weakens the notion that the expert judges are simply fallible regressions. One source of evidence is the verbal reports. The goal-driven, knowledge-intensive, and contingent search patterns are inconsistent with a regression model's focus upon one set of cues. A second source of evidence is the regression analyses, which show that a linear model fails to capture a significant part of the experts' valid

judgments. Why then do linear models do so well? Apparently because the model uses information that the expert ignores, and because that information is important.

We must be somewhat cautious about interpreting the relative importance of linear and nonlinear information. Our analysis to date has been correlational; this study does not experimentally manipulate the presence of broken-leg cues. Thus, the regression models may over- or underestimate nonlinear cue usage. More importantly, we have not separated the two types of nonlinear information usage: configurality and broken-leg cues.

In the next section we describe a study addressing these issues in a task possessing a real-world criterion, and obvious real-world incentives: the prediction of security prices. This domain allows us to experimentally manipulate the presence or absence of potential broken-leg cues—in this case, summaries of news items from the *Wall Street Journal*—allowing us to better assess the importance of such cues.

EXPERT JUDGEMENT: PREDICTING SECURITY PRICES

The Task

In a recently completed study (Johnson & Sathi, 1988), we have compared predictions of changes in security prices made by experienced security analysts and by inexperienced MBA students. Each of the subjects in this study predicted year-end closing prices for 40 securities. The securities were described on a set of 22 variables similar to those usually available to analysts engaged in predicting security prices. Half of these securities were accompanied by news items which were summaries of stories about the company that had appeared in the *Wall Street Journal*. Whereas the financial information and the news items described actual securities in 1980, the names of the securities were not revealed to the subjects, ensuring that they would have to predict, and not simply recall, the year-end closing prices. In both expert and novice groups there was, in addition to normal compensation, a sizable prize awarded to the most accurate judge. This provided an additional incentive for accurate judgment.

Process Differences

Our expert subjects represent an average of four years of experience, and were employed as research analysts in three different companies. Our novices were students studying for their Masters degrees in administration, who had taken only an introductory course in finance. Our analysis, although still in progress, replicates several of the process differences observed in the previ-

ous study. A comparison of the time required to make a prediction demonstrates that experts were faster than novices, averaging about 144 seconds per security compared to the novices' 162 seconds. A coding scheme similar to that used in the previous study has yet to show substantial differences in the types of operators used by experts and novices. Differences do appear, however, when we examine the information examined by experts and novices. Experts concentrate on certain variables: They examine certain financial descriptors such as the earnings per share and the previous year's closing price more intensely than do novices. These two variables represent 22% of all information examined by the experts, and only 5% of that examined by novices. Again experts appear to focus upon fewer cues than do novices: While novices, as a group, examined 21 out of 22 possible variables, the experts examined only 13. Apparently, like the physicians, they ignored cues they believed to be redundant. Although these results are quite tentative, differences in the processes used by experts and novices in this task are marked, and consistent with those described in the previous study.

Performance Differences

A important advantage of security pricing as a task is the existence of a clear-cut standard for the measurement of performance: actual price changes. To evaluate the experts and novices, we can compute the average absolute size of their errors, that is:

$$| \text{actual price} - \text{predicted price} | / \text{actual price}.$$

Of course with this error measure, the smaller the mean, the better the performance, and perfect performance would yield a mean error of zero. Overall, experts did perform better than novices: Their average error was 61.5% compared to a mean error of 65.3% for the novices, a statistically significant difference. More interesting is the interaction between expertise and the presence of the news items. This interaction, which is presented in Figure 7.2, demonstrates that the presence of the news items does not help the novices, but increased the accuracy of the experts. Thus the news items appear to aid expert judgment, at least in part because the experts seem able to comprehend the impact of these items.

Despite the advantage apparently provided by the news items, these experts do not approach the performance of a simple regression. A cross-validated regression of the cues, upon the criterion of percentage change in price, shows that the model is still superior to the mean of the expert judge, having an average error of only 52.4%. Thus, although experts do better in the presence of news items, they still are inferior to a relatively simple linear regression.

We next divided the experts' predictive validity into linear and nonlinear

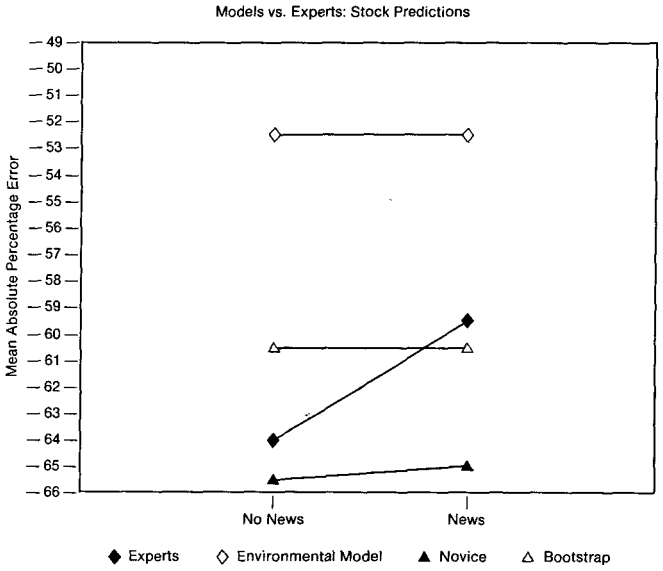

FIGURE 7.2 Experts, novices, and models performance: Security pricing with and without broken-leg cues.

components, using the bootstrapping technique used previously. Figure 7.3 shows that without the news items, the experts were well modeled as linear combination rules. However, the presence of the news items increased their ability to predict, largely through the use of nonlinear information. In contrast, the novices use nonlinear information less, and their nonlinear cue usage is largely independent of the presence of the news items. Thus, the experts seem to depend heavily upon the presence of news items to increase the accuracy of their predictions. In contrast the novices appear to use the information in some nonlinear fashion that appears to be configural and independent of the news items.

In terms of our three models of expert performance, the data from this study tend to support the different-variable hypothesis to the detriment of the configural cue hypothesis. Our experts only show significant nonlinear cue use when the news items are present. It then seems logical to attribute this nonlinearity to the presence of the news items. Over half of the experts' predictive validity is due to this information. Accurate judgment of the impact of these rare events seems essential to their expertise. The key findings of this research are therefore:

1. That experts in these ill-structured tasks behave in many ways like experts in well-structured domains.
2. That experts concentrate on the interpretation of rare events.
3. While this aids their predictions it leads, in the tasks we have exa-

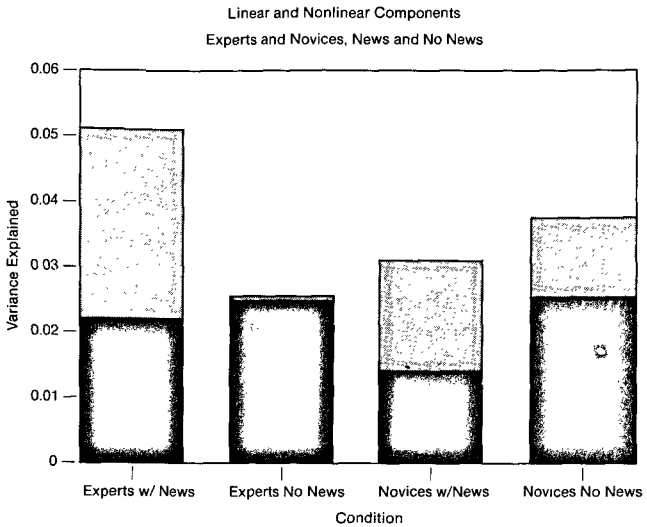

FIGURE 7.3 Experts and novice security analysts: Linear vs. nonlinear cue usage.

mined, to performance inferior to that provided by simple linear models.

DISCUSSION

Experts' Strengths, Experts' Weaknesses

These two studies clearly show experts' strengths: the interpretation of cues that apply to these particular cases. The experts' weakness, and the strength of regression models, is the ability to combine the more mundane information available for every case. Note that this distinction is analogous to another in the behavioral decision literature, that of base-rate versus case-specific data (Kahneman & Tversky, 1973). In this literature it is apparent that base-rate data is often underweighted, relative to case-specific data.

Why is case-specific data overweighted? Experts appear to interpret rare events through inference, and, more specifically, through causal reasoning. The ability to assess the impact on the price of security of a rare event such as the death of a Chief Executive Officer in a plane crash does not lie in lessons learned from experience. Instead, it seems that experts can assess the event's impact because they have some knowledge of the company and the CEO's role within it, and understand the potential market reactions. This general ability to learn better pattern-matching and reasoning

skills is consistent with the development of expertise in other domains. P. Johnson and his colleagues (1981), for example, have shown that expert physicians seem to be better able to identify the patterns of symptoms which are linked to a disease. Knowledge of the symptoms alone is not enough: "A failure in human subjects does not reflect the lack of a disease model in memory, but rather a lack of knowledge for using that model in particular situations" (P. Johnson et al., 1981, p. 274). This knowledge of patterns, and their application, seems to be the key to expertise in many domains (Larkin et al., 1980). While this is a useful skill in the domains characterized as decision under uncertainty, there is also other useful information available.

Why then is base-rate data neglected? Whereas we may be impressed with the performance of "simple" regression models, we must attribute their performance to their ability to properly weight base-rate data. Although perhaps simple in statistical terms, the cognitive processes represented by the regression model are not "simple" at all. Because the relationship between the independent variables (cues) and the dependent variables is stochastic, we need a large number of cases to evaluate their covariation. Without written records, such a task overwhelms the abilities of the decision-maker. The ability of individuals to judge covariation is severly limited, as documented in the literature on covariation (Crocker, 1981), multi-cue probability learning (Brehmer, 1980), and even judging covariation in a simple two-by-two table (Jenkins & Ward, 1965). The origins of the neglect of base-rate data seem to stem from the inherently burdensome information-processing demands of the task. In retrospect, it seems obvious that neglect of base-rate information is the key to the relative weakness of experts, as compared to models.

Implications

This tendency for decision-makers to examine broken-leg cues and to neglect other information represents an important opportunity for those who would wish to design aids to decision-making for experts. Specifically, we might design aids which help decision-makers by combining the information that they currently underweigh, and providing an estimate of its impact. An expert might then adjust this initial estimate to account for information not considered by the model, such as broken-leg cues. We are currently exploring this possibility. For details, see Johnson & Sathi (1988). Such an aid illustrates one of the payoffs of a process analysis of expert judgment. By decomposing prediction tasks into components, we can isolate those that are easily accomplished by a human judge from those which present difficulty. This then allows us to consider the design of aids which assist the judge with the more difficult components of the task.

This research also raises some interesting questions for the development of systems which capture human expertise in the form of a computer program. Much work in these expert systems seems to focus upon mimicking the performance of a human judge (Duda & Shortliffe, 1983). The present results suggest that the appropriateness of modeling a human judge may depend on the task. Because we cannot easily identify those tasks which are heavily dependent upon broken-leg cues, we need to approach the modeling of judges with caution. Specifically, if an expert system attempts to mimic a human expert, it may fail to exploit much of the information available in the task, such as the base-rate information captured by a simple regression model. For some tasks we may find ourselves in the ironic position of having developed a useful model of human performance, which also captures the experts' foibles—specifically, their tendency to emphasize broken-leg cues while neglecting other information.

This is a particularly troublesome observation since the protocols generated by our judges sound suspiciously like the productions contained in many expert systems. We might encode our physician's observation concerning students from Hopkins in the form:

If (Applicant is from Johns Hopkins and has Ossler Clerkship and grade is equal to an A)
Then (increase certainty that applicant is a good intern)

This is probably valid, leading to an increase in the accuracy of judgment. However, this may miss more valid but mundane relationships like those between grades and performance. This observation should emphasize, therefore, the importance of evaluating the performance of expert decision-makers, whether they be humans or machines, against simple alternative models. Simple linear models can serve as interesting baselines for the evaluation of expert systems.

Finally, the current research has barely started to examine the psychology of expert judgment. We have failed to examine specific facets of experts' cognitive abilities. Although there is an increasing understanding of experts' abilities in many domains, the current work has not specifically examined some of these, such as the role of their possibly superior encoding and recall skills in their performance.

ACKNOWLEDGMENT

This work has been supported by a grant from the Risk, Decision and Management Science Program at the National Science Foundation, while the author was at Carnegie-Mellon University. Comments on earlier drafts by Colin Camerer, John Payne, Anthony Pratkanis and J. Edward Russo are very much appreciated.

REFERENCES

Armstrong, J. S. (1978). *Long range forecasting: From crystal ball to computer.* New York: Wiley.

Brehemer, B. (1980). In one word: Not from experience. *Acta Psychologica, 45,* 223-241.

Camerer, C. (1980a). *The psychology of expert judgment.* Unpublished doctoral dissertation, University of Chicago, Graduate School of Business.

Camerer, C. (1980b). Conditions for the success of bootstrap models. *Organizational Behavior and Human Performance, 24,* 411-422.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55-81.

Chi, M., Feltovich, P., & Glaser, R. (1982). Categorization in experts and novices in R. Sternberg (Ed.), *The Handbook of Intelligence,* Earlbaum: Hillsdale, N.J.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90*(2), 272-292.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180-188.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95-106.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582.

Duda, R. O., & Shortliffe, E. H. (1983). Expert systems research. *Science, 220,* 261-268.

Einhorn, H. E. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7,* 86-106.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B., (1979). Linear regression *and* process-tracing models of judgment. *Psychological Review, 86,* 465-485.

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt test. *Journal of Consulting Psychology, 23,* 25-33.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence for a method of improving clinical judgment. *Psychological Bulletin, 73,* 422-432.

Goldberg, L. R. (1968). Simple or simple processes? Some research on clinical judgments. *American Psychologist, 23,* 483-496.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 79,* 11,15,19.

Johnson, E. J. (1980). *Expertise in admissions judgment,* unpublished Doctoral Dissertation, Carnegie-Mellon University, Department of Psychology.

Johnson, E. J., & Meyer, R. M. (1984). Compensatory choice models of Noncompensatory Processes: The Effect of varying Context. *Journal of Consumer Research, 11,* 528-541.

Johnson, E. J., & Russo, J. E. (1984). Product familiarity and learning new information. *Journal of Consumer Research, 11,* 542-550.

Johnson, E. J., & Sathi, A. (1988). *Expertise in security analysts.* Manuscript in preparation.

Johnson, P. E., Duran, A. S., Hassebrock, F., Moller, J., Prietula, M., Feltovich, R. J., Swanson, D. B. (1981). Expertise and error in diagnostic reasoning. *Cognitive Science, 5,* 235-283.

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

Kleinmuntz, B. (1963). Personality test interpreted by digital computer. *Science, 139,* 416-418.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980, June). Expert and novice performance in solving physics problems. *Science, 208,* 1335-1342.

Libby, R. (1976). Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance, 16,* 1-12.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Roth, A. E. (1984). The evolution of the labor market for medical interns and residents: A case study in game theory. *Journal of Political Economy, 92*, 991–1016.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 118*, 275–290.

Voss,, J. F., Vesonder, G. T., & Spilich, G. J. (1980). Text generation and recall by high-knowledge and low-knowledge individuals. *Journal of Verbal Learning and Verbal Behavior, 19*, 651–667.

Wright, W. F. (1979). Properties of judgment models in a financial setting. *Organizational Behavior and Human Performance, 23*, 73–85.

Wiggens, N., & Kohen, E. S. (1971). Man vs. model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19*, 100–106.