ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 62, No. 2, May, pp. 129–158, 1995

Eliciting Knowledge from Experts: A Methodological Analysis

ROBERT R. HOFFMAN

Adelphi University

NIGEL R. SHADBOLT

The University of Nottingham, United Kingdom

A. MIKE BURTON

The University of Glasgow, United Kingdom

AND

GARY KLEIN

Klein Associates, Inc.

The psychological study of expertise has a rich background and has recently gained impetus in part because of the advent of expert systems and related technologies for preserving knowledge. In the study of expertise, whether in the context of applications or the context of psychological research, knowledge elicitation is a crucial step. Research in a number of traditions—judgment and decision making, human factors, cognitive science, expert systems—has utilized a variety of knowledge elicitation methods. Given the diversity of disciplines, topics, paradigms, and goals, it is difficult to make the literature cohere around a methodological theme. For discussion purposes, we place knowledge elicitation techniques into three categories: (1) analysis of the tasks that experts usually perform, (2) various types of interviews, and (3) contrived tasks which reveal an expert's reasoning processes without necessarily asking about these processes. We illustrate types and subtypes of techniques, culminating in a discussion of research that has empirically evaluated and compared techniques. The article includes some recommendations about "how to do" knowledge elicitation, some cautionary tales, and a discussion of the prospects. © 1995 Academic Press, Inc.

The authors thank James Shanteau and three anonymous reviewers for their suggestions on the drafts of this article. Address correspondence and reprint requests to Robert R. Hoffman, Department of Psychology, Adelphi University, Garden City, NY 11530. E-mail: hoffman@adlibv.adelphi.edu.

INTRODUCTION

In experimental psychology, the study of expertise or proficiency has always been regarded as having merit, touching on basic questions having to do with cognition and perception, and applications involving training and the preservation of knowledge. For example, Book's (1924) study of "world champion" typists and Bryan and Harter's (1897) study of telegraphers—classics in experimental psychology—focused on basic questions about motor skill acquisition (i.e., plateaus in learning curves of complex performance; Woodworth, 1938, Ch. 7). In hindsight these studies could be embraced under the banner of "expertise."

In the areas of organizational behavior and decision making, effort has gone into the study of experts in such diverse domains as accounting, auditing, management, livestock judging, finance, and so on (Fischoff, 1989; Libby & Lewis, 1977; Shanteau, 1988). Human factors psychology is laden with studies of performance at aircraft piloting, radar operation, and air traffic control. Many of these studies have relied on highly experienced, proficient participants (cf. Alluisi, 1967; Chiles, 1967; Christensen & Mills, 1967).

Studies on topics relating to expertise appear in the archives of psychometrics. Brown and Ghiselli (1953) used a battery of standard aptitude tests in an attempt to predict the proficiency of taxicab drivers. Jenkins (1953) employed questionnaires and standardized achievement and aptitude tests to assess the charac-

0749-5978/95 $6.00

teristics of weather forecasters, comparing the data to task analyses and forecasting skill scores. Hammond's (1966) research on nursing involved field studies that disclosed the sorts of problems nurses encounter and laboratory studies using test case problems that disclosed nurses' patterns of clinical inference.

In recent years, the study of expertise has been invigorated (Hoffman & Deffenbacher, 1992). The impetus comes in part from the national emphasis on science education—leading to studies of expertise in such areas as medicine and physics. The study of expertise opens possibilities that seem inherently interesting to researchers, as suggested by studies in such areas as birdwatching (Coltheart & Walsh, 1988), traditional herbal medicine (Cox & Balick, 1994), and satellite image interpretation (Hoffman & Conway, 1989).

Research has been conducted on expertise in everything from military command and control to jurisprudence, from social policy making to dermatology, from athletics coaching to electronics trouble-shooting, from wholesale milk delivery to farming in Peru, and from wastewater treatment to the learning of archaeological categories (Chi, Glaser, & Farr, 1988; Ericsson & Smith, 1991; Hoffman, 1992a; Krovvidy & Wee, 1993; Scribner, 1984; Shanteau & Stewart, 1992; Vandierendonck, 1993).

Leaders in business, government, and the military are recognizing the value of studies of "naturalistic decision making" (Klein, Orasanu, Calderwood, & Zsambok, 1993; Zsambok & Klein, 1995). Corporate executives are realizing the value of capturing and preserving the knowledge and experience of their most skilled employees (Cross, 1988; Klein, 1992). For instance, an expert at the mass production of soups was close to retirement. As the time approached, his company realized that there was no one else who knew what he knew (Herrod & Smith, 1986). As another example, we have heard informally of a number of woeful tales of knowledge loss due to the retirement of Apollo-era NASA scientists.

With the growing importance of information technology, one focus of cognitive science has been on the acquisition of computer programming skill (Hoffman, 1992b). In addition, the advent of expert systems in the field of artificial intelligence (AI) has spawned thousands of projects in which expert knowledge is elicited and preserved (Boose, 1986; Boose & Gaines, 1991; Bramer, 1985; Coombs, 1984; Hayes-Roth, Waterman, & Lenat, 1983; Holsapple & Whinston, 1987; Keller, 1987; Neale, 1988; Stefik, Aikins, Balzer, Benoit, Birnbaum, Hayes-Roth, & Sacerdoti, 1982; Turban & Liebowitz, 1992; Waterman, 1986; Weiss & Kulikowski, 1984). Expert systems are "knowledge-based" software tools or decision aids, intended to assist experts. Seminal work includes MYCIN (Shortliffe, 1976) for diagnosing bacterial infections, PROSPECTOR (Duda, Gaschnig, & Hart, 1979) for determining site potential for geological exploration, and DENDRAL (Feigenbaum, Buchanan, & Lederberg, 1971) for chemical analysis based on mass spectrograms.

Many expert system developers, accustomed to designing and writing programs and not to conducting empirical investigations, had discovered by about 1980 that knowledge elicitation is not easy (Cullen & Bryman, 1988; Ford & Bradshaw, 1993; Wood & Ford, 1993).[1] Indeed, knowledge elicitation can be the most time-consuming and difficult stage in constructing a working program (cf. Buchanan, Sutherland, & Feigenbaum, 1969; Hayes-Roth, Waterman, & Lenat, 1983). The "knowledge acquisition bottleneck" became a focus of the introductions in books and reviews (Diaper, 1989; Hart, 1986; Kidd, 1987; McGraw & Westphal, 1990; Olson & Reuter, 1987; Rook & Croghan, 1989).

Not only did this problem lead to the suggestion that systems engineers should be trained in interview techniques (Forsyth & Buchanan, 1989), but it spawned the development of automated knowledge acquisition "shells." These are toolkits for building prototype expert systems; they do this by having the expert type in answers to questions, automatically generating from the answers a computable representation (e.g., concepts and rules in predicate logic), then integrating the representation with an "inference engine" to control processing (cf., Boose, 1986; Gaines & Boose, 1988; Gevarter, 1987; Johnson, 1985; Neale, 1988; Noble, 1989; for a bibliography, see Hoffman, 1992c).

The creation of an expert system shares something with psychological and applied research. In all cases, the expert (or novice) must be presented some sort of task that taps into their knowledge and skill, that reveals their reasoning and judgement processes, that permits assessment of their performance, and so on. This article will summarize and analyze the various methods of Knowledge Elicitation (KE) that have been used by experimental and applied psychologists and by developers of expert systems. Our concern is not with what makes for sound methodology for the purposes of cognitive research. Our concern is not just with what makes for effective methodology for building expert systems. Rather, our focus is on the question of what makes for useful methodology for the general purpose of revealing, representing, preserving, and disseminating expert knowledge.[2]

[1] For a layout of all the diverse practical and management-related problems and issues that can arise during expert systems development projects, see Brule and Blount (1989), Neale (1988), Prerau (1985, 1989), Tuthill (1990), and Salter (1988).

[2] In this article we will not review the numerous automatic knowledge acquisition toolkits (e.g., KADS, KRITON, GISMO, LAPS, KITTEN, ACQUINAS, ROGET, INFORM, MOLE, TEIRESIAS, EMYCIN, SALT, etc.). However, we will consider the KE tech-

Organization of This Article

We illustrate the methods which have been employed in KE according to three broad categories: analysis of familiar tasks, interview methods, and contrived tasks. We also summarize the results of studies that have systematically compared KE techniques. We conclude with some recommendations and some pointers to outstanding methodological, theoretical, and practical issues.

First, however, we need to say something about the definition of "expert." Clarity about the meaning of expertise can be critical when it comes to KE, e.g., since experts' time is valuable one wants to avoid inefficiency (Adelman, 1989).

DEFINITIONS OF EXPERTISE

In many applications of expert system technology, the process of identifying "the" experts has not been much of a problem in practice (Hoffman, in press). "Experts" have been selected on the basis of years of experience and on the basis of professional criteria (graduate degrees, training experience, publication record, memberships in professional societies, licensing, etc.) (Mullin, 1989). Experts have been selected by virtue of the fact that they held down jobs in operational settings, and by the simple process of asking workers to identify the experts within their organization. Only recently have attempts been made to develop broadly useful, systematic, empirical methods for identifying candidates for KA within an organization, methods borrowing on the techniques of sociogrammetry (cf. Stein, 1992).

Definitions in Psychology

Based on studies of expertise in chess, Lenat and Feigenbaum (1987) estimated that expert knowledge consists of about 50,000 "chunks" (meaningful chess game configurations). Expertise is often defined in terms of memory extent and organization (Glaser, 1987); the partial correlation of expertise with age is a reflection of the amount as well as the type of experience. In a study of expert–novice differences, Chi, Hutchinson, and Robin (1988) relied on the participation of an avid dinosaur fan, a 4-year-old child. In a similar study, Means and Voss (1985) relied on the participation of preschool children who were avid fans of the "Star Wars" films. In some research, college students have served as experts because of their knowledge of particular domains (e.g., football, wedding apparel, regional geography) (Bellezza, 1992). In some studies of mechanics problem solving, graduate students have been the "experts." In general, it takes a long time to become an expert—on the order of a decade, especially in "significant" domains (e.g., airplane piloting, livestock judging, accounting, medical diagnosis) as opposed to more common types of skill, such as reading or automobile driving.

"Expertise" is not a simple category. How individuals are selected for training, how expertise is constituted, and how it is exercised all depend on the domain. We feel that a definition of expertise should not strip the word of its conceptual richness and contextual dependence, but should nevertheless point toward operationalizations. We rely on the traditional terminology of the craft guilds of the Middle Ages, and distinguish a number of levels defined in Table 1. The levels provide context for the meaning of "expertise": Since expertise is a developmental process we cannot possibly learn what we need to know, either about expertise or about knowledge elicitation, by studying only experts (Chase & Ericsson, 1981; Dreyfus & Dreyfus, 1986; Gaeth & Shanteau, 1984; Kolodner, 1983). Indeed, the comparison of expert–novice differences is a paradigm in cognitive research on expertise.

The definition of expertise in Table 1 points in the direction of operationalizable criteria by embracing a number of general factors—experiential, social, cognitive, and performance-related. We are not satisfied with the "naivette" nomenclature since it is diminutive, but something like it is needed for descriptive purposes. Indeed, in much psychological research on expertise, the so-called novices are quite naive to the domain, except perhaps for the knowledge that the domain exists. In some research, the so-called novices have completed introductory-level instruction (i.e., they are beginning an apprenticeship). Throughout this article we focus on studies that seemed to involve experts as defined in Table 1. For studies where it is unclear how expertise was operationalized or how the experts were identified, we can nonetheless confidently refer to the participants as "experienced." In other cases we can confidently use alternative terms from Table 1 (e.g., the participants were journeymen or apprentices). We take such caution for a simple reason—tasks or problems that would challenge a top expert can be incomprehensible to the novice, and problems that novices can comprehend might be trivial to the expert.[3]

niques upon which they rely. In other words, for the purpose of this article we can "decouple" the analysis of KE from the problems of expert system implementation (Alexander, Freiling, Schulman, Rehfuss, & Messick, 1987; Breuker & Wielinga, 1985, 1987; Cleaves, 1987; Motta, Rajan, & Eisenstadt, 1989; Wielinga, & Breuker, 1985; Wiggs & Perez, 1988) and then return with the analysis to look again at implementation issues.

[3] Also, we refer to the experts as "participants" rather than subjects. Experts treated as subjects would quickly find the nearest exit. In psychological research on expertise, as well as expert systems work, the experts are collaborators or even co-investigators.

TABLE 1
A "Guild" Terminology for Development

Naivette	One who is totally ignorant of a domain.
Novice	Literally, someone who is new—a probationary member. There has been some, but minimal, exposure to the domain.
Initiate	Literally, someone who has been through an initiation ceremony—a novice who has begun introductory instruction.
Apprentice	Literally, one who is learning—a student undergoing a program of instruction beyond the introductory level. Traditionally, the apprentice is immersed in the domain by living with and assisting someone at a higher level. The length of an apprenticeship depends on the domain, ranging from about one to 12 years in the craft guilds.
Journeyman	Literally, a person who can perform a day's labor unsupervised, although working under orders. An experienced and reliable worker, or one who has achieved a level of competence. It is possible to remain at this level for life.
Expert	The distinguished or brilliant journeyman, highly regarded by peers, whose judgments are uncommonly accurate and reliable, whose performance shows consummate skill and economy of effort, and who can deal effectively with rare or "tough" cases. Also, an expert is one who has special skills or knowledge derived from extensive experience with subdomains.
Master	Traditionally, a master is any journeyman or expert who is also qualified to teach those at a lower level. Traditionally, a master is one of an elite group of experts whose judgments set the regulations, standards, or ideals. Also, a master can be that expert who is regarded by the other experts as being "the" expert, or the "real" expert, especially with regard to subdomain knowledge.

We can now lay out the palette of alternative KE techniques.

APPROACHES TO KE TAXONOMICS

Spanning such disciplines as expert systems, judgment and decision-making, cognitive science, and ergonomics, one finds a great diversity of techniques used to elicit knowledge. How one makes sense of KE is dependent on one's theoretical inclinations and one's goals or purposes. From the perspective of ergonomics or naturalistic decision-making, for example, one seeks KE methods that possess ecological validity and representativeness, and that can be transported from the laboratory to the field setting. The goals of KE include the generation of cognitive specifications for jobs or tasks, the mitigation of human error in domains involving time pressure and high risk, and the enhancement of proficiency through training, skill remediation, and technological innovation (Kaempf, Thorsden, & Klein, 1991; Klein, 1993; Woods, 1993).

In the expert systems approach, KE is conceived as being one part of the total process of Knowledge Acquisition (KA), which includes KE but also includes representation, implementation (or prototyping), evaluation, and refinement (Buchanan, Barstow, Betchal, Bennet, Clancey, Kulikowski, Mitchell, & Waterman, 1983; Fellers, 1987; Regoczei & Hirst, 1992; Rook & Croghan, 1989; Weitzel & Kirschberg, 1989). The goal for KE is to generate products in a representational format that can be readily molded into a useful implementation (Fellers, 1987; Mettrey, 1987). KE is further constrained by the intended uses of the eventual computer system. There is an emphasis on finding KE methods that allow one to get at the "important" knowledge directly, since a major consideration is the quality and validity of a knowledge base that results from KE.

For psychological research, a KE method must make sense as a way of revealing reasoning strategies and sequences, facts about knowledge organization, etc. A classification scheme for analyzing and comparing various KE methods needs to reflect cognitive functionality (i.e., tasks that are good for eliciting tacit knowledge or perceptual judgments, tasks that are good for eliciting procedural knowledge, etc.) (Breuker & Wielinga, 1984, 1987; Fischhoff, 1989; Johnson, Zualkerman, & Tukey, 1993).

Following Waterman (1986), a number system developers have divided KE methods into two simple categories, indirect (i.e. getting knowledge from texts, reports, etc.) and direct (i.e. observing expert behavior and asking questions about their reasoning) (Fellers, 1987; Geiwitz, Klatzky, & McCloskey, 1988; Olson & Reuter, 1987). Also with an eye toward simplicity, our typology is intended to facilitate discussion. We place KE methods into three categories: (a) analysis of the tasks that experts perform, (b) various types of interviews, and (c) contrived techniques. These can be paraphrased as: What do experts usually do? What do experts say they do? and What do they do when they are constrained in some new way?

ANALYSIS OF FAMILIAR TASKS

In the analysis of familiar tasks, one investigates what experts do when they conduct their usual problem solving or decision-making tasks.

Documentation Analysis

Whenever one is commencing research on a domain one must start by becoming conversant by reading

texts, manuals, by taking courses, etc. But the purposes of the researcher/learner do not preclude the analysis of documentation as a means of commencing KE. Rather than just having information flow from documents into the researcher's understanding, the researcher's analysis of the documents can involve specific procedures that generate records or analyses of the knowledge contained in the documents.

This can be a time-consuming process, but can sometimes be indispensible in KE (Kolodner, 1983). In a study of aerial photo interpreters (Hoffman, 1987), interviews about the process of terrain analysis began only after an analysis of the readily available basic knowledge of concepts and definitions. To take up the expert's time by asking questions such as "What is limestone?" would have made no sense.

Beyond documentation analysis, the analysis of familiar tasks consists of a suite of techniques with two major components: "task analysis" and "protocol analysis." Task analysis is primarily concerned with on-line activity, whereas protocol analysis is primarily concerned with reasoning during on-going performance.

Task Analysis

Human factors psychologists have studied task performance in a great variety of contexts (cf. Anastasi, 1979; Sanders & McCormick, 1987). Task analysis goes by many names, including job analysis, structural analysis, and task description, to name just a few (Fleishman, 1975; Kirwan & Ainsworth, 1992; Meister, 1985). Task analysis can have a variety of purposes: To describe jobs and identify subtasks, to study manufacturing or process control procedures, to establish ergonomic constraints on equipment design, to develop training programs, or to yield job specifications or objectives (Eastman Kodak Company, 1983; Wexley & Yukl, 1984).[4]

Although the traditional literature on task analysis did not focus on knowledge-based applications, the task analysis component of the analysis of experts' familiar tasks is illustrated in a number of the examples that will follow.

Think Aloud Problem-Solving/Protocol Analysis

In this technique, not only are job's task activities charted, but the problem solver is instructed to "think aloud." The think aloud procedure generates a protocol—a recording of the deliberations that can be transcribed and analyzed for propositional content. This method (combining the think aloud task with the protocol data analysis procedure) has been used successfully in numerous studies of problem solving (e.g., Belkin, Brooks, & Daniels, 1987; Benjafield, 1969; Johnson, Zualkerman, & Garber, 1987; Newell & Simon, 1972). An example is Bailey and Kay's (1987) study in which adults built a lifting device using a child's construction set. Performance was analyzed in terms of specific actions (e.g., bolting two parts together); verbalizations were categorized according to reference (e.g., the goals of a particular action being planned, or the evaluation of the outcome of a test of a component). On the basis of the references, actions could be grouped into behavioral episodes, revealing the pattern of goal decomposition that had been utilized by the problem solver. Also apparent were behaviors reflecting disorganized or out-of-sequence thoughts and actions. (For a discussion of the use of "timelines" or "decision trees" in the analysis of KE task performance data, see Klein, Calderwood, & MacGregor, 1989; McGraw & Riner, 1987.)

It should be noted that psychological research suggests that the process of verbalization does not typically cause dramatic interference, that is, it does not significantly affect the normal course of cognitive processes (Ericsson & Simon, 1993), and it can yield information about the reasoning sequences and goal structures in experts' problem solving (Wood & Ford, 1993). However, caution is in order regarding individual differences in verbal expressiveness (Burton, Shadbolt, Rugg, & Hedgecock, 1990; Ericsson & Simon, 1993).

The think aloud problem-solving/protocol analysis technique has been used extensively in cognitive research on expertise—medical diagnosticians (e.g., Johnson, Duran, Hassebrock, Moller, Prietula, Feltovich, & Swansson, 1981; Kuipers & Kassirer, 1984; Kuipers, Moskowitz, & Kassirer, 1988; Patel & Groen, 1986), physicists (e.g., Chi, Feltovich, & Glaser, 1981; Chi, Glaser, & Rees, 1982), computer programmers (e.g., Jeffries, Turner, Polson, & Atwood, 1981), process controllers (Bainbridge, 1979, 1981; Umbers & King, 1981), and accountants (Dillard & Mutchler, 1987).

An example of protocol analysis used explicitly for KE is a study by Fox, Myers, Greaves, and Pegram (1987) (see also Kuipers & Kassirer, 1987). Experts at diagnosing leukemia were presented with the records of a number of patients and were instructed to think aloud while coming to a diagnostic decision. From the experts' deliberations, a number of propositions were extracted, some of which referred to factual information, some referred to reasoning rules. These proposi-

[4] The present topic nearly forces one to use the word "task" in two senses. One is the sense of job requirements in human factors psychology, the other is the sense of the procedure used in knowledge elicitation or cognitive research. Although the two meanings usually disambiguate in context, we will consistently refer to "familiar task" versus "KE method" or "KE technique."

tions were then used as the basis of a prototype expert system.[5]

To conduct a study in which experts' performance in their familiar tasks is examined, one must select particular problems or cases to present to the expert.

Materials for Task Analysis and Think Aloud Problem Solving

Materials can come from a number of sources.

Test cases. Since experts often reason in terms of their memories of past experiences or cases (their "war stories") (Kolodner, 1991; Slade, 1991; Wood & Ford, 1993), hypothetical problems called test cases can be used in task analysis and think aloud problem solving (e.g., Kidd & Cooper, 1985; Prerau, 1989). Test cases can be generated from archived data or can be generated by other experts. A set of test cases is sometimes intended to sample the domain, sometimes intended to focus on prototypical cases, sometimes to sample along a range of difficulty. For example, Senjen (1988) used archived data to generate test cases of plans for sampling the insects in orchards. The test cases were presented to expert entomologists, who then conducted their familiar task—the generation of advice about the design of sampling plans.

Tough cases and atypical cases. Occasionally, experts come across a particularly difficult or challenging case. Reliance on tough cases in KE can be more revealing than observing experts solving common or routine problems (Klein & Hoffman, 1993). Mullin (1989) emphasized the need to select so-called well structured test case problems to reveal ordinary "top-down" reasoning and using so-called ill structured or novel test case problems in order to reveal flexible or "bottom-up" reasoning.

Hoffman (1987) tape recorded the deliberations of two expert aerial photo interpreters who had encountered a difficult case of radar image interpretation. The case evoked deliberate, pensive problem solving and quite a bit of "detective work." In this way, the transcripts were informative of the experts' refined or specialized reasoning.

But tough or atypical cases do not occur predictably. One can overcome this potential obstacle to KE by probing archived material for what might be tough cases. Alternatively, the expert may be asked to make tape recordings whenever "interesting" problems are encountered. A third possibility, to be discussed in more detail below, is to ask the expert to recall interesting or tough cases from their own past experience (Klein, Calderwood, & MacGregor, 1989).

The second major category of KE methods is Interviews.

INTERVIEWS

As with the analysis of familiar tasks, interview techniques come in a variety of forms (Gorden, 1987).

Unstructured Interviews

Unstructured interviews usually take the form of an open dialogue in which an interviewer asks open-ended questions about an expert's knowledge and reasoning: "Tell me everything you know about X." In one appropriate usage, initial unstructured interviews allow one to gain an overview of the domain. At successive meetings, increasing structure can be imposed on the interviews. The researcher will undertake a number of elicitation sessions, hoping to obtain a comprehensive coverage of the domain. In this sense, an unstructured interview is not disorganized or unplanned (Wood & Ford, 1993).

Interviews require some form of record-taking. Although note-taking can be sufficient, it is common to make an audio tape recording, the transcription of which is notoriously time-consuming. Each hour of interview can take as much as a full working day for an expert typist to transcribe (Hoffman, 1987). Interviews in expert KE are usually exhaustive and exhausting, a contributing factor in the KA bottleneck. Many expert system developers have utilized unstructured interviews as the main method for KE (Cullen & Bryman, 1988) (e.g., Weiss & Kulikowski, 1984), apparently taking it for granted that this is the primary or even the only way to elicit experts' knowledge (Kidd, 1987). It is certainly implied that that unstructured interviewing can constitute a sufficient program of elicitation.

However, the method does have potential pitfalls (Forsyth & Buchanan, 1989; Hoffman, 1987). For example, the expert can get side-tracked, or may assume that the elicitor has knowledge which she/he has not. Furthermore, the analysis of resultant transcripts can be especially difficult if the interview was disorganized. The widespread use of interviews in an unprincipled fashion has led a number of researchers (e.g., Basden, 1989; Forsyth & Buchanan, 1989; Hart, 1986; McGraw & Seale, 1988; Meyer & Payton, 1992; Monk, 1985; Olson & Reuter, 1987; Rolandi, 1986) to point out to the broader expert systems community that there is a large social–psychological, anthropological, and ethnomethodological literature on interviewing techniques (e.g., Benfer & Furbee, 1989; Forsyth & Buchanan, 1989; Spradley, 1979).

[5] Although Fox *et al.*'s analysis of the performance records provided a great deal of information that could be used in building the knowledge base, they were able to see that some aspects of the diagnosis task were not elicited. A documentation analysis was needed in order to complete the prototype system.

A common conclusion, long known to psychologists, is that in an interview, structure can help.

Structured Interviews

Structured interviewing can reduce time spent relative to unstructured interviewing (Sjoberg & Nett, 1968). Research also suggests that training in structured interviewing techniques can significantly raise the proficiency of an interviewer. Of course, there are also issues of social dynamics as well as the possible effects of interviewer personality on the interaction (Brown, 1989; McGraw & Seale, 1988; Waldon, 1989; Wood & Ford, 1993). (For discussions of social communication issues in interviewing for KE, see Forsyth & Buchanan, 1989; Sternberg & Frensch, 1992.)

As elaborated by many social scientists (e.g., Gorden, 1987; Lerner, 1956; Merton & Kendall, 1946), structured or "focused" interviews are designed and planned in advance. An agenda is set, and the purpose of each session and the roles of interviewer and expert are clearly defined. The structured interview is a class of techniques. The two most common formats rely on either domain-specific probe questions or on generic probe questions (Wood & Ford, 1993).

Domain-specific probe questions. In using domain-specific questions, the interviewer prepares a fixed set of questions about the domain of interest. An extensive research program using this kind of structure was conducted by Merton and his colleagues (Merton, Fiske, & Kendall, 1956; see also Herzog, 1944) in the areas of communications (e.g., radio programs, wartime propaganda films) and psychotherapy. As a result of their studies, Merton *et al.* (1956) proposed that a set of questions for structured interviews should cover a broad range of particulars within the domain and be carefully worded so as to avoid suggesting particular answers or imposing the categories or biases of the interviewer.

The creation of probe questions necessitates some prior analysis of the domain. Merton and Kendall (1946) referred to this as content analysis, the general idea being that a documentation analysis or an analysis of familiar tasks can provide information needed for structuring the interviews (Hoffman, 1987; McGraw & Riner, 1987; McGraw & Seale, 1988).

Generic probe questions. The second type of structure relies on a set of generic questions. These are not necessarily specific to a domain, and their order is not completely predetermined, but they do have specific functions (cf. Johnson & Johnson, 1987). Examples are: "What is the difference between an X and a Y?" or "Can you give me any more examples of class X?" where X and Y are domain elements previously mentioned in the interview. Shadbolt and Burton (1990a) and Wood and Ford (1993) have provided templates for KE structured interview probe questions, summarized here in Table 2.

The use of generic questions requires that the interviewer be vigilant to the interview process (as opposed to maintaining an attentional focus on their own understanding of the domain). However, analysis of the interview transcript can be facilitated by the fact that each question has a particular function. Furthermore, generic probes can be useful in the validation, refinement, and extension of knowledge that has already been elicited.

Structure can be added to an interview not only by

TABLE 2
Probe Set and Functions (after Shadbolt and Burton, 1990; and Wood and Ford, 1993)

Probe	Function
For domain overview	
Could you tell me about a typical case?	Provides an overview of the domain tasks and concepts.
Can you tell me about the last case you encountered?	Provides an overview of the domain tasks and concepts.
For domain concepts	
Can you give me an example of X?	Reveals and clarifies domain concepts.
What is the difference between X and Y?	Contrast questions clarify domain concepts.
Does X include Y?	Relationship questions clarify the interrelationships of domain concepts.
For domain procedures and reasoning rules	
Why would you do that?	*Converts an assertion into a rule.*
How would you do that?	Generates rules or procedures.
What do you do at each step in this procedure?	*Provides information about* procedural details.
When would you do that?	Reveals the scope of a rule or procedure.
Is [the rule] always the case?	As above, may generate other rules.
What alternatives [to the prescribed action or decision] are there?	Generates more rules or procedures.
What if it were not the case that [currently true condition]?	Generates rules for when current condition does not apply.
For refinements of the knowledge base and the elicitation of knowledge about special procedures	
Can you tell me about an usual case you encountered?	Refines the knowledge base to include rare cases, special procedures.
Can you tell me about an usual case you heard about from some other expert?	Refines the knowledge base to include rare cases, special procedures.

preplanning the questions, but also by preplanning the material that the questions are about.

Using test cases. A number of expert system developers have relied on test cases to add structure to interviews (e.g., Grover, 1983; Kidd, 1987; Mullin, 1989). Rolandi (1986) and Prerau (1989) recommend that KE for expert system development should consist of interviews with experts on a dozen or so test cases.

Using a "first-pass knowledge base." Another method of adding structure to interview materials is to use a knowledge representation called a first-pass knowledge base (Young & Gammack, 1987), derived from a documentation analysis, task analysis, or initial unstructured interviews. A first-pass knowledge base is essentially a meaningfully organized list of propositions that express many of the core concepts, the definitions of terms, and the procedural rules that are followed in the domain. In the interview, the expert goes over the entries in the first-pass knowledge base and comments on each one, suggesting additions, deletions, refinements, etc. (Hoffman, 1987).

Event recall interviews. Experts often have clear memories for tough or salient cases they have encountered, and they sometimes reason by analogy to past cases (Kolodner, 1991; Slade, 1991). Related to the utility of using test cases in task analysis, it can be useful to structure an interview by having the expert recall past events or cases, which can then be the focus of probe questions intended to facilitate recall (e.g., "Try to go through the events in reverse order" and "Try to recall the incident from different perspectives").[6]

Since World War II, human factors psychologists have used event recall interviews in studies of "critical incidents" of equipment failure or operator error (Flanagan, 1954). Event recall interviews are also an important method in the study of eyewitness memory and testimony, as well as police interviews (Bruce, 1988; Deffenbacher, 1988), and some important lessons are to be found in that literature. Geiselman, Fisher, MacKinnon, and Holland's (1985) comparison of alternative interview formats suggests that adding structure to event recall can make an interview more productive than a standard police interview.

A clear example of event recall-based reasoning by experts comes from avionics engineering. In "comparability analysis" an expert predicts the reliability and maintainability of new aircraft components or systems on the basis of functionally or structurally similar components on older aircraft (Tetmeyer, 1976). Klein and Weitzenfield (1982; Klein, 1987) had expert avionics engineers perform this familiar task for some test cases (e.g., the specifications for the hydraulics system on a new airplane). As the experts performed their familiar task, they were probed with a set of preplanned questions.

Klein, Calderwood, and MacGregor (1989) have found that reasoning in terms of the probed recall of past tough and salient cases can be very effective in revealing experts' knowledge, especially their tacit knowledge and reasoning strategies. Hence, Klein *et al.* have expanded this method into a KE technique they call the Critical Decision Method and have used it successfully in the study of expertise in such domains as fire fighting, design engineering, paramedicine, nursing, and military command and control. Furthermore, the method is reliable in that there are high rates of agreement between people who independently code the interview transcripts.

Group interviews. Reliance on multiple experts may be unavoidable if the experts in a domain have differing areas of specialization. Furthermore, reliance on more than one expert can be necessary to assess the reliability or importance of particular aspects of knowledge or reasoning (Hoffman, 1987; Mittal & Dym, 1985; Wolf, 1989). In the development of many expert systems, the prototype is assessed by more than one expert to suggest changes that might help ensure end-user acceptance (Cochran, Bloom, & Bullemer, 1990).

A special problem for the group interview approach in expert systems work is that an expert system must have nonconflicting rules. And yet, when two experts come together they can see it as their job to seek out areas of disagreement. Even if they agree on 99% they will quickly find and argue about the 1% (Adelman, 1989; Agnew, Brown, & Lynch, 1986; Hoffman, Slovic, and Rorer, 1968; Libby and Lewis, 1977). Although judgment scaling methods (to be described later) can help one deal with conflicting viewpoints, for the purpose of constructing a coherent knowledge base there can be a trade-off between group size and the optimality of consensus. Based on their work with experts in aviation systems, McGraw and Seale (1988) recommended using groups of only two or three experts.

This concludes the illustration of interview methods.

[6] We use the term event recall rather than the term "retrospection." Technically, introspections are verbal statements that represent judgments about mental phenomena. The concept of retrospection emphasizes the fact that all introspection depends on memory. The term retrospection does not denote a task that is somehow different from introspection. In KE procedures such as think aloud problem solving, an expert actually spends little time making judgments *about* his or her mental phenomena or memories. During a problem-solving task, the participant mostly describes the things that are involved in the problem (Newell and Simon, 1972; Woodworth, 1938). Sometimes in a KE procedure an expert may spontaneously make an introspective judgment, say, about the confidence of a particular memory. Some KE methods, to be discussed below, require the expert to make such judgments. However, in event recall, the expert simply recalls events.

We now turn to the third general category of KE methods—"contrived techniques."

CONTRIVED TECHNIQUES

Psychological research on expertise shows that deliberate modification of the familiar task can reveal the expert's knowledge and reasoning. This has been demonstrated, for example, by asking chess masters to recall game boards in which the pieces had been randomly arranged (Chase & Simon, 1973; DeGroot, 1966), and by making bridge players adhere to altered rules (Sternberg & Frensch, 1992). However, there is disagreement on how much departure is legitimate or fruitful in KE. Unfamiliar tasks can make the expert feel uncomfortable (Hoffman, 1987; Klein & Weitzenfield, 1982; Schweickert, Burton, Taylor, Corlett, Shadbolt, & Hedgecock, 1987). Furthermore, some contrived techniques may reveal reasoning strategies that have little to do with the experts' usual modus operandi (Fischhoff, 1989; Salter, 1988).

However, research by Shadbolt and his colleagues (Burton, Shadbolt, Rugg, & Hedgecock, 1988; Shadbolt & Burton, 1989, 1990b) (to be described in detail below) has shown that departure from familiar tasks can be very informative in KE. Moreover, contrived techniques can yield knowledge efficiently.

Decision Analysis

Decision analysis is a set of procedures including decision aiding, risk analysis, and the sorts of analyses that involve utility and probability modeling based on inputs provided by the expert (Fischhoff, 1989). In most applications of decision analysis, experts yield evidence about the sequence of steps in their usual decision making by generating a number of lists: (a) the components or elements of problems, (b) the causal relations of the components, (c) the different kinds of problems encountered, (d) the characteristics (i.e, parameters or boundary conditions) of each problem type, (e) the kinds of decisions each problem type involves, (f) the problem solver's confidence in judgments or hypotheses, (g) the possible consequences of each decision, and (h) evaluations of the quality of the analyses.

From all of the lists and judgments one can generate not only a mathematical model of reasoning but also a dictionary of key concepts and an inference network or decision tree. Both of these can be incorporated in decision aids or expert systems (Fischhoff, 1989; Hart, 1986; Tolcott, Marvin, & Lehner, 1989).

Group Decision Making

Some contrived KE methods involve the use of small groups of participants. One group method that has been used in expert KE is "brainstorming" (Osborn, 1953). In a brainstorming session, the participants are challenged to generate creative decisions. The trick is to have a group "facilitator" and to separate the process of generating solutions from the process of criticizing and refining solutions. McGraw and Seale (1988) used the brainstorming method to structure an interview with a group of experts, and they examined two other group KE methods as well. The second was "consensus decision making." In this kind of problem-solving group, the group's goal is to find the "best" solution to a problem by assessing the advantages and disadvantages of alternative solutions. A third group technique studied by McGraw & Seale is called the "nominal group." In such a group, the individuals are given a list of alternative solutions and they perform independent (confidential) ratings of advantages and disadvantages. McGraw and Seale reported that all of these small group problem-solving methods were useful in eliciting experts' knowledge (see also Adelman, 1989).

Rating and Sorting Tasks

In some domains, ratings or rankings are performed as a part of the familiar tasks. But in many domains they are not, and so we include ratings and rankings in the contrived category of KE methods. A number of researchers have used rating tasks in the psychological study of expertise (e.g., Einhorn, 1974; Gammack, 1987; Johnson, Hassebrock, Duran, & Moller, 1982; Shaw & Gaines, 1987). For example, Hammond, Hamm, Grassia, and Pearson (1987) studied expert highway engineers by using a rating familiar to experts. They judged particular roads' aesthetic values (an intuitional judgment), predicted the roads' accident rates (a rating requiring subjective judgment and formal analysis), and estimated the roads' carrying capacity (a judgment which usually relies mostly on calculation). Some of the participants were required to think aloud while they generated judgments. Information about roads was presented in one of two forms, either slides showing views of roadways or a bar graph depicting a number of road variables (e.g., lane width, angle of the steepest gradient, etc.). By manipulating stimulus and judgment type, Hammond *et al.* were able to explore the ways in which various combinations of tasks and materials could induce either "intuitive" or analytical reasoning strategies (e.g., presenting information via a film strip was more supportive of aesthetic judgment). They also found that intuitive reasoning did not lead to poorer performance than analytical reasoning (see also Howell, 1984; Prietula, Feltovich, & Marchak, 1989).

In a study of expert livestock judges, Phelps and Shanteau (1978) used a rating task to reveal those

variables on which experts relied. Although experts' decisions can be very complex (Ceci & Liker, 1986), some research has shown that decision makers often do not rely on all of the relevant and important information that is available (Slovic, 1969). The usual task of livestock judges is to inspect livestock (or photographs) and make ratings of breeding quality. In the experiment, one condition simulated the experts' familiar task, that is, experts were presented with photographs of gilts (female pigs) and had to rate them for breeding quality. In another condition, the experts were presented only with descriptions of gilts in terms of 11 variables (e.g., weight, freeness of gait, etc.).

Comparison of the two conditions revealed that the judges relied on most of the variables when given the descriptions. Interviews with the experts revealed their underlying strategy. The variables are naturally grouped into three sets, those referring to size, those referring to meat quality, and those referring to breeding quality. In making their judgments, the experts would first make an assessment for each of the three groups, and then generate an overall judgment. By being given the descriptions alone, the experts were forced to consider each variable systematically.

The rating task has been used in a number of psychological investigations of expert–novice differences (e.g., Butler & Corter, 1986; Chi *et al.,* 1981; Weiser & Shertz, 1984). Schvaneveldt, Durso, Goldsmith, Breen, Cooke, Tucker, and DeMaio (1985) have developed an algorithm for graphical representation of rating results, which has been used to reveal expert-novice differences in fighter pilots. The core finding is that experts classify domain elements by meaningful dimensions, whereas novices tend to classify along dimensions of more superficial characteristics (see also McKeithen, Reitman, Rueter, & Hirtle, 1981).

A variation on the rating task is the sorting task. The usual mode is to have the participant sort cards bearing the names of problem domain elements. The piles are then labeled according to particular features or dimensions (e.g., similarity, shared functions, causal links, etc.). Chi *et al.* (1982) used this technique to examine expert–novice differences in physics problem solving. Similarly, Weiser and Shertz (1984) used the technique in the study of expert and novice computer programmers. Overall, research results have been similar to those of Schvaneveldt *et al.* (1985): Experts sort cards along semantically important dimensions of the domain whereas novices often sort by similarity of literal or superficial features (for research reviews, see Chi *et al.,* 1988; Cooke, 1992; Ericsson & Smith, 1991).

Another rating task is Kelly's (1955) repertory grid technique. In the original application in personality research, participants were asked to provide a list of people who were important in their lives. They were then asked to provide dimensions on which these people may be rated, for example, friendliness, laziness, height, and so on. These dimensions (or "constructs") are often elicited using a triadic method: Participants are asked to pick out three people, two of whom are regarded as similar on some dimension.

The repertory grid has been used in a number of psychological investigations (see Fransella & Bannister, 1971, for a review of the method; Collett, 1979, for a review of applications). For example, the technique has been used to explore students' choice of university (Reid & Holley, 1972), people's perceptions of environment and architectural constructions (Harrison & Sarre, 1975), and the subjective consequences of urbanization (DuPreez & Ward, 1970).

Most psychological experiments that have relied on judgment, rating, or sorting tasks have been designed to explore particular hypotheses about expertise or subjective judgment, rather than to elicit expert knowledge. However, these tasks are being used in knowledge engineering: Rating tasks form the heart of many automated KA tools (e.g., Boose & Bradshaw, 1987; Ford & Adams-Webber, 1992). In constructing a ratings grid, experts first list the basic concepts that apply to the case (e.g., different types of disease). They then rate each concept on a number of dimensions (e.g., the symptoms or features of diseases).[7]

Rating and sorting data can be analyzed and data from different experts compared using a variety of statistical techniques including classical procedures such as analysis of variance and χ^2. Analysis methods also include multidimensional scaling, factor analysis, cluster analysis, and principal component analysis (see Cordingley, 1989, or Gammack, 1987, for reviews). In such analyses, the ratings provided by experts are taken to represent measurements of the semantic proximity of domain elements. Thus, from scaling or rating data a conceptual map of the domain can be derived. This can be useful for KE when there is no preexisting specification of the domain content (Regoczei & Hirst, 1988). But as with all KE methods, there can be difficulties.

One is that the data can yield "spurious accuracy." For example, it may be unimportant that two rocks are 2.57 units apart on some scale; the important linking features may categorical (Cooke & MacDonald, 1986). Another difficulty is the creation of spurious categories arising from the interpretation of multidimensional

[7] Examples of the use of repertory grids for KE in expert systems development can be found in Boose (1985, 1986); DeMantras, Cortes, Manero, Plaza, Salra, and Agusti (1983); Kim and Courtney (1988); and Shaw and Gaines (1987). General discussions of the use of ratings for KE in expert systems development are provided in Chignell and Peterson (1988), Garg-Janardan and Salvendy (1987), Hart (1987), and Shaw and Gaines (1987).

data. For example, in cluster analysis (the most popular technique for analyzing repertory grids), it can sometimes be difficult to label the clusters with any assurance of content validity. This can be particularly true in domains of high dimensionality (Gammack, 1987).

The various statistical techniques each make numerical assumptions which can be inappropriate for KE. However, interpretation problems are not always catastrophic. For example, Cooke and McDonald (1987) presented a set of Unix commands to expert programmers and performed a cluster analysis on the results of ratings. The cluster analysis was then given to expert programmers who were quite consistent in assigning labels to the clusters. Similarly, Schvaneveldt *et al.* (1985) have shown that the output from their clustering algorithm is easily understood (and remembered) by domain experts. In demonstrating that it is possible to use rating and clustering methods to elicit and meaningfully describe the knowledge of experts, these studies serve as a reminder that in the examination of expertise, statistical comparisons of groups or conditions are subsidiary to the goal of understanding.

Constrained Processing and Limited Information Problems

In the constrained processing method, familiar routines are constrained in some way. The expert may be explicitly instructed to adopt a particular strategy, for example. Hoffman (1987) had expert terrain analysts inspect aerial photos for only 2 min—aerial photo interpretation ordinarily takes hours, even days. Following the brief inspection period, the experts had to recall everything they could about the photos and provide their interpretation (e.g., "this region is a tropical climate with shallow soils overlying tilted interbedded sandstone and limestone"). Results from this constrained processing problem revealed the extent to which the experts achieve immediate perceptual understanding of terrain when viewing aerial photos.

A constrained processing technique that has been recommended for use in KE involves combining on-line task performance with the use of probe questions, a technique called interruption analysis (Salter, 1988). During task performance, the expert can be asked, for example, "What were you just doing?" or "What was just going on?" or "What would you have done just then if . . . ?"

The limited information technique does not necessarily yield an overview of domain knowledge, but it can elicit detailed or subdomain knowledge. In one version, experts are asked to solve problems given incomplete information. Selective withholding of information can be used to reveal experts' strategies and reasoning sequences in different situations (e.g., Hoffman, 1987). For example, Tolcott, Marvin, and Lehner (1989) had expert Army battlefield intelligence analysts think aloud while reasoning about scenarios. On the first presentation of a scenario, the information was limited, but over a series of trials additional information was provided to see if it would lead to the formation of alternative hypotheses, changes in confidence judgments, etc.

A contrived technique illustrating both constrained processing and limited information is called "20 Questions" (Grover, 1983). The expert is provided with little or no information about a particular problem to be solved and must ask the elicitor for information needed to solve the problem. The information which is requested along with the order in which it is requested can provide the researcher with an insight into the experts' problem-solving strategy. One difficulty with this method is that the researcher needs a firm understanding of the domain in order to make sense of the experts' questions and provide meaningful responses. A way around this is to use two experts, one as interviewer and one as interviewee. The 20 Questions method has been used successfully in expert KE (Schweickert *et al.*, 1987; Shadbolt & Barton, 1990a).

Graph Construction

A "conceptual graph" is a representation of a domain or problem in terms of the relationships (links) between domain elements (concept/nodes). Such representations are useful in AI and interface design (Sowa, 1984)—an instance of the widely recognized utility of using graphical displays to convey information (Bauer & Johnson-Laird, 1993; Wickens, Merwin, & Lin, 1994). Gordon, Schmierer, and Gill (1993) compared the conceptual graph technique (along with the use of probe questions) to the use of explanatory text to assess the materials in terms of facilitation of problem-solving performance on test cases (in the domain of engineering, using students as participants). The conceptual graph materials won out.

In using conceptual graphs (or "laddered grids") in KE, the expert and the elicitor work together to construct a graphical representation of the domain in terms of the relationships (links) between domain elements (concept nodes) (Adelman, 1989; Hinkle, 1965; Shadbolt & Burton, 1990b). As the KE process continues, attributes of nodes can be used as the basis for starting new graphs. Conceptual graphing has formed the heart of some automated KA tools that support the transformation of diagrammed information into rules or other knowledge representation formats (Berg-Cross & Price, 1989; Major and Reichgelt, 1990; Motta, Eisenstadt, West, Pitman, & Evertsz, 1987; Motta, Ra-

jan, & Eisenstadt, 1989; Solvberg, Nordbo, Vestli, Aakvik, Amble, Eggen, & Amodt, 1988).

This concludes our illustration of each of the various methods that have been used either to elicit or study the knowledge of experts.

OVERVIEW OF TECHNIQUES

The multidisciplinary literature on KE suggests that most researchers and system developers have used combinations of methods. For example, the Hammond *et al.* (1987) study of expert highway engineers combined think aloud problem solving with a judgment task. Fox *et al.* (1985) combined think aloud problem solving with the analysis of familiar tasks. Klein's (1987) Critical Decision Method combines a probe question-structured interview with the recall of tough or salient past cases. Hoffman (1987) combined think aloud problem solving with limited information problems. Wood and Ford (1993) conducted a think aloud problem-solving task and followed it with a structured interview. Gordon *et al.* (1993) combined conceptual graphing with probe questions.

There seem to be countless possibilities, but each of the methods can be regarded as involving one or more particular type of materials and one or more particular procedures. This has been implicit in the organization of our discussion and is specified in Table 3.

This sets the stage for the next sections, in which we discuss empirical evaluations of KE methods in two distinct categories: (1) case studies of uses of methods in the development of expert systems and (2) systematic empirical studies.

TABLE 3
A KE Methods Classification

Participants
Experience level
Naivette, Novice, Trainee, Journeyman, Expert, Master
Groupings
Individuals, small groups, working groups
Procedure
Familiar task activities
Task analysis, unobtrusive observation, simulated familiar tasks
Interviews
Unstructured, structured (by probe questions, test cases, first-pass knowledge base)
Contrived techniques
Event recall, think aloud problem solving, creative problem solving, decision analysis, scaling tasks, sorting tasks, rating tasks, constrained processing tasks, limited information tasks, graph generation tasks
Materials
Familiar task materials, limited information materials, probe questions, first-pass knowledge bases (e.g., results from a documentation analysis), archive-based test cases, test cases generated by experimenter, tough case materials, salient case materials, critical incident records

CASE STUDIES ON KNOWLEDGE ELICITATION METHODS

Reports on expert system development projects sometimes contain information about the developers' experiences. Such reports usually appear in AI sources such as technical reports, conference proceedings and books on "how to do" KE (see Hart, 1986; McGraw & Harbison-Briggs, 1989; Footnote 1). System developers sometimes reflect on the KE techniques they used. For every technique we have mentioned so far there is at least one case study (see Dhaliwal & Benbasat, 1990; Fellers, 1987). Example reports, on either the development of expert systems or on automated KA, spanning the variety of KE methods are Clarke (1987); DeGreef, Breuker, Schreiber, and Wielemaker (1988); de Mantaras, Cortes, Maero, Plaza, Salra, and Agusti (1986); Gale (1987); Kidd and Cooper (1985); Kolodner, (1991); Mitchell (1987); Prerau (1989); Smith and Baker (1983); and Trimble and Cooper (1987).

We Believe It Worked Well

Despite the wide coverage, it is difficult to make comparisons. Many recommendations are idiosyncratic and quite a few are contradictory. One system developer might recommend the use of small groups, but another may claim that the use of groups can be disastrous. Another may claim that experts should never interview other experts. One system developer may argue that the knowledge engineer should become a domain expert (cf. Friedland, 1981; Taylor, 1985), while another developer may argue the converse, that the expert should be trained to be a knowledge engineer. Yet, another developer may believe that the more an expert knows about AI, the more likely it will be that the expert will develop inappropriate biases (e.g., about which knowledge representation format should be favored) (cf. McIntosh, 1986).

Typically, system developers "name their methods without providing much information about how to apply them" (Forsyth & Buchanan, 1989, pp. 1–2). It is not enough to know that a particular technique performed to someone's satisfaction in the development of an expert system. The success may have been due to a variety of social and psychological factors. The efficacy of a technique may depend on the particular kind of knowledge to be elicited, on the personality or psychometric characteristics of the expert or elicitor, on the way in which the knowledge was to be implemented in

a program, and so on (Adelman, 1989; Fleishman, 1975).

> . . . most expert systems are developed for one problem domain using only one expert, one knowledge engineer, and one elicitation method for a predetermined knowledge representation scheme. The generalizability or validity of such systems is questionable, for there is minimal (if any) research demonstrating that these sources of variability do not significantly affect the quality of the knowledge base. (Adelman, 1989, p. 483)

(Human) KE versus (Machine) KA

KE techniques involving human interaction (e.g., interviews, think aloud problem solving, etc.) have also been contrasted with automated KA tools, the focus being on the relative advantages and disadvantages of alternative tools (e.g., some elicit knowledge more easily, some are less user-friendly, some must be run on a LISP machine, some allow systems to be built easily but the systems require more testing, etc.) (Mettrey, 1987). For example, Neale (1988) compared a modest palette of KE methods with some approaches to automated KA and concluded that: (a) human-on-human KE methods (interviews, protocol analysis) place an "unjustified faith in textbook knowledge and what experts say they do" (p. 135) and (b) are also time-consuming. Hence, Neale sees "a strong trend towards eliminating the knowledge engineer and getting the expert to work directly with the computer" (p. 136).

A number of expert system developers have utilized a KE method and an automated KA tool and have then asked experts to make judgments about the results (e.g., ratings of knowledge- or rule-based quality, complexity, and completeness; ratings of interface quality, etc.). Dhaliwal and Benbasat (1990) reviewed a number of technical reports on such work, and concluded that automated KA is more efficient, less subject to the effects of the skill level of the human interviewer, more likely to yield valid knowledge, preferable because it requires the expert to decompose their domain into elements, and, of course, preferable because the elicited knowledge comes in a form ready to implement in a prototype. Echoing this, Kim and Courtney (1988) argued that the need for bootstrapping and the chance for miscommunication make the human-oriented methods less preferred than automated KA.

Michalski and Chilausky (1980) compared the performance of two expert systems for the diagnosis of soybean diseases, one system based on rules derived from an unstructured interview with an expert plant pathologist, and the other based on rules "learned" by algorithmic induction from a set of examples. In terms of correct diagnosis of a set of test cases, the algorithm performed better (about 97% versus 72% correct), a result that is, alas, clouded by the lack of experimental controls.

TABLE 4
Some Potential Advantages and Disadvantages of KE Methods, Based on Case Study Experience

Methods	Advantages and disadvantages
Analysis of familiar tasks	+ Instructive to system developer – Reveals what experts do but does not necessarily reveal what experts know – Can be time consuming
Unstructured interviews	+ Useful in social facilitation + Useful in determining user needs – Can be time consuming and laborious +/– Can yield some information about domain concepts and reasoning
Structured interviews	+ Can be efficient and productive – Can be time consuming +/– Can yield some information about domain concepts and reasoning
Contrived tasks	+ Can inform about refined reasoning + Can be tailored to probe particular or subdomain knowledge or particular reasoning strategies – Expert can initially feel uncomfortable

Consensus from the Case Studies

Among others, Dhaliwal and Benbasat (1990) made a plea for controlled systematic research on KE and KA, research that is sensitive to issues of operationalization, individual differences (e.g., interviewer style and skill), and the expert's opinions about expert systems. Furthermore, they argued for systematic treatment of fundamental methodological problems, such as, How can one discriminate the contribution to an expert system made by an expert from that made by a KA tool?

Despite the lack of adequate empirical analysis of the problems of KE and KA with the expert systems field, some general consensus can be gleaned from the case experiences, which we summarize in Table 4.[8]

Recently, some experimental comparisons of KE methods have been conducted.

[8] We will not discuss the relative merits of KE versus automated KA. Reference in Table 4 is to human-oriented KE and only indirectly to automated KA systems—insofar as such systems rely upon the KE methods (such as repertory grids, etc.). Hence, advantages and disadvantages listed in Table 4 do not refer, for example, to interface quality issues.)

EXPERIMENTAL COMPARISONS OF KE METHODS

Experimental comparisons have focused on two issues: the relative efficiency of techniques and the complex interactions of KE technique with knowledge types, reasoning strategies, and domain characteristics.

The Relative Efficiency of KE Methods

Hoffman (1987) compared five methods: a documentation-based task analysis, unstructured interviews, structured interviews reliant on a first-pass knowledge base, familiar tasks with think aloud processing constraints and tough case materials, and problems that combined limited information with processing constraints. The application domains were aerial photo interpretation and military airlift scheduling. Hoffman began by generating operational definitions of variables that could capture potential advantages and disadvantages of the various methods. Table 5 presents the dimensions used in that study.

Focusing on the overall efficiency of KE methods, Hoffman asked how many "informative propositions" each KE method produced per "total task minute" (TTM). Informative propositions were defined as those which were not in the initial documentation-based first-pass knowledge base. TTM referred to the total amount of time it took the elicitor to prepare for the session, plus the time of the session, plus the time taken to analyze the transcripts for propositional content. In other words, the measure reflected the total effort on the part of the elicitor. The efficiency ratio expressed the average rate at which informative propositions were elicited.[9]

TABLE 5
Some Operational Dimensions on Which KE Methods Can Be Compared

Dimension	Operational definition
Simplicity of materials	The number of stimuli or other materials and their complexity relative to the familiar task
Simplicity of the task	Brevity of the instructions that are necessary to specify precisely what the expert is expected to do
Brevity of the task	Total time on task, or total time relative to the duration of the familiar task.
Flexibility of the task	Is it adaptable to different materials, different experts, variations in instructions, etc.?
Artificiality of the task	How much, and in what ways, does it depart from the familiar tasks?
Data format	Do the data records come out of the task in a format ready to be represented in a computer?
Data validity	Do the data records provide correct and important knowledge?
Method efficiency	How many informative propositions are produced per total task minute?

Hoffman's study showed that the unstructured interview produced less than one informative proposition per TTM. The structured interview, on the other hand, yielded about one informative proposition per TTM. The contrived task and the analysis of tough cases were the most efficient, yielding between two and three informative propositions per TTM.

Hoffman recommended that throughout all KE projects (on diverse domains, from any paradigm or perspective), records should be kept of effort. And if one finds that a KE technique is yielding useful information at a rate of only one informative proposition per TTM, one might consider switching methods. On the other hand, if a KE method is yielding information at a rate closer to two or three informative propositions per TTM, then one can have some confidence that one is proceeding effectively.

The conclusion from Hoffman's research is that KE methods can differ in the relative efficiency with which they yield knowledge. This leads to research that bears on a second issue in the evaluation of KE methods.

The Differential Access Hypothesis

This issue reflects a belief by a number of philosophers, psychologists, and system developers that different KE techniques may elicit different types of knowledge (e.g., declarative versus procedural, explicit versus tacit, verbal versus perceptual, etc.) and different kinds of strategies (top-down versus bottom-up reasoning, convergent versus divergent thinking, etc.) (Berry & Broadbent, 1984, 1988; Dhaliwal & Benbasat, 1990; Dreyfus & Dreyfus, 1986; Gordon, 1992; Jackson, Reichgelt, & Harmelen, 1989; Olson & Reuter, 1987). Gammack and Young (1985) asserted that domain concepts could be best elicited by documentation analysis, domain concept interactions could be best elicited by sorting and scaling tasks, and that procedural rules and heuristics could best be elicited by think aloud problem solving, task analysis, and interviews based on memory probe questions (see also Geiwitz, Klatzky, & McCloskey, 1988).

Differential access could arise because of strategy effects or because of limitations on memory access.

[9] An "informative" proposition would not necessarily be "important" when it comes to building an expert system. For instance, a system developer would be delighted if a single general or more powerful rule could effectively substitute for a number of more specific or less powerful rules. Hence, "rules per pound" of KE may be appropriate for exploratory experimental comparison of KE methods, but only approximates what knowledge engineers need to know.

Differential access due to strategy effects. Different KE techniques (the procedures or the materials) might actually mold cognitive processes in such a way that knowledge may interact with reasoning strategies and goals. This was shown in the study by Hammond *et al.* (1987), in which highway engineers performed a road rating task given different formats and types of information about the roads. It was also suggested by the study by Phelps and Shanteau (1978) in which expert livestock judges rated breeding quality based on either photographs of gilts versus descriptions of gilts in terms of the key variables.

Prietula *et al.* (1989) demonstrated a strategy effect in another way, by varying feedback. Experts at controlling steam plants for industrial electrical power generation interacted with a graphical depiction of a plant (showing boilers, pumps, pipes, etc.) by manipulating such variables as flow rates, fuel consumption, the amount of steam produced, and the distribution of power. The goal set for them was to minimize operating costs. In one condition, each parameter change the expert made was followed by feedback concerning the effect on operating costs. In another condition no feedback was provided. This relatively simple manipulation had a considerable effect. Providing feedback led the experts to adopt an "hypothesis-and-test" strategy whereby they attempted to balance a number of tradeoffs in plant operation. Lack of feedback led the experts to adopt a "plan-and-implement" strategy whereby they first analyzed each component device and alternative device settings and then implemented what was believed to be an optimal set of parameters based on heuristic rules and the pertinent equations. Their analysis of the system was conceptually deeper and richer even though they were less able to achieve the minimum-cost goal.

These findings show that a simple manipulation of the KE procedure can have an effect on strategy, which in turn can effect the kinds and extent of the knowledge elicited. The cognition of experts is very flexible and both goal- and context-dependent (Spiro, Vispoel, Schmitz, Samarapungavan, & Boerger, 1987).

Differential access due to nonverbalizability. Another possible cause of differential access hinges on the notion that some aspects of expertise might not be verbalizable because they involve knowledge that is not available to consciousness (Agnew, Brown, & Lynch, 1986; Bainbridge, 1979; Kim & Courtney, 1988; Olson & Reuter, 1987; Salter, 1988). Hypothetically, knowledge that may initially be explicit becomes automatic with practice (Woodworth, 1938, Ch. 29). In AI terminology, declarative knowledge becomes compiled down or proceduralized and loses the form in which it can be accessed by consciousness (Anderson, 1983, 1987). This is easily recognizable in such common skills as automobile driving, but is also believed to characterize the shift from novice to expert (Dreyfus & Dreyfus, 1986; Glaser, 1987; Klein & Hoffman, 1993).

Social psychologists have argued that it is possible to act purposefully with little awareness of the reasons for one's actions (Nisbett & Wilson, 1977). Cognitive research has shown that some knowledge can be more readily characterized as declarative and some more readily as procedural (Anderson, 1987; Gordon, 1992). Similarly, it has been demonstrated that in the learning of process control procedures, improvements in performance can be partly dissociated from the ability to verbalize knowledge about how the process is controlled. That is, successful performance can go along with an inability to verbalize procedures, and unsuccessful performance can go along with an ability to verbalize procedures (Berry & Broadbent, 1984, 1988; Buchner, Funke, & Berry, 1995). In light of such research, it has been argued that contrived KE techniques are necessary in order to reveal experts' tacit or unconscious knowledge and strategies (i.e., think aloud problem solving cannot reveal implicit knowledge, conceptual graphs can elicit "deep" knowledge, etc.) (e.g., Kim & Courtney, 1988; Klein & Weitzenfield, 1982; McGraw & Riner, 1987; Olson & Reuter, 1987; Salter, 1988).

Recent research on expertise has addressed the question of verbalizibility-based differential access. The first goal of such research has been to show whether differential access actually occurs in the KE context.

Does differential access actually occur in KE? Crandall and Klein (1987; Crandall, 1989) interviewed experienced fire fighters about their interpretations of what was happening during urban fire scenarios. Two KE methods were unstructured interviews (i.e., "What are you noticing here?") and the Critical Decision Method. The latter yielded much more information, including a wider variety of specific details and more information about underlying causal linkages among the core concepts. Indeed, it facilitated the revelation of knowledge that, one might suppose, was tacit. For example, in describing one of his experiences, a firefighter initially explained that he had a "sixth sense" for judging the safety of a fireground (i.e., a burning roof). Upon re-analysis of the recalled event using the probe questions, the expert "discovered" the perceptual cues that he relied upon, such things as smoke color and the feel of a "spongy" roof.

Cooke and MacDonald (1986, 1987) recognized some of the problems of protocol analysis and unstructured interviews and so compared a number of alternative KE methods. In one study they elicited knowledge about automobile driving from a number of experi-

enced drivers using one of four techniques: (1) Small group interviews involving an "instructor," a "student," and a person whose job it was to extract concepts from the instructor-student dialogue; (2) Concept listing in which participants had to "list all the elements pertinent to driving a car"; (3) Step listing, in which participants had to "list all the steps involved in driving a car"; and (4) Chapter listing, in which participants had to "Pretend you are writing a book and need to first list all the chapters." Cooke and McDonald classified the knowledge elicited by the four methods (e.g., general rules, concepts, procedures, facts, etc.). The results suggested that the small group interviews and the chapter listing procedure generated mostly concepts (e.g., "skidding," "brakes"), whereas step listing and concept listing elicited more rules or procedures (e.g., "wear a seat belt").

Some recent studies do not show differential access. For example, Adelman (1989) conducted an experiment involving multiple experts, two KE methods, and multiple knowledge elicitors. Participants were graduates of Marine Corps training in combat readiness evaluation (i.e., they were advanced apprentices). The knowledge elicitors were six knowledge engineers who had already utilized a variety of KE methods in their private-sector work in other domains. The two KE methods were attribute listing and graph construction, both of which focused on the factors and attributes of combat preparation (i.e., tasks, requirements, mission standards, etc.). The two methods were used to structure interviews (3–4 h long) with small groups of participants. Data were compared to the Marine Corps' formal performance standard, itself an hierarchical mission planning guide—which had *not* been taught explicitly in the courses that these participants had taken. Results showed no difference in knowledge elicited by the different elicitors or by the two KE methods. It should be noted, however, that the two KE methods were both types of structured interview, both involved groups, and both involved generating data in the form of a multiattribute hierarchy.

A major attempt to empirically compare KE methods was conducted at the University of Nottingham (Burton, Shadbolt, Hedgecock, & Rugg, 1987; Burton, Shadbolt, Rugg, & Hedgecock, 1988, Burton *et al.,* 1990; Schweickert *et al.,* 1987; Shadbolt & Burton, 1989, 1990b). The first study (Burton *et al.,* 1987) involved 32 advanced students of geology who were skilled in the classification of igneous rocks. Burton *et al.* utilized four KE methods: a structured interview, think aloud problem solving, a laddered grid, and a card sort. The first two methods were considered to be Familiar or natural to the experts, and the second two were considered to be Contrived. In the experiment, each participant served in two sessions, one with a Familiar and one with a Contrived method. Dependent measures included the time taken in the sessions, the time taken to transcribe sessions into usable rules, the number of rules elicited, and the complexity of rules. A senior geologist provided a "gold standard" rule set against which the completeness of sessions could be assessed. It was hypothesized that the interview and think aloud problem solving would produce more procedural knowledge, whereas the Contrived techniques would elicit declarative knowledge.

The differential access hypothesis was not found to be predictive in this domain, that is, there was considerable overlap of knowledge elicited by each of the techniques. Although each participant used both types of technique, almost all the same knowledge was elicited, i.e., there was no interaction of technique and knowledge type. In a second experiment, Burton *et al.* (1988) studied experts' identification of the geographical features associated with glaciation. The design of the experiment was the same as the first study, though this time 32 highly experienced geographers served as participants. The pattern of results in this study closely resembled that of the first. Think aloud problem solving (including the protocol analysis) was once again the least efficient technique. Once again, the Contrived tasks yielded data similar to that from the interviews, that is, there was no pronounced differential access.

It is possible that the failure to find an interaction of KE methods and knowledge types was due to the nature of the domains studied. The domain task is classificational in nature and also fairly simple. It may turn out that domains involving familiar tasks with other characteristics might show a pattern of differential efficacy across KE techniques. Shadblot and Burton (1990b) and Burton *et al.* (1990) provided some support for the differential access hypothesis by looking at domains in which the classification tasks are of substantial complexity. Eight experts (i.e., highly published academic and museum professionals) were recruited from each of two archeological domains: identification of Stone Age flint artifacts and classification of Medieval pottery shards. A great amount and variety of knowledge is required for the solution of classification problems in these domains. As in the other experiments, a variety of KE methods was used. The results confirmed the earlier conclusion that protocol analysis was less effective (i.e., more time-consuming) than the Contrived techniques. However, the Contrived techniques needed to be used in conjunction with an interview since they elicited specific knowledge and did not yield an overview of the domain knowledge.

General Assessment of the Differential Access Hypothesis

The studies of Burton and his colleagues showed that protocol analysis was the more time-consuming tech-

nique in terms of data analysis. Furthermore, this technique yielded less information and less complete information than any other. The Contrived tasks took less time than the interview but yielded about the same amount of information. This fits with Hoffman's (1987) findings, and the general conclusion seems inescapable–think aloud problem solving (with protocol analysis) can be inefficient in KE. Yet, this technique is widely used by expert system developers (Cullen & Bryman, 1988). We suspect that the reason for its widespread use is that most experts usually feel fairly comfortable with the think aloud procedure.

A second conclusion from the experiments is that differential access is not pervasive. Some psychological research does show that consistency can obtain between verbal reports and task performance, in a way suggesting that the accuracy of verbal reports depends not only on the procedures used to elicit verbal reports but also on the procedures used to assure that the knowledge is potentially accessible in the first place (i.e., the use of within-subjects vs between-subjects designs) (Ericsson & Simon, 1993).

The strong version of the hypothesis might simply be incorrect. One weaker version states that *Although KE methods may all have the potential to elicit many knowledge and strategy types, methods may differ in terms of the kinds of knowledge or strategies that they elicit most effectively or most readily.* This version dovetails with the finding that contrived KE techniques are sometimes useful in the elicitation of refined reasoning or subdomain knowledge (Hoffman, 1987) and correspondingly can be less useful in eliciting an overview of a domain (Burton *et al.,* 1990).

The weaker version of the differential access hypothesis dovetails with some of the conclusions from available case study reports on expert systems work. For instance, Kim and Courtney (1988) analyzed KE according to a scheme that distinguished "knowledge engineer-driven" methods (interviews, repertory grids, think aloud problem solving) with "expert-driven" methods (in which the expert interacts with a toolkit) and "machine-driven" methods (automated learning by induction from examples). Kim and Courtney concluded that all three methods or approaches can yield information about practices and heuristics as well as domain concepts. However, they argued that conceptual graphing is better at eliciting "deep" knowledge, that ratings tasks are better at eliciting information about domain concepts than about domain procedures, and that think aloud problem solving yields information about domain heuristics more readily than about domain concepts (see their Table 4).

An alternative explanation for the partial overlap of data content between KE techniques hinges on a notion of domain-dependence rather than any form of differential access. In some domains, expertise depends heavily on particular reasoning strategies—any dependence on particular knowledge types may be incidental to this. In chess, all the data are available at a glance whereas in clinical medicine, data are uncovered over time. In chess, actions are linked heavily to pattern recognition, whereas in clinical medicine actions are also linked to complex hierarchical knowledge structures (Prietula *et al.,* 1989). In the classification domain/tasks studied by Burton *et al.,* (1988) there may be little use for procedural knowledge.

If this is the case, tasks within domains would need to be classified by strategy type (rather than knowledge type) in order to permit recommendations about which KE technique might prove most efficient. It is also possible that choice of a KE method could depend on the extent to which a domain is "well structured" (i.e., it involves standardized or normative procedures, well-defined concepts, etc.) versus being "ill structured" (i.e., it involves many interrelated and causally connected concepts, uncertainty in decision-making, ambiguous or missing data, numerous decision paths, conflicting goals, etc.) (Kim & Courtney, 1988).

An acid test of the strong differential access hypothesis would involve attempting KE with individuals who are naturally inarticulate and whose skills (apparently) rely on unconscious processes or tacit knowledge. Neisser (1983) performed such a test by studying an expert (savant) mental calendar calculator. By extensive use of test case problems (to generate reaction time data) and the use of frequent probe questions, Neisser was able to reveal the savant's extensive knowledge base and set of flexible strategies.

In general, research has failed to support a strong version of the differential access hypothesis. Indeed, it has never actually been demonstrated that experts actually possess knowledge that is *in principle* nonverbalizable. The burden of proof falls on the shoulders of those who claim that knowledge exists in different "kinds" relative to its verbalizability. In the context of KE, it is always "premature to conclude that knowledge about a given term or topic is tacit and completely unavailable to the expert when he or she fails to respond in a single questioning context" (Wood & Ford, 1993, p. 80). Furthermore, the chance that knowledge can be more or less verbalizable has not seemed to be much of a problem in practice.

Our discussion of empirical comparisons has focused so far on the two questions of efficiency and differential access, but additional comparisons are noteworthy.

The Interaction of Personality Variables and KE Methods

In their experiments on KE, Burton *et al.* took individual differences into account. In the 1987 study, they found that there was a significant correlation between

participants' performance in the interview and their scores on a scale of introversion–extraversion. As one might expect, extroverts provided information more quickly and easily than introverts. However, introverts ended up generating as much if not more knowledge than extroverts. The dimension was not correlated with performance in any of the other KE conditions.

A measure of cognitive style was also taken, that of "field dependence," measured by the Embedded Figures Test (Witkin, Oltman, Raskin, & Karp, 1971). It was found that field-dependent participants performed significantly worse than field-independent participants on the laddered grid technique. Recall that this is a graphical technique which requires participants to study complex two-dimensional information, and so the effect is in a predictable direction.

These findings raise the interesting and potentially important possibility that particular KE techniques may suit participants with differing psychometric profiles (i.e., verbal expressiveness), a conclusion supported by research on think aloud problem solving and individual differences in verbal expressiveness (Ericsson & Simon, 1993).

We are now in a position to present some recommendations for KE and then conclude with a discussion of challenges and prospects.

"HOW TO DO" KE

Bootstrapping

Whether for the purpose of preserving corporate knowledge, building an expert system, or studying cognition, the researcher has to be bootstrapped into the domain prior to KE (Clancey, 1988; Newell, 1981; Wood & Ford, 1993). How far one should progress along the developmental continuum is an open question. Neale (1988) argues that the researcher must develop a deep conceptual model of the domain, but informal consensus seems to be that the researcher should at least enter the apprentice stage.

Useful methods for bootstrapping are documentation analysis and unstructured interviews. This seems to be an obvious or trivial point until one encounters a domain where there is only one expert, or a domain in which there is no documentation.

Stages for KE

One should not expect unstructured interviews to work efficiently in the long haul, that is, going from a first-pass knowledge base to a richer or implementable knowledge base. Rather, one should consider such methods as task analysis, structured interviewing, and contrived techniques in order to flesh out a first-pass knowledge base. A number of psychologists and system developers have proposed what they regard as generally useful stages for KE (Brule & Blount, 1989; Cochran *et al.*, 1989; Diaper, 1989; Hoffman, 1997; Mullin, 1989; Olson & Reuter, 1987; Salter, 1988; Wood & Ford, 1993). The common theme to the proposals is that knowledge that is acquired first is used to constrain subsequent KE; early elicitation of domain concepts is used to guide the elicitation of knowledge about the interactions of domain concepts and knowledge about domain tasks. The common theme to proposals in expressed in Table 6.

After Stage 4, KE is largely accomplished, but the work goes on, of course, to generate training manuals, build a system prototype, generate cognitive experiments that test alternative hypotheses, etc. (Cohen & Howe, 1989; Neale, 1988; Rook & Croghan, 1989; Weitzel & Kerschberg, 1989).

Along with others, we recommend that one should avoid reliance on a single KE method (Alluisi, 1967; Gordon, 1992; Mullin, 1989; Salter, 1988; Wood & Ford, 1993; Wright & Ayton, 1987). The weak version of the differential access hypothesis is a cautionary tale, reminding us that any KE session might provide partial information.

Should One Use Familiar Tasks or Contrived Techniques?

The materials and procedures used in KE can interact negatively with those used in the exercise of knowledge (the expert's familiar task) (Breuker & Wielinga, 1984, 1987; Fischhoff, 1989; Shanteau, 1992). Any KE project could yield flawed results if the KE method misled participants, forced them into conforming their knowledge and reasoning strategies to unfamiliar formats, encouraged them to form inadequate problem

TABLE 6
Some Consensus on Stages for KE

Stage	Methods	Purposes
1	Documentation analysis, unstructured interviewing, or observation/analysis of the expert's familiar tasks.	To bootstrap the researcher.
2	Same as Stage 1.	To generate a first-pass knowledge base.
3	Structured interviewing, think aloud problem-solving or other contrived tasks.	To validate, refine, or extend the knowledge base.
4	Same as Stage 3.	To instantiate the refined knowledge base in documents or implementable systems.

representations, or forced them into response modes that reflect poorly on their problem-solving abilities, say, by artificially changing the workload (Cleaves, 1987; Fischhoff, 1989; Hoch & Loewenstein, 1989; Johnson & Thompson, 1981; Norman, Rosenthal, Brooks, Allen & Muzzin, 1989; Salter, 1988). Hence, Salter (1989) uses "intrusiveness" as the main dimension to characterize KE methods that are intended to study experts' normal performance in their familiar tasks.

In selecting a palette of KE methods one must make some attempt to sample the reasoning strategies, problem types, goals, etc. that are involved in the domain's familiar tasks (Chiles, 1967). Woods (1993) argues that any context and method of "process tracing" (the description of on-going reasoning and task behavior) should be representative of the context in which knowledge and skill are exercised.

On the other hand, some evidence suggests that contrived techniques can be effective in eliciting subdomain knowledge or aspects of refined reasoning (e.g., reasoning about how to develop new methods to how to handle novel or rare cases) (Hoffman, 1987; Klein *et al.*, 1989; Mullin, 1989; Olson & Reuter, 1987). Thus, there can be a positive side to the interaction of the KE method and familiar task.

Hoffman (1987) found that experts can initially be "put off" by contrived problem situations. It is important for experts to feel that KE is not evaluative or a challenge to authority. A striking example of this occurred in KE with expert aerial photo interpreters. In their familiar task they always rely on available ancillary information, such as maps. The notion of interpreting photos without such contextual information was anathema. But they came to see the limited information problem as a useful and even interesting exercise. At that point, the method began to yield a wealth of data about knowledge and reasoning.

Based on their own comparative analysis of KE techniques, Burton *et al.* (1987, 1988, 1990) reached a similar conclusion: The a priori impression a person (expert, system developer, etc.) has about the meaningfulness or potential worth of a session may not reflect the information eventually obtained. In particular, techniques that force the expert's knowledge into an unfamiliar format can be more useful than experts (or researchers) might initially suppose.

Furthermore, whether a given KE method seems contrived can depend on the perspective (researcher versus domain expert). In some domains, the expert's familiar task itself may involve completing sets of ratings or rankings. Some experts may routinely perform tasks in which they verbalize some of their reasoning or behavioral strategies (e.g., medical diagnosis), but if they do not, verbalization might influence their reasoning processes (positively or negatively). And so on. The issue is not a choice between the familiar and the contrived, but a choice of an appropriate mix.

To this point, our discussion has centered on illustrating KE methods and their combinatorics and on generating some recommendations based on what is currently known. What are some of the outstanding challenges and issues? What are some prospects for further research? We discuss two lingering issues, one having to do with the possibility of bias in KE, and the other centering on the problem of taskonomy.

BIAS IN KE?

Bias in KE may come from the expert, the researcher, or the method. For example, the use of probe questions always entails a potential danger that the questions might "lead" the interviewee. Loftus (1975) has shown strong effects of question formulation in event recall. Following a slide presentation depicting an automobile accident, participants were asked to recall the speed of the collision, and their responses showed systematic variation depending on which intensifier was used to ask: "How fast was the car going when it (crashed) (smashed) (collided) (ran into) the wall?" Throughout all KE there is the possibility of "reductive bias" when the researcher/apprentice misconceives or overly simplifies the domain, creates artificial distinctions, or misinterprets domain concept terms that happen to also be everyday terms (Feltovich, Spiro, & Coulson, 1989; Wood & Ford, 1993).

Reductive bias can accompany the use of certain measures of expert performance. Studies of "hit rate" and "skill score" have sometimes shown that fairly simple linear regression models can outperform the expert (Dawes, 1979; Meehl, 1959), even when experts insist that the problems are complex and configural. This finding generally obtains for domains in which the expert's task is the prediction of human behavior under dynamic conditions involving a lack of feedback and a lack of standardized procedures or decision aids (Mullin, 1989; Shanteau, 1988, 1992).

Whether statistical prediction outperforms the expert can depend critically on the job task and the amount and kind of contextual information that is available to the human (Meehl, 1954; Yaniv & Hogarth, 1993). This serves as an important reminder that single, simple outcome measures of performance do not come close to revealing the depth and detail of experts' knowledge or reasoning, however fallible it may be.

Bias can also come from the expert. Psychological research on reasoning bias can be divided into two paradigms, that focusing on biases in logical reasoning and that focusing on bias in probabilistic reasoning.

Bias in Probabilistic Reasoning

This has been especially well documented (Kahneman, Slovic, and Tversky, 1982; Kahneman & Tversky, 1982; Lichtenstein & Fischhoff, 1980; Tversky & Kahneman, 1983), and types of bias have been cataloged (Fraser, Smith, & Smith, 1992). Examples are: (a) to be unduly swayed by the cognitive availability of information, to mistake this characteristic for frequency; (b) to anchor judgments on initial estimates; (c) to assess the likelihood of an event based on familiarity or stereotypy rather than objective frequency; and (d) to overestimate the frequency of rare events.

Following the demonstrations of Tversky *et al.*, some researchers speculated that various biases might be manifest in experts (Cleaves, 1987; Evans, 1988; Fischhoff, 1989; Jacob, Gaultney, & Salvendy, 1986). At first glance, the fallibility of people (in general) seemed so severe that pundits wondered whether information systems should be designed to mimic human expertise at all (Dreyfus & Dreyfus, 1986; Tolcott *et al.*, 1989), while some insiders suggested that knowledge engineers should avoid the use of probabilistic or statistical judgments in KE altogether (Hink & Woods, 1987).

The work on probabilistic reasoning bias has been a red flag because the notion of uncertainty is crucial in many expert systems (cf. Deane & Kanazawa, 1989; Doyle, 1983; Fox, 1986; Hall & Kandel, 1988; Kuipers, Moskowitz, & Kassirer, 1988; Mullin, 1989; Neapolitan, 1990; Zadeh & Kacprzuk, 1992). For example, in diagnostic domains one may need to formulate such rules as: "If the patient has spots, then the patient has measles with certainty *X*." If experts provide biased probability estimates, there could be considerable problems for those building expert systems containing rules that are triggered when certain probability values are in effect for certain variables.

In many applications, statistical judgment and the sorts of judgments involved in decision analysis are contrived in that they can take experts away from their usual way of thinking about problems. As Fischhoff (1989) cautioned:

> . . . the elicitation process must allow respondents to admit ignorance and encourage them to assess their full knowledge, rather than take the first number that comes to mind. . . . [yet] the need for numbers often forces analysts to extract judgments from experts that strain their capability and credibility. (pp. 454, 457)

Yet, Fischhoff argued that people (in general) have little trouble in giving probabilities, and that decision analysis can be used in KE—and indeed should be used whenever the focus is on improving judgment by making decision processes and judgment criteria explicit.

There are, of course, domains in which expert reasoning in familiar tasks is explicitly reliant on probabilistic or statistical judgment and expressions of uncertainty. This is perhaps best exemplified in some weather forecasts (e.g., rain is "likely"; skies will be "partly" cloudy). ("Best" because short-term public weather forecasts produced by the Weather Service are actually quite accurate.) In such domains, it is possible or at least more likely that expert could comfortably and easily express some of their reasoning in terms of quantitative judgments of uncertainty (Laskey, Cohen, & Martin, 1989). For such domains, one can question whether errorful data necessarily represent "errors" or "biases," or actually represent the exercise of strategies or procedures that are appropriate for certain situations (Fraser *et al.*, 1992).

Furthermore, in experiments in which error-prone expert reasoning has been induced, experts seem more likely than novices to achieve a correct solution once the inadequacies of their initial problem representations have been pointed out or discovered (Johnson & Thompson, 1981). Some researchers have expressed doubt that the biases in probabilistic reasoning that have been observed in laboratory research (i.e., college-age participants in contrived problems using limited information and artificial materials) occur with the same frequency and magnitude in any real-world problem-solving situations (Beyth-Marom & Arkes, 1983; Carroll & Siegler, 1977; Christensen-Szalanski & Beach, 1983; Fraser, Smith, & Smith, 1992; Olson, 1976; Tolcott *et al.*, 1989; Wright, 1984; Zakay & Wooler, 1984).

Bias in Logical Reasoning

Among the types of bias in logical reasoning that have been observed in the laboratory and in the history of science are: (a) the tendency to assign undue weight to the first evidence obtained, (b) overreliance on variables that have taken on extreme values, (c) the tendency to seek evidence that confirms the current hypothesis, (d) the tendency to reason about only one or two hypotheses at a time, (e) the tendency to be overconfident, (f) the desire to maintain consistency even if that means devaluing or ignoring important information, (g) belief in illusory correlations, (h) being overly conservative, and (i) basing conclusions on hindsight (i.e., "I knew it all along.") (Edwards, 1968; Evans, 1989; Fischhoff, 1989; Fraser *et al.*, 1992; Johnson-Laird, 1983; Tolcott *et al.*, 1989; Tweney, Doherty, & Mynatt, 1991).

Evidence concerning the extent of cognitive bias in general or everyday reasoning is actually mixed. In some cases, certain types of cognitive bias occur and in other cases they do not, or they occur only to a moderate extent. Often when bias does occur in laboratory

research it is because the participants have been forced to

> . . . make an unreasonable interpretation of the problem as stated by the experimenter, but their interpretation is reasonable in a more realistic version of the problem . . . the experimenter has created a situation in which the subject's natural tendency to seek explanation of seemingly random variation is incorrect; however, such a tendency is appropriate in many real situations. (Fraser, Smith, & Smith, 1992, pp. 299–300)

In their studies of medical decision-making, Schwartz and Griffin (1986) cited over 20 relevant papers supposedly demonstrating that experts rely on heuristics or "rules of thumb." However, Schwartz and Griffin argued that experts do not seem to be prone to biases to such an extent as to have practical import in KE. Similarly, in studies of auditing there are reports of both biased (e.g., Fischhoff, 1989; Holt, 1987) and unbiased (e.g., Kinney & Uecker, 1982) expert reasoning, yet in a majority of the studies bias effects are much smaller than those of the novices (Ashton, 1983; Olson, 1976; Shields, Solomon & Waller, 1987).

Despite the mixed results, the list of reasoning biases is so formidable that one must wonder whether unbiased KE is possible. Assuming bias is rampant, Cleaves (1987) offered a few suggestions, e.g., the biased anchoring of judgments can be counterbalanced by asking experts about extreme cases before asking about prototypical cases. Tolcott, Marvin, and Lehner (1989) suggested that just prior to KE, the experts should be explicitly informed about the biases that may operate. Such suggestions involve tackling the biases one at a time; neither approach deals directly with the practicalities of KE or the tradeoffs and interactions that can be involved.

For example, certain tasks may be ideal for inducing behavior that could be attributed to certain cognitive biases. Tolcott *et al.* (1989) tried to see if cognitive biases would occur in the reasoning of expert analysts of battlefield intelligence. In the first KE session, the experts were engaged in think aloud problem solving of battlefield scenarios that were described with limited information. In the subsequent sessions, the experts engaged in decision analysis based on additional information about battle outcome. This is a problem situation which could encourage a nonanalytical process of initial hypothesis formation (with the potential for biased deemphasis of base rate information) followed by attempts to confirm the hypothesis (with the potential for biased deemphasis of contradictory information). That is exactly how the experts performed, with the codicil that the experts were not prone to overconfidence or other biases. In this case, the experts demonstrated performance similar to (i.e., "biased" like) that of novices in contrived and less dynamic laboratory tasks.

The Challenge

We currently have little evidence about the extent of cognitive biases in expert reasoning across diverse domains and tasks (Fischhoff, 1989; Levi, 1989; Shanteau, 1992), especially perhaps in so-called ill-structured domains (Spiro *et al.*, 1987). Despite the suggestion that bias is not much of a problem for KE, caution is still in order since reasoning biases can conceivably make KE unreliable. Further research on KE should be aimed at enabling researchers to identify and deliberately utilize bias effects (Fischhoff, 1989; Gaeth & Shanteau, 1984; Meyer, Booker, & Bradshaw, 1990).

This call for further research points to some additional empirical challenges.

EMPIRICAL CHALLENGES FOR PSYCHOLOGY AND AI

In the area of expert systems there exist literally thousands of reports on projects involving diverse KE methods, multiple projects on the same or highly related domains, and so on. Most KE procedures go undocumented, and published reports do not include an adequate discussion of KE methods. With others, we feel that this practice should change for the long-term benefit of those who seek answers to what are, essentially, empirical questions (Cohen & Howe, 1989; Neale, 1988). For instance, Adelman (1989) suggests that

> By using two or three knowledge engineers, knowledge representation schemes, and elicitation methods when working with two or more domain experts [one should] be able to identify which, if any, of these sources of variability result in disagreement . . . Although the sample size will still be small, it is substantially better than N = 1 . . . We would urge developers to document disagreements and resolution procedures in order to accumulate experientially-based knowledge. (p. 488)

With regard to conceptual or theoretical analysis within AI, the task(s) that experts perform have been categorized by a number of system developers (e.g., Chandrasekaran, 1983, 1986; Duda & Shortliffe, 1983; Madni, 1988; Stefik, Aikins, Balzer, Benoit, Hayes-Roth, Waterman, & Lenat, 1982). The conception of computer systems as falling into basic categories, such as planning systems, diagnosis systems, control systems, etc., is reflected by a conception that domains are composed of tasks that fall into generic categories (cf. Bylander & Chandrasekaran, 1987) or decisions that fall into basic types (operating, coordinating, strategic) (Kim & Courtney, 1988). It is further believed that the categorization of a domain or task according to generic types will define or constrain the KE methods that are appropriate (e.g., Dhaliwal & Benbasat, 1990; Kim & Courtney, 1988).

The generic task scheme has been applied in the

analysis of expert systems (Chandrasekaran, 1986, 1990), but has yet to be applied in the empirical analysis of diverse domains (Johnson *et al.,* 1993). At a cognitive level, generic tasks each involve a variety of specific perceptual, conceptual, and reasoning skills (Breuker and Wielinga, 1987; Clancey, 1985). As the research of Burton *et al.* (1988) suggests, familiar tasks need to be classified according to both the reasoning strategies and the types of knowledge involved, but the generic task scheme has not yet been interfaced with psychological characterizations of tasks in terms of abilities and processes such as recognition, discrimination, categorization, vigilance, workload, etc. (as taxonomized in Alluisi, 1967; Fleishman, 1967, 1975; Gagne, 1964).

In addition to the attempt to map KE methods onto domain/task types, some computer scientists have become sensitized to the need to facilitate communication and social interaction in KE (cf. Basden, 1989; Brown, 1989; Clancey, 1993; Forsythe & Buchanan, 1989; Hart, 1986; Hayes, Ford, & Agnew, 1994; McGraw & Seale, 1988; Meyer & Payton, 1992; Monk, 1985). They have become sensitized to the individual differences—characterizing both the expert and the knowledge elicitor—that make for effective communication (cf. Fellers, 1987; Forsyth & Buchanan, 1989; Neale, 1988; Olson & Reuter, 1987; Prerau, 1989; Rolandi, 1986). In this article we have referred to "elicitation" as if knowledge were extracted like gold from ore. But in many contexts, KE is better described as knowledge "co-creation" (Agnew *et al.,* 1986; Clancey, 1993; Ford & Bradshaw, 1993; Hayes *et al.,* 1994; LaFrance, 1992; Meyer & Payton, 1992; Neale, 1988; Regoczei & Hirst, 1992).

The available psychological research comparing KE methods is a first step in establishing a methodology that will permit a fuller determination of how to effectively elicit expert knowledge (Dahlstrom, 1989; Fellers, 1987; Honeck & Temple, 1992; Kim & Courtney, 1988; Rook & Croghan, 1989; Shanteau, 1988, 1992; Wiggs & Perez, 1988). A goal of researchers—both system developers and psychologists—is to generate empirically sound advice about which KE methods work best for which domains, and advice about which methods to use at various stages of a project (Kim & Courtney, 1988). As Fleishman (1975) anticipated: "We need several task classification systems for several purposes, with the linkage between them understood and specified" (p. 1147). One must be able to map individual differences, problem-solving strategies, domain types, and types of familiar tasks onto KE techniques and show how to put KE techniques together to form coherent series of elicitation activities. Regarding the major sources of variability, ". . . major research programs varying [them] all are required in order for us to better understand the extent to which they affect knowledge base quality" (Adelman, 1989, p. 487).

Controlled empirical studies should continue to offer insights into expertise and practical applications in KE. This entails continuing interdisciplinary collaboration. The technology of knowledge-based systems continues to evolve and be applied in diverse domains. In many ways, this is driving cognitive science and its applications. For instance, AI researchers have realized that expert systems need to have a number of "second-generation" capabilities if they are to be accepted. For one, expert systems should be able to explain their reasoning (Jackson, 1990; Neches, Swartout, & Moore, 1984; Shortliffe, 1976; Swartout, 1983). This dovetails with cognitive research on the question of what makes for a "good" explanation (Buchanan & Shortliffe, 1984; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Clancey, 1983; Wick & Slagle, 1989).

In some applications, expert systems could benefit from the capability to form "user models" (Kidd, 1985) and reason at a conceptual level (Alexander, Freiling, Shulman, Rehfuss, & Messick, 1987; Bylander & Chandrasekaran, 1987; DeGreef, Breuker, & Wielinga, 1986; Reddy, 1988; Shadbolt & Wielinga, 1990; Steels, 1990; Woods & Hollnagel, 1987). This dovetails with cognitive research on mental models, or the general idea that knowledge must be accessible at a "higher" or more conceptual level than the languages traditionally used to program expert systems (Neale, 1988; Salter, 1988).

Achieving such goals will require further psychological research, continuing effort on behalf of AI researchers to solve the riddles of symbolic computation, continuing effort on the part of applied psychologists and expert system developers to explore and document diverse domains, and continuing cross-disciplinary collaboration.

REFERENCES

Adelman, L. (1989). Management issues in knowledge elicitation. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 483–488.

Agnew, N. M., Brown, J. L., & Lynch, J. G. (1986). Extending the reach of knowledge engineering. *Future Computing Systems,* **1,** 115–141.

Alexander, J. H., Freiling, M. J., Shulman, S. J., Rehfuss, S., & Messick, S. L. (1987). Ontological analysis: An ongoing experiment. *International Journal of Man–Machine Studies,* **26,** 473–485.

Alluisi, E. A. (1967). Methodology in the use of synthetic tasks to assess complex performance. *Human Factors,* **9,** 375–384.

Anastasi, A. (1979). *Fields of applied psychology.* New York: McGraw–Hill.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard Univ. Press.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak method problem solutions. *Psychological Review,* **94,** 192–210.

Ashton, R. H. (1983). Research in auditing and decision making: Rationale, evidence, and implications. *Canadian Certified General Accountant's Monographs,* No. 6. Vancouver, BC: Canadian General Accountants Association.

Bailey, W. A., & Kay, D. J. (1987). Structural analysis of verbal data. In J. M. Carroll & P. Tanner (Eds.), *Human factors in computing systems and graphics interfaces* (pp. 297–301). London: Academic Press.

Bainbridge, L. (1979). Verbal reports as evidence of the process operator's knowledge. *International Journal of Man-Machine Studies,* **11,** 411–436.

Basden, A. (1984). AI: Cognition as composition. In F. Machlup & U. Mansfield (Eds.), *The study of information: Interdisciplinary messages* (pp. 237–262). New York: Wiley.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science,* **4,** 372–378.

Belkin, N. J., Brooks, H. M., & Daniels, P. J. (1987). Knowledge elicitation using discourse analysis. *International Journal of Man-Machine Studies,* **27,** 127–144.

Bellezza, F. S. (1992). Mnemonics and expert knowledge: Mental cueing. In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 204–217). Hillsdale, NJ: Erlbaum.

Benfer, R. A., & Furbee, L. (1989, November). Knowledge acquisition in the Peruvian Andes. *AI Expert,* 22–29.

Benjafield, J. (1969). Evidence that "thinking aloud" constitutes an externalization of inner speech. *Psychonomic Science,* **15,** 83–84.

Berg-Cross, G., & Price, M. E. (1989). Acquiring and managing knowledge using a conceptual structures approach: Introduction and framework. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 513–527.

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

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

Beyth-Marom, R., & Arkes, H. R. (1983). Being accurate but not necessarily Bayesian: Comments on Christensen-Szalanski and Beach. *Organizational Behavior and Human Performance,* **31,** 255–257.

Book, W. F. (1924). Voluntary motor ability of the world's champion typists. *Journal of Applied Psychology,* **8,** 283–308.

Boose, J. H. (1985). A knowledge acquisition program for expert systems based on personal construct psychology. *International Journal of Man-Machine Studies,* **23,** 495–525.

Boose, J. H. (1986). *Expertise transfer for expert system design.* Amsterdam: Elsevier.

Boose, J. H., & Bradshaw, J. M. (1987). Expertise transfer and complex problems: Using ACQUINAS as a knowledge-acquisition workbench for knowledge-based systems. *International Journal of Man-Machine Studies,* **26,** 3–28.

Boose, J. H., & Gaines, B. (Organizers) (1991). *The Sixth Banff Knowledge Acquisition for Knowledge-Based Systems Workshop.* Banff, Canada (Sponsored by the University of Calgary and Boeing Computer Services).

Bramer, M. (Ed.) (1985). *Research and development in expert systems.* Cambridge: Cambridge Univ. Press.

Breuker, J., & Wielinga, B. (1984). *Techniques for knowledge elicitation and analysis.* Report 1.5, ESPRIT Project 12, Amsterdam: University of Amsterdam.

Breuker, J., & Wielinga, B. (1985). KADS: Structured knowledge acquisition for expert systems. In Rault, J.-C. (Ed.), *Proceedings of the Fifth International Workshop on Expert Systems and their Applications.* Paris: Agence de l'Informatique.

Breuker, J., & Wielinga, B. (1987). Use of models in the interpretation of verbal data. In A. L. Kidd (Ed.), *Knowledge acquisition for expert systems: A practical handbook* (pp. 17–44). New York: Plenum Press.

Brown, B. (1989). The taming of an expert: An anecdotal report. In C. R. Westphal & K. L. McGraw (Eds.), *Special issue on knowledge acquisition, SIGART Newsletter* (No. 108, pp. 133–135). Special Interest Group on Artificial Intelligence, Association for Computing Machinery, New York.

Brown, C. W., & Ghiselli, E. E. (1953). The prediction of proficiency of taxicab drivers. *Journal of Applied Psychology,* **37,** 437–439.

Bruce, V. (1988). *Recognizing faces.* Brighton, England: Lawrence Erlbaum Associates.

Brule, J. F., & Blount, A. (1989). *Knowledge acquisition.* New York: McGraw–Hill.

Bryan, W. L., & Harter, N. (1897). Studies in the physiology and psychology of the telegraphic language. *Psychological Review,* **4,** 27–53.

Buchanan, B. G., Barstow, D., Betchal, R., Bennet, J., Clancey, W., Kulikowski, C., Mitchell, T., & Waterman, D. (1983). Constructing an expert system. In F. Hayes-Roth, D. Waterman, & D. Lenat (Eds.), *Building expert systems* (pp. 127–168). Reading, MA: Addison-Wesley.

Buchanan, B. G., & Shortliffe, E. H. (1984). Explanation as a topic of AI research. In B. G. Buchanan & E. H. Shortliffe (Eds.), *Rule-based expert systems: The MYCIN experiments on the Stanford Heuristic Programming Project* (pp. 331–337). Reading. MA: Addison-Wesley.

Buchanan, B. G., Sutherland, G. L., & Feigenbaum, E. A. (1969). Rediscovering some problems in artificial intelligence in the context of organic chemistry. In B. Meltzer & D. Michie (Eds.), *Machine intelligence 4* (pp. 209–254). Edinburgh: Edinburgh Univ. Press.

Buchner, A., Funke, J., & Berry, D. (1995). Negative correlations between control performance and verbalizable knowledge: Indicators for implicit learning in process control tasks? *Quarterly Journal of Experimental Psychology,* **48A,** 166–187.

Burton, A. M., Shadbolt, N. R., Hedgecock, A. P., & Rugg, G. (1987). A formal evaluation of a knowledge elicitation techniques for expert systems: Domain 1. In D. S. Moralee (Ed.), *Research and development in expert systems, Vol 4.* (pp. 35–46). Cambridge: Univ. Press.

Burton, A. M., Shadbolt, N. R., Rugg, G., & Hedgecock, A. P. (1988). Knowledge elicitation techniques in classification domains. In Y. Kodratoff (Ed.), *ECAI-88: Proceedings of the 8th European conference on artificial intelligence* (pp. 85–93). London: Pittman.

Burton, A. M., Shadbolt, N. R. Rugg, G., & Hedgecock, A. P. (1990). The efficacy of knowledge elicitation techniques: A comparison across domains and levels of expertise. *Journal of Knowledge Acquisition,* **2,** 167–178.

Butler, K. E., & Corter, J. E. (1986). Use of psychometric tools for knowledge acquisition: A case study. In W. A. Gale (Ed.), *Artificial intelligence and statistics* (pp. 295–319). Cambridge, MA: Addison-Wesley.

Bylander, T., & Chandrasekaran, B. (1987). Generic tasks for knowledge-based reasoning: The "right" level of abstraction for knowledge acquisition. *International Journal of Man-Machine Studies,* **26,** 231–243.

Carroll, J. S., & Siegler, R. S. (1977). Strategies for the use of base-

rate information. *Organizational Behavior and Human Performance,* **19,** 392–402.

Ceci, S. J., & Liker, J. K. (1986). A day at the races: A study of IQ, expertise, and cognitive complexity. *Journal of Experimental Psychology: General,* **115,** 255–266.

Chandrasekaran, B. (1986). Generic tasks in knowledge-based reasoning: High-level building blocks for expert system design. *IEEE Expert,* **1,** 23–30.

Chandrasekaran, B. (1983, Spring). Towards a taxonomy of problem solving types. *The AI Magazine,* **11,** 9–17.

Chandrasekaran, B. (1990, Winter). Design problem solving: A task analysis. *The AI Magazine,* **11,** 59–71.

Chase, W. G., & Ericsson, K. A. (1981). Skilled memory. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 141–189). Hillsdale, NJ: Erlbaum.

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

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science,* **13,** 145–182.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representations of physics problems by experts and novices. *Cognitive Science,* **5,** 121–152.

Chi, M. T. H., Glaser, R., & Farr, M. L. (Eds.) (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 1* (pp. 7–75). Hillsdale, NJ: Erlbaum.

Chignell, M. H., & Peterson, J. G. (1988). Strategic issues in knowledge engineering. *Human Factors,* **30,** 381–394.

Chiles, W. D. (1967). Methodology in the assessment of complex performance: Discussion and conclusions. *Human Factors,* **9,** 385–392.

Christensen, J. M., & Mills, R. G. (1967). What does the operator do in complex systems? *Human Factors,* **9,** 329–340.

Christensen-Szalanski, J. J., & Beach, L. R. (1984). The citation bias: Fad and fashion in the judgment and decision making literature. *American Psychologist,* **39,** 75–78.

Clancey, W. J. (1983). The epistemology of a rule-based expert system: A framework for explanation. *Artificial Intelligence,* **20,** 215–251.

Clancey, W. J. (1985). Heuristic classification. *Artificial Intelligence,* **27,** 215–251.

Clancey, W. J. (1988). The knowledge engineer as a student: Metacognitive bases for asking good questions. In H. Mandl & A. Lesgold (Eds.), *Learning issues for intelligent tutoring systems* (pp. 80–113). New York: Springer-Verlag.

Clancey, W. J. (1993). The knowledge level reinterpreted: Modeling socio-technical systems. In K. M. Ford & J. M. Bradshaw (Eds.), *Knowledge acquisition as modeling* (pp. 33–49, Pt. 1). New York: Wiley.

Clarke, B. (1987). Knowledge acquisition for real time knowledge based systems. In *Proceedings of the first European workshop on knowledge acquisition for knowledge-based systems.* Reading, England: Reading University.

Cleaves, D. A. (1987). Cognitive biases and corrective techniques: Proposals for improving elicitation procedures for knowledge-based systems. *International Journal of Man-Machine Studies,* **27,** 155–166.

Cochran, E. L., Bloom, C. P., & Bullemer, P. T. (1990). Increasing end-user acceptance of an expert system by using multiple experts: Case studies in knowledge acquisition. In C. R. Westphal & K. L. McGraw (Eds.), *Readings in knowledge acquisition: Current practices and trends* (pp. 73–89). London: Ellis Horwood.

Cohen, P. R., & Howe, A. E. (1989). Toward AI research methodology: Three case studies in evaluation. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 634–646.

Collett, P. (1979). The repertory grid in psychological research. In G. P. Ginsburg (Ed.), *Emerging strategies in social psychological research.* Chichester, England: Wiley.

Coltheart, V., & Walsh, P. (1988). Expert knowledge and semantic memory. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues, vol. 2* (pp. 524–530). Chichester, England: Wiley.

Cooke, N. M. (1992). Modeling expertise in expert systems. In R. R. Hoffman (Ed.). *The psychology of expertise: Cognitive research and empirical AI* (pp. 29–60). Hillsdale, NJ: Erlbaum.

Cooke, N. M., & McDonald, J. E. (1986). A formal methodology for acquiring and representing expert knowledge. *Proceedings of the IEEE,* **74,** 1422–1430.

Cooke, N. M., & McDonald, J. E. (1987). The application of psychological scaling techniques to knowledge elicitation for knowledge-based systems. *International Journal of Man-Machine Studies,* **26,** 533–550.

Coombs, M. J. (Ed.) (1984). *Developments in expert systems.* New York: Academic Press.

Cordingley, S. (1989). Knowledge elicitation techniques for knowledge-based systems. In D. Diaper (Ed.), *Knowledge elicitation: Principles, techniques, and applications* (pp. 89–175). Chichester, England: Ellis-Horwood.

Cox, P. A., & Balick, M. J. (1994, June). The ethnobotanical approach to drug discovery. *Scientific American,* **270,** 82–87.

Crandall, B. W., & Klein, G. A. (1987). Key components of MIS performance. Report, Klein Associates, Yellow Springs, OH.

Crandall, B. W. (1989). A comparative study of think-aloud and Critical Decision knowledge elicitation methods. In C. R. Westphal & K. L. McGraw (Eds.), *SIGART Newsletter: Special issue on knowledge acquisition, No. 108* (pp. 144–146). New York: Special Interest Group on Artificial Intelligence, Association for Computing Machinery.

Cross, T. B. (1988). *Knowledge engineering: The use of artificial intelligence in business.* New York: Bradley Books.

Cullen, J., & Bryman, A. (1988). The knowledge acquisition bottleneck: Time for reassessment? *Expert Systems,* **5,** 216–225.

Dahlstrom, D. O. (1989). Worlds of knowing and nonmonotonic reasoning. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 626–633.

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

Deane, T., & Kanazawa, K. (1989). Persistence and probabilistic projection. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 574–585.

DeGreef, P., & Breuker, J. (1985). A case study in structured knowledge acquisition. In A. Joshi (Ed.), *Proceedings of the 9th International Joint Conference on Artificial Intelligence* (pp. 390–392), Los Altos, CA: Kaufman.

DeGreef, P., Breuker, J., & Wielinga, B. (1986). *Statcons-1 design document.* ESPRIT Deliverable E2.2, Amsterdam: University of Amsterdam.

DeGreef, P., Breuker, J., Schreiber, G., & Wielemaker, J. (1988). StatCons: Knowledge acquisition in a complex domain. In *ECAI-88: Proceedings of the 8th European Workshop on Artificial Intelligence.*

DeGroot, A. D. (1965). *Thought and choice in chess.* The Hague: Mouton.

de Mantaras, L. R., Cortes, U., Manero, J., Plaza, E., Salra, X., & Agusti, J. (1986). Knowledge elicitation using personal constructs: Application to document classification. In *ECAI-86: Proceedings of the 7th European Workshop on Artificial Intelligence.*

Deffenbacher, K. (1988). Eyewitness research: The next ten years. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory, Vol. 2: Current research and issues* (pp. 20–26). Chichester, England: Wiley.

Dhaliwal, J. S., & Benbasat, I. (1990). A framework for the comparative evaluation of knowledge acquisition tools and techniques. *Knowledge Acquisition,* **2,** 145–166.

Diaper, D. (Ed.) (1989). *Knowledge acquisition: Principles, techniques, and applications.* New York: Wiley.

Dillard, J. F., & Mutchler, J. F. (1987). Expertise in assessing solvency problems. *Expert Systems,* **4,** 170–179.

Doyle, J. (1983, Summer). Methodological simplicity in expert system construction: The case of judgments and reasoned assumptions. *The AI Magazine,* **4,** 39–43.

Dreyfus, H., & Dreyfus, S. E. (1986). *Mind over machine.* New York: Free Press.

DuPreez, P. D., & Ward, D. G. (1970). Personal constructs of modern and traditional Xhosa. *Journal of Social Psychology,* **82,** 149–160.

Duda, J., Gaschnig, J., & Hart, P. (1979). Model design in the PROSPECTOR consultant system for mineral exploration. In D. Michie (Ed.), *Expert systems in the micro-electronic age* (pp. 153–167). Edinburgh: Edinburgh University Press.

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

Eastman Kodak Company, Inc. (1983). *Ergonomic design for people at work.* New York: Van Nostrand Reinhold.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–267). New York: Wiley.

Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology,* **59,** 562–571.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd. ed.). Cambridge, MA: MIT Press.

Ericsson, K. A., & Smith, J. (Eds.). (1991). *Toward a general theory of expertise.* Cambridge: Cambridge Univ. Press.

Evans, J. St. B. T. (1988). The knowledge elicitation problem: A psychological perspective. *Behavior and Information Technology,* **7,** 111–130.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Erlbaum.

Feigenbaum, E. A., Buchanan, B. G., & Lederberg, J. (1971). On generality and problem solving: A case study using the DENDRAL program. In B. Meltzer & D. Michie (Eds.), *Machine intelligence 6* (pp. 165–190). Edinburgh: Edinburgh Univ. Press.

Fellers, J. W. (1987). Key factors in knowledge acquisition. *Computer Personnel,* **11,** 10–24.

Feltovich, P. J., Spiro, R. J., & Coulson, R. L. (1989). The nature of conceptual understanding in biomedicine: The deep structure of complex ideas and the development of misconceptions. In D. Evans & V. Patel (Eds.), *Cognitive science in medicine* (pp. 113–172). Cambridge, MA: MIT Press.

Fischhoff, B. (1989). Eliciting knowledge for analytical representation. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 448–461.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin,* **51,** 327–358.

Fleishman, E. A. (1967). Performance assessment based on an empirically derived task taxonomy. *Human Factors,* **9,** 349–366.

Fleishman, E. A. (1975). Toward a taxonomy of human performance. *American Psychologist,* **30,** 1127–1149.

Ford, K. M., & Adams-Webber, J. R. (1992). Knowledge acquisition and constructive epistemology. In R. Hoffman (Ed.), *The cognition of experts: Psychological research and empirical AI* (pp. 121–136). Hillsdale, NJ: Erlbaum.

Ford, K. M., & Bradshaw, J. M. (Eds.). (1993). *Knowledge acquisition as modeling.* New York: Wiley.

Forsyth, D. E., & Buchanan, B. G. (1989). Knowledge acquisition for expert systems: Some pitfalls and suggestions. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 345–442.

Fox, J. (1986). Knowledge, decision making, and uncertainty. In W. A. Gale (Ed.), *Artificial intelligence and statistics* (pp. 57–76). Reading, MA: Addison-Wesley.

Fox, J., Myers, C. D., Greaves, M. F., & Pegram, S. (1985). Knowledge acquisition for expert systems: Experience in leukemia diagnosis. *Methods of Information in Medicine,* **24,** 65–72.

Fransella, F., & Bannister, D. (1971). *Inquiring man: The theory of personal constructs.* Harmondsworth: Penguin.

Fraser, J. M., Smith, P. J., & Smith, J. W. (1992). A catalog of errors. *International Journal of Man-Machine Studies,* **37,** 265–307.

Friedland, P. (1981). Acquisition of procedural knowledge from domain experts. In A. Drinan (Ed.), *Proceedings of the 7th international joint conference on artificial intelligence* (pp. 856–861). Los Altos, CA: Kaufman.

Gaeth, G. J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance,* **33,** 263–282.

Gagne, R. M. (1964). *Conditions of human learning.* New York: Holt, Rinehart and Winston.

Gaines, B. R., & Boose, J. H. (Eds.) (1988). *Knowledge acquisition tools for expert systems.* London: Academic Press.

Gale, W. A. (1987). Knowledge-based knowledge acquisition for a statistical consulting system. *International Journal of Man-Machine Studies,* **26,** 55–64.

Gammack, J. G. (1987). Different techniques, and different aspects of declarative knowledge. In A. L. Kidd (Ed.), *Knowledge acquisition for expert systems: A practical handbook* (pp. 137–164). New York: Plenum Press.

Gammack, J. G., & Young, R. M. (1985). Psychological techniques for eliciting expert knowledge. In M. Bramer (Ed.), *Research and development in expert systems* (pp. 105–112). Cambridge: Cambridge Univ. Press.

Garg-Janardan, C., & Salvendy, G. (1987). A conceptual framework for knowledge elicitation. *International Journal of Man-Machine Studies,* **26,** 521–531.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1985). Eyewitness memory enhancement in the police interview: Cognitive retrieval mnemonics versus hypnosis. *Journal of Applied Psychology,* **70,** 401–412.

Geiwitz, J., Klatzky, R. L., & McCloskey, B. P. (1988). *Knowledge acquisition techniques for expert systems: Conceptual and empirical comparisons.* Report No. DAAB07-87-C-A405, U.S. Army Communications Electronics Command, Fort Monmouth, NJ.

Gevarter, W. B. (1987). The nature and evaluation of commercial expert system building tools. *Computer,* **20,** 24–41.

Glaser, R. (1987). Thoughts on expertise. In C. Schooler & W. Schaie

(Eds.), *Cognitive functioning and social structure over the life course* (pp. 81–94). Norwood, NJ: Ablex.

Gorden, R. L. (1987). *Interviewing: Strategy, techniques, and tactics* (4th ed.). Chicago, IL: Dorsey Press.

Gordon, S. E. (1992). Implications of cognitive theory for knowledge acquisition. In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 99–120). Hillsdale, NJ: Erlbaum.

Gordon, S. E., Schmierer, K. A., & Gill, R. T. (1993). Conceptual graph analysis: Knowledge acquisition for instructional system design. *Human Factors,* **35,** 459–481.

Grover, M. D. (1983). A pragmatic knowledge acquisition methodology. In A. Bundy (Ed.), *IJCAI-83: Proceedings of the 8th international joint conference on artificial intelligence* (pp. 436–438). Los Altos, CA: Kaufmann.

Hall, L. O., & Kandel, A. (1988). Toward a methodology for building expert systems for imprecise domains. *International Journal of Expert Systems,* **1,** 237–252.

Hammond, K. R. (1966). Clinical inference in nursing, II: A psychologist's viewpoint. *Nursing Research,* **15,** 27–38.

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

Harrison, J. A., & Sarre, P. V. (1975). Personal construct theory in the measurement of environmental images: Applications. *Environment and Behavior,* **7,** 3–58.

Hart, A. (1986). *Knowledge acquisition for expert systems.* London: Kogan Page.

Hart, A. (1987). Role of induction in knowledge elicitation. In A. L. Kidd (Ed.), *Knowledge acquisition for expert systems: A practical handbook* (pp. 165–189). New York: Plenum Press.

Hayes, P. J., Ford, K. M., & Agnew, N. (1994, Fall). On babies and bathwater: A cautionary tale. *The AI Magazine,* **15,** 15–26.

Hayes-Roth, F., Waterman, D. A., & Lenat, D. B. (1983). *Building expert systems.* Reading, MA: Addison-Wesley.

Herrod, R., & Smith, M. (1986). The Campbell Soup story: An application of AI technology in the food industry. *Texas Instruments Engineering Journal,* **3,** 16–19.

Herzog, H. (1944). What do we really know about daytime serial listeners? In P. F. Lazarsfeld & N. F. Stanton (Eds.), *Radio research* (pp. 1942–1943). New York: Ovell, Sloan and Pearce.

Hink, R. F., & Woods, D. L. (1987, Fall). How humans process uncertain knowledge. *The AI Magazine,* **8,** 41–53.

Hinkle, D. N. (1965). *The change of personal constructs from the viewpoint of a theory of implications.* Ph.D. thesis, Department of Psychology, University of Ohio. Cited in Fransella & Bannister (1971).

Hoch, S., & Loewenstein, G. F. (1989). Outcome feedback: Hindsight and information. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **15,** 605–619.

Hoffman, P. J., Slovic, P., & Rorer, L. G. (1968). An analysis of variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin,* **69,** 338–349.

Hoffman, R. R. (1987, Summer). The problem of extracting the knowledge of experts from the perspective of experimental psychology. *The AI Magazine,* **8,** 53–66.

Hoffman, R. R. (1991). Human factors psychology in the support of forecasting: The design of advanced meteorological workstations. *Weather and Forecasting,* **6,** 98–110.

Hoffman, R. R. (Ed.). (1992a). *The psychology of expertise: Cognitive research and empirical AI.* Hillsdale, NJ: Erlbaum.

Hoffman, R. R. (1992b). Bibliography: Expertise in programming. In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 359–362). Hillsdale, NJ: Erlbaum.

Hoffman, R. R. (1992c). Bibliography: Automated knowledge elicitation, representation, and instantiation ("knowledge acquisition"). In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 346–358). Hillsdale, NJ: Erlbaum.

Hoffman, R. R. (in press). Can experts be trusted? How can expertise be defined? In J. Fleck & R. Williams (Eds.), *Exploring expertise.* London: Macmillan Press.

Hoffman, R. R. & Conway, J. A. (1989). Psychological factors in remote sensing: A review of some recent research. *GEOCARTO International,* **4,** 3–21.

Hoffman, R. R., & Deffenbacher, K. A. (1992). A brief history of applied cognitive psychology. *Applied Cognitive Psychology,* **6,** 1–48.

Holsapple, C. W., & Whinston, A. B. (Eds.) (1987). *Business expert systems.* Homewood, IL: Irwin.

Holt, D. L. (1987). Auditors' base rates revisited. *Accounting, Organizations, and Society,* **12,** 571–578.

Honeck, R. P., & Temple, J. G. (1992). Metaphor, expertise, and a PEST. *Metaphor and Symbolic Activity,* **7,** 237–252.

Howell, W. C. (1984). *Task influences in the analytic-intuitive approach to decision making.* Report, Contract No. N00014-82-001. Office of Naval Research, Bethesda, MD.

Jackson, P. (1990). *Introduction to expert systems* (2nd ed.). Workingham, England: Addison-Wesley.

Jacob, V. S., Gaultney, L. D., & Salvendy, G. (1986). Strategies and biases in human decision making and their implications for expert systems. *Behavior and Information Technology,* **5,** 119–140.

Jeffries, R., Turner, A., Polson, P., & Atwood, M. (1981). The processes involved in designing software. In R. J. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 255–283). London: Springer-Verlag.

Jenkins, J. J. (1953). Some measured characteristics of Air Force weather forecasters and success in forecasting. *The Journal of Applied Psychology,* **37,** 440–444.

Jenkins, J. J. (1978). Four points to remember: A tetrahedral model of memory experiments. In L. Cermak & F. Craik (Eds.), *Levels of processing and human memory* (pp. 429–446). Hillsdale, NJ: Erlbaum.

Johnson, L., & Johnson, N. (1987). Knowledge elicitation involving teachback interviewing. In A. L. Kidd (Ed.), *Knowledge elicitation for expert systems: A practical handbook* (pp. 91–108). New York: Plenum Press.

Johnson, N. E. (1985). Varieties of representation in eliciting and representing knowledge in IKBS. *International Journal of Systems Research and Information Science,* **1,** 69–90.

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

Johnson, P. E., Hassebrock, F., Duran, A. S., & Moller, J. H. (1982). Multimethod study of clinical judgment. *Organizational Behavior and Human Performance,* **30,** 201–230.

Johnson, P. E., & Thompson, W. B. (1981). Strolling down the garden path: Error prone tasks in expert problem solving. In A. Drinan (Ed.), *Proceedings of the 7th international joint conference on artificial intelligence* (pp. 215–217). Los Altos, CA: Kaufman.

Johnson, P. E., Zualkerman, I. A., & Garber, S. (1987). Specification

of expertise. *International Journal of Man-Machine Studies,* **26,** 161–181.

Johnson, P. E., Zualkerman, I. A., & Tukey, D. (1993). Types of expertise: An invariant of problem solving. *International Journal of Man-Machine Studies,* **39,** 641–665.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard Univ. Press.

Kaempf, G. L., Thorsden, M. L., & Klein, G. (1991). *Application of an expertise-centered taxonomy to training decisions.* Report No. MDA903-91-C-0050, U.S. Army Research Institute, Alexandria, VA.

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

Keller, R. (1987). *Expert systems technology: Development and applications.* Englewood Cliffs, NJ: Prentice–Hall.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition,* **11,** 123–141.

Kelly, G. A. (1955). *The psychology of personal constructs.* New York: Norton.

Kidd, A. L. (Ed.). (1987). *Knowledge acquisition for expert systems: A practical handbook.* New York: Plenum Press.

Kidd, A. L., & Cooper, M. B. (1985). Man–machine interface issues in the construction and use of an expert system. *International Journal of Man–Machine Studies,* **22,** 91–102.

Kim, J., & Courtney, J. F. (1988). A survey of knowledge acquisition techniques and their relevance to managerial problem domains. *Decision Support Systems,* **4,** 269–284.

Kinney, W. R., & Uecker, W. C. (1982). Mitigating the consequences of anchoring in auditor judgment. *Accounting Review,* **57,** 55–69.

Kirwan, B., & Ainsworth, L. K. (1992). *A guide to task analysis.* London: Taylor & Francis.

Klein, G. A. (1987). Applications of analogical reasoning. *Metaphor and Symbolic Activity,* **2,** 201–218.

Klein, G. A. (1992). Using knowledge engineering to preserve corporate memory. In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 170–190). Hillsdale, NJ: Erlbaum.

Klein, G. (1993). *State-of-the-art report: Naturalistic decision making: Implications for design.* Report DLA-900-88-0393, Crew Systems Ergonomics Information Analysis Center, Wright-Patterson Air Force Base, OH.

Klein, G. A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 462–472.

Klein, G. A., & Hoffman, R. R. (1993). Perceptual-cognitive aspects of expertise. In M. Rabinowitz (Ed.). *Cognitive science foundations of instruction* (pp. 203–226). Hillsdale, NJ: Erlbaum.

Klein, G. A., & Weitzenfield, J. (1982). The use of analogues in comparability analysis. *Applied Ergonomics,* **13,** 99–104.

Kolodner, J. L. (1983). Towards an understanding of the role of experience in the evolution from novice to expert. *International Journal of Man–Machine Studies,* **19,** 497–518.

Kolodner, J. L. (1991, Summer). Improving decision making through case-based decision aiding. *The AI Magazine,* **12,** 52–68.

Krovvidy, S., & Wee, W. G. (1993). Wastewater treatment systems for case-based reasoning. In *Machine learning 10* (pp. 341–363). Dordrecht, The Netherlands: Kluwer.

Kuipers, B., & Kassirer, J. P. (1987). Knowledge acquisition by analysis of verbatim protocols. In A. L. Kidd (Ed.), *Knowledge acquisition for expert systems: A practical handbook* (pp. 45–71). New York: Plenum Press.

Kuipers, B., Moskowitz, A. J., & Kassirer, J. P. (1988). Critical decisions under uncertainty. *Cognitive Science,* **12,** 177–210.

LaFrance, M. (1992). Excavation, capture, collection, and creation: Computer scientists' metaphors for eliciting human expertise. *Metaphor and Symbolic Activity,* **7,** 135–156.

Laskey, K. B., Cohen, M. S., & Martin, A. W. (1989). Representing and eliciting knowledge about uncertain evidence and its implications. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 536–545.

Lerner, D. (1956). Interviewing Frenchmen. *American Journal of Sociology,* **1,** 187–194.

Levi, K. (1989). Expert systems should be more accurate than human experts: Evaluation procedures from human judgment and decision making. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 647–657.

Libby, R., & Lewis, B. L. (1977). Human information processing research in accounting: The state of the art. *Accounting, Organizations, and Society,* **21,** 245–268.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology,* **7,** 560–572.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance,* **26,** 49–171.

Madni, A. M. (1988). The role of human factors in expert system design and acceptance. *Human Factors,* **30,** 395–414.

Major, N., & Reichgelt, H. (1990). ALTO: An automated laddering tool. In B. Wielinga, J. Boose, B. Gaines, G. Schreiber, & M. van Someren (Eds.), *Current trends in knowledge acquisition.* Amsterdam: IOS Press.

McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 651–696). Chicago: Rand–McNally.

McGraw, K. L., & Harbison-Briggs, K. (1989). *Knowledge acquisition: Principles and guidelines.* Englewood Cliffs, NJ: Prentice–Hall.

McGraw, K., & Riner, A. (1987). Task analysis: Structuring the knowledge acquisition process. *Texas Instruments Technical Journal,* **4,** 16–21.

McGraw, K. L., & Seale, M. R. (1988). Knowledge elicitation with multiple experts: Considerations and techniques. *Artificial Intelligence Review,* **2,** 31–44.

McIntosh, P. S. (1986). *Knowledge acquisition for THEO: An expert system for solar flare forecasting.* Paper presented at the Conference on Artificial Intelligence Research in Environmental Science, National Oceanic and Atmospheric Administration, Boulder, CO.

McKeithen, K. B., Reitman, J. S., Rueter, H. H., & Hirtle, S. C. (1981). Knowledge organization and skill differences in computer programmers. *Cognitive Psychology,* **13,** 307–325.

Means, M. L., & Voss, J. F. (1985). Star wars: A developmental study of expert and novice knowledge structures. *Journal of Memory and Language,* **24,** 746–757.

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

Meehl, P. E. (1959). A comparsion of clinicians with five statistical methods of identifying psychotic MMPI profiles. *Journal of Counseling Psychology,* **6,** 102–109.

Meister, D. (1985). *Behavioral analysis and measurement methods.* New York: Wiley.

Merton, R. K., Fiske, M., & Kendall, P. L. (1956). *The focused interview.* Glencoe, IL: The Free Press.

Merton, R. K., & Kendall, P. L. (1946). The focused interview. *American Journal of Sociology,* **51,** 541–557.

Meyer, M. A., Booker, J. M., & Bradshaw, J. M. (1990). A flexible six-step program for defining and handling bias in knowledge elicitation. In B. Wielinga, J. Boose, B. Gaines, G. Schreiber, & M. van Someren (Eds.), *Current trends in knowledge acquisition.* Amsterdam: IOS Press.

Meyer, M. A., & Payton, R. C. (1992). Towards an analysis and classification of approaches to knowledge acquisition from examination of textual metaphor. *Knowledge Acquisition,* **4,** 347–369.

Michalski, R. S., & Chilausky, R. L. (1980, June). Knowledge acquisition by encoding expert rules versus computer induction from examples: A case study involving soybean pathology. *International Journal of Man–Machine Studies,* **12,** 63–87.

Mitchell, A. A. (1987). The use of alternative knowledge acquisition procedures in the development of a knowledge-based media planning system. *International Journal of Man-Machine Studies,* **26,** 399–411.

Mittal, S., & Dym, C. L. (1985, Summer). Knowledge acquisition from multiple experts. *The AI Magazine,* **6,** 32–36.

Monk, A. (1985). How and when to collect behavioral data. In A. Monk (Ed.), *Fundamentals of human–computer interaction* (pp. 69–79). New York: Academic Press.

Motta, E., Rajan, T., & Eisenstadt, M. (1989). A methodological tool for knowledge acquisition in KEATS-2. In J. Boose & B. Gaines (Eds.), *Proceedings of the Third Knowledge Acquisition for Knowledge-Based Systems Workshop* (pp. 21.1-21.20). Alberta, Canada: Department of Computer Science, University of Alberta.

Motta, E., Eisenstadt, M., West, M., Pitman, K., & Evertsz, R. (1987). *KEATS: The knowledge engineer's assistant.* Technical Report No. 20, HCRL, Open University, Milton Keynes, England.

Mullin, T. M. (1989). Experts' estimation of uncertain quantities and its implications for knowledge acquisition. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 616–625.

Neale, I. M. (1988). First generation expert systems: A review of knowledge acquisition methodologies. *Knowledge Engineering Review,* **3,** 105–146.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems.* New York: Wiley.

Neches, R., Swartout, W. R., & Moore, J. (1984). Enhanced maintenance and explanation of expert systems though explicit models and their development. In *Proceedings of the IEEE workshop on principles of knowledge-based systems* (pp. 173–183). New York: IEEE.

Neisser, U. (1993). Toward a skillful psychology. In D. Rogers & J. A. Sloboda (Eds.), *The acquisition of symbolic skills* (pp. 1–17). New York: Plenum.

Newell, A. (1981). The knowledge level. *Artificial Intelligence,* **18,** 87–127.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice–Hall.

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

Noble, D. F. (1989). Schema-based knowledge elicitation for planning and situation assessment. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 473–482.

Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of Dermatology,* **125,** 1063–1068.

Olson, C. L. (1976). Some apparent violations of the representativeness heuristic in human judgment. *Journal of Experimental Psychology: Human Perception and Performance,* **2,** 599–608.

Olson, J., & Reuter, H. (1987). Extracting expertise from experts: Methods for knowledge acquisition. *Expert Systems,* **4,** 152–168.

Osborn, A. (1953). *Applied imagination: Principles and procedures of creative thinking.* New York: Scribner.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science,* **10,** 91–116.

Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance,* **21,** 209–219.

Prerau, D. S. (1985, Summer). Selection of an appropriate domain for an expert system. *The AI Magazine,* **6,** 26–30.

Prerau, D. (1989). *Developing and managing expert systems: Proven techniques for business and industry.* Reading, MA: Addison Wesley.

Prietula, M. J., Feltovich, P. J., & Marchak, F. (1989). A heuristic framework for assessing factors influencing knowledge acquisition. In D. Blanning & D. King (Eds.), *Proceedings of the 22nd Hawaii international conference on systems science, Vol. 3: Decision support and knowledge-based systems* (pp. 419–426). New York: IEEE.

Reddy, R. (1988, Winter). Foundations and grand challenges of artificial intelligence. *The AI Magazine,* **9,** 9–21.

Regoczei, S. B., & Hirst, G. (1988). *Knowledge acquisition as knowledge explication by conceptual analysis.* Report No. 205, Computer Systems Research Institute, University of Toronto, Toronto, Ontario, Canada.

Regoczei, S. B., & Hirst, G. (1992). Knowledge and knowledge acquisition in the computational context. In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 12–28). Hillsdale, NJ: Erlbaum.

Reid, W. A., & Holley, B. J. (1972). An application of repertory grid techniques to the study of choice of university. *British Journal of Educational Psychology,* **42,** 52–59.

Rolandi, W. G. (1986). Knowledge engineering in practice. *AI Expert,* **1,** 58–62.

Rook, F. W., & Croghan, J. W. (1989). The knowledge acquisition activity matrix: A systems engineering conceptual framework. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 586–597.

Salter, W. J. (1988). Human factors in knowledge acquisition. In M. Helander (Ed.), *Handbook of human–computer interaction* (pp. 957–968). Amsterdam: North-Holland.

Sanders, M. S., & McCormick, E. J. (1987). *Human factors in engineering and design.* New York: McGraw–Hill.

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., & DeMaio, J. C. (1985). Measuring the structure of expertise. *International Journal of Man-Machine Studies,* **23,** 699–728.

Schwartz, S., & Griffin, T. (1986). *Medical thinking: The psychology of medical judgment and decision making.* New York: Springer Verlag.

Schweickert, R., Burton, A. M., Taylor, N. K., Corlett, E. N., Shadbolt, N. R., & Hedgecock, A. P. (1987). Comparing knowledge elicitation techniques: A case study. *Artificial Intelligence Review,* **1,** 245–253.

Scribner, S. (1984). Studying working intelligence. In B. Rogoff & S. Lave (Eds.), *Everyday cognition: Its development in social context.* (pp. 9–40). Cambridge: Harvard Univ. Press.

Senjen, R. (1988). Knowledge acquisition by experiment: Developing test cases for an expert system. *AI Applications in Natural Resource Management,* **2,** 52–55.

Shadbolt, N. R., & Burton, A. M. (1990a). Knowledge elicitation. In E. N. Wilson & J. R. Corlett (Eds), *Evaluation of human work: Practical ergonomics methodology* (pp. 321–345). London: Taylor and Francis.

Shadbolt, N. R., & Burton, A. M. (1990b). Knowledge elicitation techniques: Some experimental results. In K. L. McGraw & C. R. Westphal (Eds.) *Readings in knowledge acquisition* (pp. 21–33). New York: Ellis Horwood.

Shadbolt, N. R., & Wielinga, B. (1990). Knowledge-based knowledge acquisition: The next generation of support tools. In B. Wielinga, J. Boose, B. Gaines, G. Schreiber, & M. van Someren (Eds.), *Current trends in knowledge acquisition.* Amsterdam: IOS Press.

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologia,* **68,** 203–215.

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

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

Shaw, M. L. G., & Gaines, B. R. (1987). An interactive knowledge elicitation technique using personal construct technology. In A. L. Kidd (Ed.), *Knowledge acquisition for expert systems: A practical handbook* (pp. 109–136). New York: Plenum Press.

Shields, M. D., Solomon, I., & Waller, W. S. (1987). Effects of alternative sample space representations on the accuracy of auditors' uncertainty judgments. *Accounting, Organizations, and Society,* **12,** 375–385.

Shortliffe, E. H. (1976). *Computer-based medical consultations: MYCIN.* New York: Elsevier.

Sjoberg, G., & Nett, R. (1968). *A methodology for social research.* New York: Harper and Row.

Slade, S. (1991, Spring). Case-based reasoning: A research paradigm. *The AI Magazine,* **12,** 42–55.

Slovic, P. (1969). Analyzing the expert judge: A description of stockbrokers' decision processes. *Journal of Applied Psychology,* **53,** 255–263.

Smith, R. G., & Baker, J. D. (1983). The Dipmeter advisory system: A case study in commercial expert system development. In A. Bundy (Ed.), *IJCAI-83: Proceedings of the 8th International Joint Conference on Artificial Intelligence* (pp. 122–129). Los Altos, CA: Kaufman.

Solvberg, I., Nordbo, I., Vestli, M., Aakvik, G., Amble, T., Eggen, J., & Amodt, A. (1988). *METAKREK: Methodology and toolkit for knowledge acquisition.* Report No. STF14-A88046, Computing Centre, University of Trondheim, Trondheim, Norway.

Sowa, J. F. (1984). *Conceptual structures: Information processing in mind and machine.* New York: Addison-Wesley.

Spiro, R. J., Vispoel, W. P., Schmitz, J. G., Samarapungavan, A., & Boerger, A. E. (1987). Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In B. K. Britton & S. M. Glynn (Eds.), *Executive control processes in reading* (pp. 177–199). Hillsdale, NJ: Erlbaum.

Spradley, J. P. (1979). *The ethnographic interview.* New York: Holt, Rinehart and Winston.

Steels, L. (1990, Fall). Components of expertise. *The AI Magazine,* **11,** 2.

Stefik, M., Aikins, J., Balzer, R., Benoit, J., Birnbaum, L., Hayes-Roth, F., & Sacerdoti, E. (1982). The organization of expert systems: A tutorial. *Artificial Intelligence,* **18,** 135–173.

Stein, E. (1992). A method to identify candidates for knowledge acquisition. *Journal of Management and Information Systems,* **9,** 161–178.

Sternberg, R. J., & Frensch, P. J. (1992). On being an expert: A cost-benefit analysis. In R. R. Hoffman (Ed.), *The psychology of expertise: Cognitive research and empirical AI* (pp. 191–203). Hillsdale, NJ: Erlbaum.

Swartout, W. R. (1983). XPLAIN: A system for creating and explaining expert consulting programs. *Artificial Intelligence,* **21,** 285–325.

Taylor, E. C. (1985, Summer). Developing a knowledge engineering capability in the TRW Defense Systems Group. *The AI Magazine,* **6,** 58–63.

Tetmeyer, D. C. (1976). *Comparable item approach to establishing frequency of maintenance and maintenance tasks for a new aircraft.* Report, Escape and Human Factors Branch, Aeronautical Systems Division, Wright Patterson Air Force Base, OH.

Tolcott, M. A., Marvin, F. F., & Lehner, P. E. (1989). Expert decision making in evolving situations. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 606–615.

Trimble, G., & Cooper, C. (1987). Experience with knowledge acquisition for expert systems in construction. SERC Report RAL-87-055. In *Proceedings of the SERC workshop on knowledge acquisition for engineering applications.* London: Didcot.

Turban, E., & Liebowitz, J. (Eds.) (1992). *Managing expert systems.* Harrisburg, PA: Idea Group Publishing.

Tuthill, G. S. (Ed.) (1990). *Knowledge engineering: Concepts and practices for knowledge-based systems.* Blue Ridge Summit, PA: Tab Professional and Reference Books.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review,* **4,** 293–315.

Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (Eds.) (1981). *On scientific thinking.* New York: Columbia Univ. Press.

Umbers, I. G., & King, P. J. (1981). An analysis of human decision-making in cement kiln control and the implications for automation. In E. H. Mamdani & B. R. Gaines (Eds.), *Fuzzy reasoning and its applications* (pp. 369–381). London: Academic Press.

Vandierendonck, A. (1993). *Effects of acquisition method and similarity in category learning of archaeological objects.* Paper presented at the 34th meeting of the Psychonomic Society, Washington DC, November 5–7.

Waterman, D. A. (1986). *A guide to expert systems.* Reading, MA: Addison-Wesley.

Weiser, M., & Shertz, J. (1984). Programming problem representation in novice and expert programmers. *International Journal of Man-Machine Studies,* **19,** 391–398.

Weiss, S., & Kulikowski, C. (1984). *A practical guide to designing expert systems.* Totowa, NJ: Rowman and Allanheld.

Weitzel, J. R., & Kerschberg, L. (1989). A system development methodology for knowledge-based systems. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 598–605.

Wexley, K. N., & Yukl, G. A. (1984). *Organizational behavior and personnel psychology.* Homewood, IL: Irwin.

Wick, M. R., & Slagle, J. R. (1989). The partitioned support network for expert system justification. *IEEE Transactions on Systems, Man, and Cybernetics,* **19,** 528–534. New York: IEEE.

Wickens, C. .D., Merwin, D. H., & Lin, E. L. (1994). Implications of graphics enhancements for the visualization of scientific data. *Human Factors,* **36,** 44–61.

Wielinga, B. J., & Breuker, J. A. (1985). Interpretation of verbal data for knowledge acquisition. In T. O'Shea (Ed.), *Advances in artificial intelligence* (pp. 3–12). Amsterdam: North-Holland.

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

Wolf, W. A. (1989). Knowledge acquisition from multiple experts. In C. R. Westphal & K. L. McGraw (Eds.), *Special issue on knowledge acquisition, SIGART Newsletter,* No. 108, pp. 138–140. New York: Special Interest Group for Artificial Intelligence, Association for Computing Machinery.

Wood, L. E., & Ford, J. M. (1993). Structuring interviews with experts during knowledge elicitation. In K. M. Ford & J. M. Bradshaw (Eds.), *Knowledge acquisition as modeling* (pp. 71–90, Pt. 1). New York: Wiley.

Woods, D. D. (1993). Process-tracing methods for the study of cognition outside the experimental psychology laboratory. In G. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok, C. E. (Eds.) *Decision making in action: Models and methods* (pp. 228–251). Norwood, NJ: Ablex.

Received: July 9, 1993

Woods, D. D., & Hollnagel, E. (1987). Mapping cognitive demands in complex problem-solving worlds. *International Journal of Man-Machine Studies,* **26,** 257–275.

Woodworth, R. S. (1938). *Experimental psychology.* New York: Holt.

Wright, G. (1984). *Behavioral decision theory: An introduction.* Beverley Hills, CA: Sage.

Wright, G., & Ayton, P. (1987). Eliciting and modeling expert knowledge. *Decision Support Systems,* **3,** 13–26.

Young, R. M., & Gammack, J. (1987). The role of psychological techniques and intermediate representations in knowledge elicitation. In *Proceedings of the First European Workshop on Knowledge Acquisition for Knowledge-based Systems.* Reading, England: Reading University.

Zadeh, L. A., & Kacprzuk, J. (Eds.) (1992). *Fuzzy logic for the management of uncertainty.* New York: Wiley.

Zakay, D., & Wooler, S. (1984). Time pressure, training, and decision effectiveness. *Ergonomics,* **27,** 273–284.

Zsambok, C. E., & Klein, G. A. (Eds.) (1995). *Naturalistic decision making.* Hillsdale, NJ: Erlbaum.