The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice

By: L. Richard Ye
Department of Accounting and
  MIS, BA&E
California State University,
  Northridge
Northridge, CA 91330-8245
U.S.A.
rye@vax.csun.edu

Paul E. Johnson
Department of Information and
  Decision Sciences
University of Minnesota
271 19th Ave. So.
Minneapolis, MN 55455
U.S.A.
pjohnson@csom.umn.edu

Abstract

Providing explanations for recommended actions is deemed one of the most important capabilities of expert systems (ES). There is little empirical evidence, however, that explanation facilities indeed influence user confidence in, and acceptance of, ES-based decisions and recommendations. This paper investigates the impact of ES explanations on changes in user beliefs toward ES-generated conclusions. Grounded on a theoretical model of argument, three alternative types of ES explanations—trace, justification, and strategy—were provided in a simulated diagnostic expert system performing auditing tasks. Twenty practicing auditors evaluated the outputs of the system in a laboratory setting. The results indicate that explanation facilities can make ES-generated advice more acceptable to users and that justification is the most effective type of explanation to bring about changes in user attitudes toward the system. These findings are expected to be generalizable to application domains that exhibit similar characteristics to those of auditing: domains in which decision making tends to be judgmental and yet highly consequential, and the correctness or validity of such decisions cannot be readily verified.

Introduction

Expert systems (ES) are computer programs capable of performing specialized tasks based on an understanding of how human experts perform the same tasks. Few ESs, however, are targeted at replacing their human counterparts; usually they are intended to function as assistants or advisers to professional people with different technical background and problem-solving experience (Berry and Hart, 1990; Feigenbaum, et al., 1988; Leonard-Barton and Sviokla, 1988). To be useful and acceptable, it has been argued, an ES must not only perform at a level comparable to a human expert's, but also must be able to explain, in a form understandable to users, the reasoning processes it employs to solve problems and make recommendations (Duda and Shortliffe, 1983; Moore and Swartout, 1988; Teach and Shortliffe, 1981).

Central to the issue of explanation are two unique characteristics of ES applications. First, ESs are often developed to help make relatively unstructured decisions, and a time lag may exist between when such decisions must be made and when their quality can be assessed. As a result, the acceptance of ES-generated advice is more likely to be determined by its reasonableness than by its correctness. Second, real-world decisions have practical—financial, legal, political, and social—consequences. If users are to remain responsible for the decisions made, they are unlikely to accept a system's recommendation if they do not understand its underlying reasoning processes (Hollnagel,

1987). An explanation facility provides the potential to make an ES more useful and acceptable by increasing user understanding of, and confidence in, its decisions and recommendations.[1]

Past research in ES explanations has focused on establishing "existence proofs," i.e., solving the theoretical and practical problems of explanation *generation* (cf. AAAI, 1990; AAAI, 1992; Moore and Swartout, 1988). Given the potentially high costs of making these explanations available, as exemplified by the difficulties earlier research has encountered, there appears to be little guarantee that automated explanations will produce the positive impact on the uses of ESs as expected. Indeed, there have been documented cases in which users question the utility of having an explanation facility (Slatter, et al., 1988) or ignore the explanation facilities altogether (Hart and Wyatt, 1990).

In this light, our paper reports an empirical investigation of the value of ES explanations to users. Specifically, a laboratory setting was employed to study the impact of alternative types of explanations on user acceptance of ES-generated advice. The scope of the study was limited to diagnostic problem-solving, a process in which a system's true states are inferred from observable but noisy data called symptoms. At the conceptual level, diagnosis represents an important class of problems across a variety of domains, and it was among the earliest candidates for ES applications (Clancey, 1985; Davis, 1993; Stefik, et al., 1982; Torasso and Console, 1989).

The next section provides the conceptual basis for the study and develops the research hypotheses. The following section employs a detailed discussion of the research method employed. The balance of the paper presents the results and discusses the findings.

Conceptual Basis of Research

According to the *American Heritage Dictionary*, to *explain* is "1. to make plain or comprehensi-

ble; 2. to define, expound; and 3. to offer reasons for or a cause of, to justify." Within the context of using expert systems to solve problems, the term "explanation" has been used very loosely to cover almost any request for further information. For example, users might have a need for further operational instruction, more data, explication of terms, feedback, or justification of the reasoning methods used or advice given. Clearly, in each case both the form and the content of the information needed could be different, and grouping them all together under the same term, "explanation," would not be helpful.

This study adopts a typology of ES explanations that has been the focus of prior research. Together, three types of ES explanations are identified (Chandrasekaran, et al., 1988; Clancey, 1983; 1993; Neches, et al., 1985; Swartout, 1983; Wick, 1992):[2]

1. **Trace,** or **Line of Reasoning,** which refers to a record of the inferential steps taken by an ES to reach a conclusion.

2. **Justification,** which is an explicit description of the causal argument or rationale behind each inferential step taken by the ES.

3. **Strategy,** which is a high-level goal structure that determines how the ES uses its domain knowledge to accomplish a task.[3]

---

expanding ES are non-trivial tasks, and an explanation facility can help developers recall the encoded problem-solving knowledge, identify faulty reasoning steps, and locate errors (Neches, et al., 1985).

[2] While the user's need for explanations stems from the unique characteristics of ES applications, the ability to provide explanations is admittedly technology driven. The typology adopted in this study is strongly influenced by the rule-based tradition, one that remains the most popular and intuitive method of representing and reasoning about domain knowledge. Newer technologies, such as *belief networks* and *neural networks*, have been able to achieve expert-level performance without a cognitive model of the human expert. However, because the internal structures of these systems bear little resemblance to problem-solving expertise understood in human terms, it has been substantially more difficult to explain their reasoning processes to the user (Zehedi, 1993; but also see Suermondt, 1992 and Klein, 1994, for progress being made on explaining quantitative decision-making models).

[3] *Definitions,* which include descriptions of terms and descriptions of differences between terms, are sometimes referred to as explanations (see Nettler, 1970 for a detailed discussion of *explanation by definition*). However, since a definition of something usually involves straightforward information retrieval, not reasoning, the field has not considered it a major issue of explanation.

---

[1] While the focus of this paper is the impact of explanations on ES users, an explanation facility has obvious value to ES *developers*. Validating and maintaining an evolving and

The next section introduces a conceptual model developed by Toulmin (1958) for the process of argumentation. This model is then applied to the three types of explanations to form the theoretical basis for our research.

Toulmin's model of argument

Stephen Toulmin (1958) formalizes a model of argument to reflect the "rational process" (p. 7) characteristic of human reasoning. Drawing an analogy from claims and argument made in the courts, Toulmin proposes the structure and procedures by which "claims-in-general" can be argued for and accepted.

Toulmin makes a clear distinction between the "field-invariant" and "field-dependent" aspects of argument (p.15). Regardless of the context that surrounds the argument, its field-invariant aspect consists of six elements (also see Ehninger and Brockriede, 1960; 1978; Toulmin, et al., 1984):

1. *Claims (C)*—assertions or conclusions put forward for general acceptance. A claim is always of a potentially controversial nature.

2. *Data (D)*—statements specifying the particular facts or previously established beliefs about a situation based on which a claim is made.

3. *Warrants (W)*—statements that justify (or certify) the reasonableness of leaping from data to a claim.

4. *Backing (B)*—the general body of information or experience that assures the trustworthiness of a warrant. Backing is not needed unless the validity of the warrant is challenged.

5. *Qualifiers (Q)*—phrases expressing the degree of certainty placed on a claim. No qualifier is needed if a claim is considered indisputable.

6. *Possible Rebuttals (R)*—extraordinary or exceptional circumstances that might defeat the warranted claim. The function of a rebuttal is analogous to a safety valve and therefore is optional.

Argument is typically made in such a manner that the following relationships among the six

elements exist (Ehninger and Brockriede, 1960, p. 45):
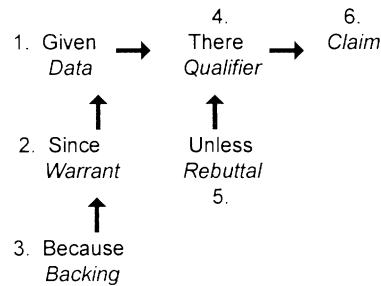
Despite this basic, context-free structure said to be at work in any argument, according to Toulmin (1958), an analytical framework of argument is not complete unless it includes the larger human enterprise whose purposes argument serves; hence the field-dependent aspect of argument. For example, in different domains, there are differences in the degree of formality and precision that argument must satisfy in order to be acceptable. The fundamental force of medical argument is realized only to the extent that the enterprise of medicine itself is understood, and the same is true in business, politics, law, or any other fields.

Explanation as argumentation

Toulmin's model of argument is significant in the way it highlights the discrete response steps that an ES explanation facility should follow in order to answer user queries in a convincing way. Increasingly, researchers of ES explanation facilities have drawn upon studies in argumentative dialogues to generate explanation dialogues (e.g., Cavalli-Sforza and Moore, 1992; Moore and Swartout, 1989; Quillici, 1991). The relevance of the model to each type of explanation is discussed next, with examples drawn from the domain of auditing.

Trace

The opportunity of using trace as a source of explanation first appeared with the introduction of rule-based ES (Buchanan and Shortliffe, 1984). A rule is a representation of a data-conclusion association, where a conclusion can be inferred from a logical combination of some premises.

Moreover, *certainty factors* (CF) can be added to a rule to indicate how strongly the premises have been confirmed or how probable the conclusion can be inferred from the premises. Because the conclusion clause of one rule can be used to confirm the premises of another rule, a collection of rules used together become a network of individual inference steps. A trace is a record of the system's run-time, rule-invocation history. Properly presented, a trace allows interactive question-answering about the system's reasoning steps: "How a request for data is related to a goal, how one goal leads to another, and how a goal is achieved (Clancey, 1983, p. 217)."[4]

A rule represented as a *Premise-CF-Conclusion* triad corresponds to a *Data (D)-Qualifier (Q)-Claim (C)* structure in Toulmin's model of argument. The other three elements of the model are missing, however: warrant, backing, and rebuttals. Trace, which provides a chain of rules invoked by an ES, is insufficient to explain the system's reasoning processes, unless for each rule in the chain the user already understands exactly why the premises (*D*) necessarily, or conditionally if the CF (*Q*) is less than 100 percent, lead to the conclusion (*C*).
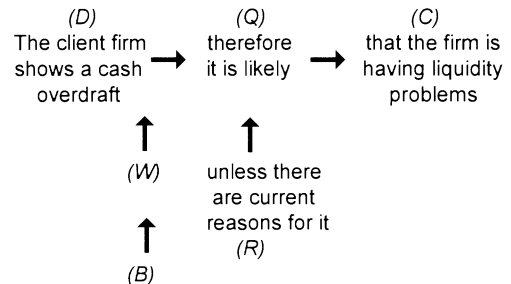
Justification

By encoding large chunks of knowledge as empirical associations, ESs are constructed for efficient problem solving at the expense of omitting the intermediate reasoning steps on which these associations are based. Consequently, such systems lack the support knowledge necessary for justifying to the user why, for example, a conclusion follows naturally from its premises. To illustrate, consider the following excerpt from the verbal protocol generated by an auditor while performing an audit task:

> The cash overdraft always raises red flags. It's one of the items that we always look at. Or maybe I shouldn't say that it always raises a warning flag, but when people get into a cash overdraft situation then there should be current reasons for it. Otherwise you should be aware that they're having liquidity problems.

Viewed as argument, the auditor's line of reasoning corresponds to the following Toulmin structure.
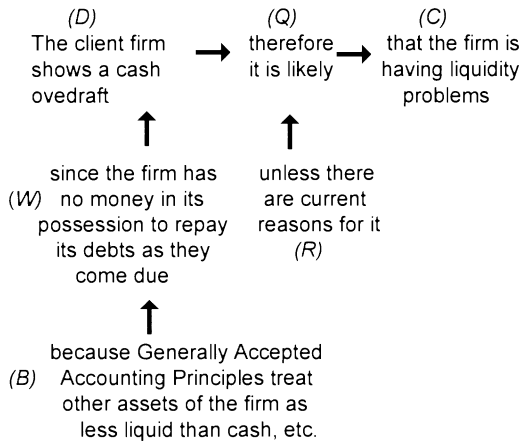

(D)		(Q)		(C)
The client firm shows a cash overdraft	→	therefore it is likely	→	that the firm is having liquidity problems
↑		↑		
(W)		unless there are current reasons for it		
↑		(R)		
(B)				

Because the auditor does have some reservations about the conclusion, a qualifier (*Q*) and a rebuttal (*R*) are included to serve as an escape clause. On the one hand, it would be rather straightforward to encode this chunk of knowledge into a rule that, if reproduced by an ES to answer questions, becomes part of a *trace*. On the other hand, the reason that a cash overdraft should signal a liquidity problem in the first place is less than obvious. To interested observers, the problem-solver's "inferential leap" from the data to the claim needs to be justified.

Probed further, the auditor may offer new information to support his/her claim by adding a warrant (*W*) and perhaps even a backing (*B*). The first diagram on the next page completes an example of Toulmin's model of argument at work.

There are two practical reasons that *justification*, which requires a deeper understanding of the domain, can be important. First, by demonstrating that the conclusions developed by the system are based on sound reasoning, justification

---

[4] Many ES development tools use a trace facility to answer two types of questions: WHY and HOW. The WHY question is typically interpreted as "why did you (the system) request that data," and is answered by showing the conclusion of a rule that can be inferred from the data requested (the goal to achieve). The HOW question, on the other hand, typically means "how did you reach that conclusion," and is answered by showing all the data used to lead to the conclusion. These interpretations, however, do not account for all possible user intentions and are therefore inadequate as a method of classifying explanations. See, for example, Lehnert (1978), who proposes 13 types of questions for story understanding, including different kinds of *why* and *how* questions, and Graesser and Murachver (1985), who identify seven question functions and three statement categories to form a taxonomy of 21 question types.

```
       (D)              (Q)            (C)
  The client firm    therefore   that the firm is
  shows a cash       it is likely  having liquidity
  ovedraft                          problems
        ↑                ↑
  since the firm has  unless there
(W)  no money in its  are current
  possession to repay  reasons for it
  its debts as they       (R)
  come due
        ↑
  because Generally Accepted
(B)  Accounting Principles treat
  other assets of the firm as
  less liquid than cash, etc.
```

helps increase ES users' confidence in the system's problem-solving competence and hence, the acceptability of the conclusions. Second, because ESs can only achieve high performance within relatively narrow problem areas, justification enables users to make more informed decisions on whether the system's advice should be followed. For example, in boundary cases or unusual situations, the lack of specific knowledge may cause an ES to respond to a problem inappropriately, and justification forces it to reveal such limitations.
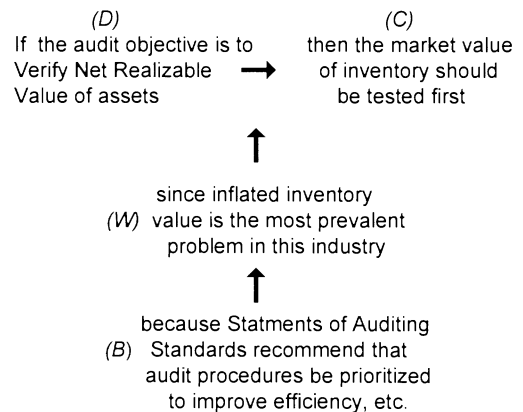
Both the warrant and the backing of argument serve the function of justification, but at different stages. If the warrant is challenged, the backing becomes the next line of defense, opening a new round of argument. The process does not end until either the warrant is accepted, or no further backing can be offered.[5] Toulmin's framework shows where justification for a line of reasoning should be focused. Examined within the present context, it also points out the type of information a knowledge engineer needs to extract from a problem solver, if the knowledge obtained is to be used later for explanation purposes.

Strategy

Instead of asking what knowledge is being applied by an ES (trace) and its underlying rationale (justification), users may be interested in

knowing why information is gathered in a certain order, why one particular chunk of knowledge is invoked before others, and how individual reasoning steps contribute to a high-level objective. To answer these questions, the system must have access to knowledge about its problem-solving *strategy*. Moreover, because strategic knowledge often involves general principles in problem solving, it also finds use in knowledge-based tutoring systems that teach, among other things, high-level reasoning and problem-solving skills within specific domains (e.g., Clancey, 1986; Nickerson, et al., 1985; Regian and Shute, 1992).

Structurally, explaining the system's strategy is similar to providing justification in the sense that the system must clarify why it solves a problem by following a specific procedure (Clancey, 1993). The Toulmin diagram below provides such an example (qualifiers and rebuttals omitted).

```
        (D)                     (C)
If the audit objective is to   then the market value
Verify Net Realizable    →     of inventory should
Value of assets                be tested first
                    ↑
       since inflated inventory
(W)  value is the most prevalent
       problem in this industry
                    ↑
       because Statments of Auditing
(B)  Standards recommend that
       audit procedures be prioritized
       to improve efficiency, etc.
```

Cost of automated explanations

ES explanation facilities that provide users with access to justification or strategic knowledge are more difficult and more costly to develop than those relying primarily on a trace of the system's execution paths. The raw materials from which a trace is constructed are part of the system's knowledge base, and providing a trace as explanation is largely a problem of presentation. Rules, for example, are often encoded internally in a succinct way for efficient storage and execution, and making them comprehensible to a user typically involves simple translation using a standardized natural-language template (Barr, et al., 1989).

---

[5] As a result of this recurring relationship, Ehninger and Brockriede (1978) propose a simplified model known as *a unit of proof*, which includes every element of Toulmin's model except backing.

On the other hand, because strategic knowledge tends to be buried implicitly in the knowledge base, and an ES does not need a representation of justification knowledge in order to execute, neither type of information is readily available for explanation purposes. Not only must additional efforts be spent uncovering the knowledge to be used for explanation, but more fundamental issues may arise concerning the psychological validity of the explanation knowledge so acquired.

It is common practice, for example, to study decision making by a variety of process tracing techniques, i.e., observing the human expert performing a task and then analyzing a record of the entire problem-solving process (Todd and Benbasat, 1987). It is considered rather disruptive, however, to ask the expert for an explanation of his/her reasoning while the decision-making process is under way (Ericsson and Simon, 1984; Nisbett and Wilson, 1977). An alternative is to ask the expert to provide a *retrospective* rationalization of the decisions made, but there is evidence that people are not always conscious of their belief structures and therefore have marked difficulties in providing post hoc justification that is consistent with the quality of their performance or with their problem-solving behavior (Berry and Broadbent, 1984; Norman, 1983; Wason and Evans, 1975).

Contrary to common beliefs, a truly informative explanation facility does not come free with the acquisition of some associative knowledge and the encoding of that knowledge in a computer program for problem solving. The resolution of many non-trivial theoretical and practical difficulties in automated explanations demands a significant amount of research and development resources. It is important that we establish the relative merits and usefulness of this technology, lest we run the risk that the resources devoted to constructing "existence proofs" might not be justified.

Development of hypotheses

The purpose of this study is to examine the impact of ES explanations on users' acceptance of ES-developed conclusions and to identify the type of explanation that is most effective in producing such an impact.

Unlike conventional information systems that function in relatively structured task environments, the nature of ES applications suggests that the conclusions and recommendations developed by an ES can be potentially controversial. To reduce the amount of uncertainty associated with ES-based decisions and increase the system's acceptability to users, an explanation facility is necessary.

ES explanations serve the function of argument. Because the primary purpose of argument is to advance a problematic claim and have it accepted, for the argument to be considered successful, there must be a change from an existing belief to the adoption of a new belief on the part of the audience (Achinstein, 1983; Dretske, 1988; Ehninger and Brockriede, 1978; Toulmin, et al., 1984). The change is unlikely to take place, however, unless the audience agrees with the evidence (data) and endorses the principle that is expressed or implicit in a warrant (Ehninger and Brockriede, 1978). It follows that, by making available the data an ES uses to reach conclusions and the rationale that justifies such inferential leaps, an explanation facility has the potential to affect users' beliefs in the system's conclusions. This expectation leads to the following hypothesis:

H1: ES explanations can affect users' beliefs in the system's conclusions.

Of the three types of explanations discussed above, justification appears to have the highest potential to change user beliefs in an ES-generated conclusion. *Strategy*, while explaining why things are done in a certain order, provides information at the wrong level. The fact that there is a separate body of "domain-general" knowledge (Clancey, 1993, p. 199), which governs the use of domain-specific knowledge, suggests that ES conclusions must be expressed at a level consistent with the advice the user is seeking (i.e., as shown in the earlier example, an ES for **financial statements analysis** will advise that the client might have liquidity problems, while timing the analysis of inventory market value might be an internal task of that system; an ES for **audit planning,** on the other hand, will most likely offer the timing of specific task performance as advice to the user). If strategy does not provide information at the system's

conclusion level, it is unlikely to impact users' beliefs in these conclusions.

Judged from the perspective of argument, if the system's conclusion is indeed controversial, additional information provided by *trace*, i.e., data used to reach the conclusion, may be insufficient to bring about changes in users' beliefs. By definition, data are pre-established facts or beliefs. Data alone are unlikely to provide much added value, unless (1) the user is unaware of the existence of such data, and (2) the user will interpret the data in the same way and reach the same conclusion as the system does.

If the user does not understand the underlying rationale of an inferential leap that the system is making (hence the controversy), only *justification* provides the right type of information that can help bridge the gap. Providing justification for the system's conclusions may or may not completely resolve the controversy, but making it available at least presents the opportunity to allow the system's side of the story to be understood (Toulmin, et al., 1984). If users find justification most useful in helping them judge the acceptability of the system's conclusions, it is expected that they will use this type of explanation more often and will spend more time examining it. These expectations lead to the following hypothesis:

H2: The most effective type of explanation in helping to change users' beliefs is justification.

　　a. Users will request justification more frequently than they will other types of explanations.

　　b. Users will spend more time examining justification than they will with other types of explanations.

Research Method

This research was conducted in a laboratory setting using a simulated ES. This section describes the domain, research design, development of the stimulus material and the system, dependent measures, participants, and experimental procedures.

Domain

The domain of auditing was chosen for the study of ES explanations. First, auditing is diag-

nostic in nature, since the overall objective of auditing is to detect discrepancies between the financial statements of an enterprise and its true financial health. Second, audit decisions are often judgmental and notably consequential; the validity of many audit decisions is either impossible to evaluate in a short term, or is too costly to verify. As a result, auditors must be prepared to defend, before their clients and their colleagues, every conclusion they make (Ashton, et al., 1989).

Design

In this study, participants evaluated the outputs of an ES performing an *analytical review*, a procedure widely employed by auditors to collect evidence (Blocker and Willingham, 1985). Because of the lack of access to a real ES for analytical review, a computer program was developed to simulate the user interface of such a system. As a tool frequently employed in the study of human-computer interaction (e.g., Good, et al., 1984; Gould, et al., 1983), simulation allows various design alternatives to be explored in a realistic research environment. At the same time it avoids the use of such technologies as natural language processing and text generation, which are yet to become widely available.

A one-group pretest-post-test design was employed in this study. *Explanation* was the independent variable: the overall presence or absence of explanation was manipulated, and the participants' belief in the system's conclusions *before* and *after* they received explanation were measured. There are potential confounding effects associated with a one-group design, specifically the history effect and the maturation effect (Emory, 1991). In this study, however, the duration between the pretest and the post-test was considered too short to cause these effects.

Material

The stimulus material for the experiment consisted of an audit case and the outputs of the simulated ES (the system). Originally designed for the study of auditing expertise (see Johnson, et al., 1989; 1991), the case involved a publicly traded medical products company in which a series of frauds were perpetrated by senior manage-

ment. Adapted for our study, the case material included a narrative description of the business and a set of financial statements and associated notes. Having been tested and improved over time, the case appeared to reproduce the complexity and the challenge of the participants' usual task environment.

The outputs of the system included (1) a series of audit conclusions related to various parts of the financial statements, and (2) three alternative explanations for each conclusion. Two domain consultants assisted with the development of these outputs. First, using data collected from prior research, a series of conclusions were developed. Next, using Toulmin's model of argument, templates were developed to constrain the syntactic structure that each explanation type would conform to and to define the nature of its information content. Three explanation texts (trace, justification, and strategy) were then developed for each conclusion based on input from one of the consultants. To ensure a high quality of explanation material, the initial set of explanations so constructed was presented to the second consultant for critique, and a revised version was shown to the first consultant. This process was repeated until all major disagreements were resolved.

To ensure the three explanations pertaining to the same conclusion could be considered equivalent in terms of their *readability*, the explanation material was further calibrated. First, the material was rated for subjective readability by 15 MBA students completing a course in auditing. Their ratings were used to make readability improvement to the material. The explanation texts were then empirically calibrated by 32 upper-division business students completing an auditing course. Reading rate and recall accuracy were employed as indirect measures of readability (Kintsch and Vipond, 1979). For each conclusion developed, if its explanations were found to have statistically significant differences in readability, both the conclusion and the explanations would be excluded from the material. The final version of the material included 14 conclusions and their corresponding explanations.

System

A typical user-ES dialogue may include the system's requests for data, the user's data entry as

requested, the system's presentation of conclusions and recommendations, and the user's requests for explanations pertaining to any specific data request or conclusion. Because of the large quantity of data required by an auditor for the analytical review task, interactive data entries were deemed too time-consuming to be part of the experiment. It was assumed, therefore, that the data entry routine would be completed in advance and that the system would not ask the participants for interactive data inputs.[6]

The resulting dialogue included the following system outputs: (a) a client company profile at the beginning of the review, (b) a series of audit conclusions related to various parts of the financial statements, (c) three types of explanation for each conclusion, and (d) a review summary to highlight the system's recommendations. Prior studies (Johnson, et al., 1989; 1991) indicate that a typical strategy used by auditors on analytical review tasks is to first develop a model of the client company, and then identify potential audit risks associated with such a model. Because auditors following the strategy tend to examine client financial statements in their standard presentation sequence (i.e., a narrative followed by a balance sheet, an income statement, a statement of changes in financial position, etc.), the dialogue presented various conclusions in the same order to reflect the use of that strategy.

To make the system look "real," intentional delays were embedded in the user-system dialogue, especially prior to the presentation of each conclusion. This would create the impression that the system needed time to develop the output. The amount of keyboard inputs required on the part of participants was minimal. In addition, an online help facility was available in anticipation of potential operational difficulties. Finally, there was a built-in routine that would capture all user keystrokes and elapsed time (in 1/100 seconds) between these keystrokes. The system went through several rounds of pilot tests with academic and practicing auditors. The results of these tests indicated that the system would appear realistic to participants.

---

6 Commercial expert systems for financial planning applications typically use such an input strategy because of large quantities of client data.

Dependent measures

Change in belief was the primary dependent variable. For each conclusion produced by the system, participants were asked to indicate, on a seven-point scale, the extent to which they believed that the conclusion was true or reasonable, *before* and *after* they received explanations offered by the system. The before-explanation score measures participants' existing belief in the system's conclusion, and the after-explanation score measures their newly adopted belief. The difference between these two scores was used to measure whether explanations had any impact on participants' belief in, and hence their acceptance of, the system's conclusion (Cohen, 1989; Dretske, 1988).

Two other dependent variables—choice of explanation and elapsed reading time—were used to measure the degree to which participants actually found uses for individual types of explanations. Choice of explanation was recorded as the frequency of requests participants made for each explanation type. Elapsed reading time was measured by recording elapsed time between keystrokes and calculating the average time participants spent examining individual explanations (in seconds per 100 words of text). These usage measures provide a more direct indicator of the relative value of alternative explanation types to users.

Participants

Twenty practicing auditors from a large public accounting firm participated as subjects. They ranged from staff-level accountants with two years of practice to audit managers with six or more years of practice in the public accounting profession. While it was not practical to employ randomization procedures in selecting the participants, participation was voluntary, and there was no obvious reason to believe that the sample selected was systematically biased in any respect.

Procedures

A two-hour experiment was administered individually to participants in a microcomputer laboratory. All instructions were provided in written form. The experiment was divided into two sessions: a review session followed by an evalu-

ation session. During the review session, participants were asked to assume the role of a manager on an audit team. Their task was to conduct an analytical review on a set of draft financial statements of a client company (the audit case) and to provide a written summary identifying any unusual account relationships that might lead to further investigation. The purpose was to engage participants in active problem-solving and to allow them to become familiar with the case material.

At the beginning of the evaluation session, participants were presented with a different context: they were asked by their firm to evaluate a computer-based decision aid for analytical review. Their task was to provide critiques on the conclusions and explanations developed by the system and to make a recommendation as to whether their firm should consider adopting such a system. Throughout the experiment, participants were unaware that the system they evaluated was a simulation.

The first half of the evaluation session served a training purpose. The system presented eight conclusions. Participants practiced reviewing each conclusion and the corresponding three types of explanations. During the second half, the system presented six conclusions. For each conclusion presented by the system, participants were first asked to indicate to what extent they believed that the conclusion was true or reasonable. They were then allowed to either proceed to the next conclusion, or freely request and review any of the three types of explanations. If participants requested at least one explanation, before they proceeded to the next conclusion, they were asked again to indicate to what extent they *now* believed the system's present conclusion was being true or reasonable. Throughout the evaluation session, participants were also invited to provide, in writing, free-form comments about the system's outputs.

Results

Change in belief

Change in belief was the difference between participants' post-explanation belief and pre-

explanation belief scores. Prompted with the statement, "I believe the system's conclusion is true or reasonable," prior to and after receiving explanations, participants provided ratings on identical seven-point scales, ranging from "1" for "strongly disagree" to "7" for "strongly agree." Table 1 presents the summary statistics on these measures and the results of a paired T-test performed on change in belief. There was a significant increase in participants' belief in the system's conclusions, after they reviewed explanations provided by the system. The result thus provided support for H1.

Table 1. Summary Statistics and Paired T-test Results on Change in Belief

	Pre-Expl. Belief	Post-Expl. Belief	Change in Belief*
Mean	5.56	6.13	0.57
Std. Deviation	1.38	1.00	1.01
		t value	5.69
		D.F.	101**
		2-tail Prob.	.000

*Post-explanation belief **minus** Pre-explanation belief.

**There were a total of 120 cases (20 participants x 6 conclusions). In 18 of these cases, participants did not request an explanation, and therefore did not have post-explanation belief scores, resulting in 102 data points.

Choice of explanation

Table 2 shows the number of times a request was made for each type of explanation, as well as its share (in percentage) of the total number of requests made. The results indicate that justification was the most frequently requested explanation type, followed by trace. Strategy was the least frequently requested explanation type. Also shown in Table 2, a one-dimensional x^2 test found significant differences among the total number of requests made for each type of explanation. This result provided support for H2.a.

Table 2. Summary Statistics and Chi-Square Test Results on Choice of Explanation

	Number of Requests	Percentage of Total Requests Made
Justification	95	41.9%
Trace	80	35.2%
Strategy	52	22.9%
Total	227	100%
c2	D.F.	Significance
12.59	2	.002

Elapsed reading time

Table 3 presents the average elapsed time participants spent examining each type of explanation (adjusted for the amount of text read), as well as the results of a repeated-measures ANOVA test. Participants spent relatively little time reading strategy, substantially more time examining trace, and the greatest amount of time inspecting justification. The ANOVA test found significant differences among the three average reading times. The result provided support for H2.b.

Table 3. Summary Statistics and ANOVA Test Results on Elapsed Reading Time*

	Mean (Sec./100 Words)	Standard Deviation
Justification	21.3	2.35
Trace	13.1	1.96
Strategy	4.6	1.78
F	D.F.	Significance
827.9	2.92	.000

*Calculations based on 47 cases in which participants requested all three types of explanation.

Discussion

This section discusses the findings and the limitations of the study, and the implications of the results for research and practice.

Interpretations of findings

The results of this study are consistent with the view that explanations can have a positive impact on user acceptance of an expert system. After they had the opportunity to learn about the system's reasoning processes through the explanations, participants appeared more convinced about the soundness of the system's conclusions, as demonstrated by their increased belief in those conclusions. Indeed, 17 (85 percent) of the 20 participants' overall impression of the system was so positive that their final recommendations of the evaluation were all in favor of adopting the system.[7]

---

[7] The remaining three participants, all senior-level auditors, were more skeptical about the value of the system, mainly because they felt they could do better than what the system could.

More specifically, the study found justification to be the most effective type of explanation in making the system's conclusions more acceptable, as evidenced both by the highest frequency at which justification was requested, and by the unproportionately large amount of time participants spent analyzing it. Participants' informal comments also provided support for their discrete usage patterns, as a number of them suggested that they would always want to see the justification for a conclusion, and they wondered why conclusion and justification were not presented together. These findings are consistent with Toulmin's prediction in his model of argument that a potentially controversial claim cannot be supported by data alone, and that a warrant will be necessary in order for the claim to be considered acceptable (Toulmin, 1958; Toulmin, et al., 1984).[8]

In addition to providing support for the research hypothesis concerning the usefulness of justification, the results indicate other usage patterns. Participants also made a relatively high number of requests for explanations based on trace, which provided information on the specific data items the system used to reach conclusions. Considering the fact that the system did not request user data inputs, participants' need for trace seems understandable. Although they were exposed to the case material during the review session, they might have either ignored some case data or failed to remember all the details. One question naturally results: would users still need trace if they were asked to provide all data inputs? In this case, users most likely would want to know why certain data items were being requested, and trace would instead provide information in terms of the conclusions the system would try to draw based on the data supplied (see an earlier discussion on page 4, and footnote 4, of this paper). In view of Toulmin's model of argument, trace is useful because it makes explicit both the data used and the claim inferred from the data, but it is insufficient to cause changes in users' beliefs.

Despite their role as critics of the system, participants did not appear to be concerned about the system's strategic knowledge. In particular, they spent a brief amount of time examining strategic explanation: they seemed disappointed with the information provided and lost interest quickly. In their written comments, participants suggested that one role of strategic explanations was to remind the auditors about the higher-level audit objectives to be accomplished (i.e., the "bigger picture" to be maintained) while analyzing a specific situation. This finding is consistent with the expectation discussed earlier—that information not directly supporting the system's conclusions will not affect users' beliefs. Beyond its role of providing an overall picture of the task at hand, participants also commented on the potential value of strategic explanations in auditor training. This observation again supports the view that strategic knowledge might be most useful in knowledge-based tutoring systems (Clancey, 1986; Nickerson, et al., 1985; Regian and Shute, 1992).

Finally, for exploratory purposes the impact of participants' domain experience on their need for explanations was also examined. The collected data were divided into two groups: data from auditors with two to four years of practice (novice group) and data from auditors with six or more years of practice (expert group). The data were then compared. The change in belief scores for the two groups was comparable, indicating they were equally likely to be influenced by explanations. Both groups also requested justification most frequently, followed by trace and strategy, in that order. These informal results seem to suggest that level of domain experience is not a factor in assessing the impact of explanations on users nor in determining the most desirable explanation type.

On the other hand, the novice group did make a significantly higher number of requests for explanations than the expert group. Novices made a total of 133 requests (56 for justification, 46 for trace, and 31 for strategy), compared to experts' 94 requests (30 for justification, 34 for trace, and 21 for strategy). The difference came as no surprise, however: one would expect that more

---

[8] Note that a warrant is still a *necessary* condition, but not a *sufficient* condition (see Mackie, 1980).

experienced auditors should be able to understand the system's conclusions better and therefore have a lesser need for additional information.

Limitations

The main limitations of the study center on the scope of the experiment, the limited user-system dialogs, and the experimental task used. First, the experiment focused on a single problem domain. The generalizability of the results would likely increase, for example, if two or three different domains were studied. The resource requirements, on the other hand, would be substantial in terms of developing and testing the stimulus materials for each domain, which would demand large amounts of domain consultants' time. An argument can be made, however, that ES applications in many other domains share the characteristics of those in auditing and therefore can benefit from the study's findings. For example, medical diagnosis, credit appraisal, insurance underwriting, and trouble-shooting complex mechanical or physical systems all involve the use of judgments. Decision making in these domains can be highly consequential, and the correctness or validity of these decisions is either impossible or too expensive to quickly verify (Torasso and Console, 1989).

Second, the user-system interface used in the experiment did not allow different problem context for explanation to be explored. For example, the lack of data entry requirements would not prompt potential user questions on why the system needed specific data (trace would have been used to answer such questions), and it was not possible for participants to volunteer information. While the user-system dialog of the study was designed to emphasize ease of operation and controlled data collection, the interpretation of the results must take these conscious design choices into consideration.

Third, in this experiment the participants' main task was to *evaluate* a decision aid, not to *use* it to aid in decision making. While studying such "first-time" user behavior is important because it helps assess the acceptability of an ES as judged from its explanation capability, it is nevertheless artificial. Caution must be exercised in deciding whether the findings could be generalized to experienced users, or to the real-world task environment in which users will truly be held responsible for any decisions they make or adopt.

Implications for future research

The efficacy of ES explanations is often intuitively assumed by the research community in explanation generation technologies, but not formally tested (except for a survey of *potential* ES users by Teach and Shortliffe, 1981). By providing empirical support for the claim that explanations can influence users' acceptance of ES outputs, the results of this study enhance the legitimacy of research in explanation development technologies. The results also suggest that better and more in-depth understanding of the process of ES explanations is warranted. In addition, further investigation should be conducted on the domain characteristics that make ES explanations necessary and helpful. In our research, the nature of the auditing domain might explain a strong need for explanation facilities. Future research should focus on a more precise identification and definition of the factors in an ES task environment that dictate the usefulness of an explanation facility. For example, such a facility may be superfluous to users if the quality of the system's advice can be tested immediately, as might be the case in ESs developed for software debugging. After all, there appear to be many ESs in use now that offer at best very primitive explanation facilities. It is also possible, however, that ESs have not been more widely deployed in some domains because they are considered unacceptable without truly useful explanation capabilities.

This study found justification to be the most preferred explanation type. As a result, a different set of research questions relating to the effectiveness of justification can be introduced. Specifically, more in-depth inquiries on justification may benefit from earlier research in *warrant*, the key construct in Toulmin's model of argument. Rhetoricians, for instance, have studied warrants that are based on a rich variety of reasoning strategies, such as causal reasoning, classification, generalization (case-based reasoning), analogical reasoning, and the use of a combination of these strategies for different types of argument (e.g., Ehninger and Brock-

riede, 1960; 1978). Given the importance of justification to ES users, one future research direction may be to assess the impact of different kinds of justification and to identify the appropriate conditions under which each should be employed. For example, while justification based on causal reasoning—more likely referred to as *scientific explanations* (Achinstein, 1983; Mackie, 1980; Nettler, 1970)—might be most convincing, it will not always be available. In many domains, such as in stock market investment, the court of law, or career counseling, reasoning based on statistical evidence, past cases, or analogies may be a human problem-solver's best explanation strategy.

This study is an initial step toward formally examining the impact of ES explanation on users. A potentially more fruitful but also more challenging direction of research will be to study the impact of ES explanations in real work environments. More fundamental questions should be asked about the utility patterns of ES explanations in such settings, and about the impact of explanation facilities on users' decision-making behavior and performance. Obviously, these questions are of greater significance to ES developers and user organizations—questions that cannot be answered readily until they are studied in field settings and over an extended period of time.

Implications for practice

The practical problem addressed in this study was the following: can ES explanation facilities influence users' decisions on whether to adopt the advice of the system, and what kind of explanation is most effective in influencing users' decisions? From an ES developer's perspective, the study's findings suggest that explanations have the potential to make the system's conclusions more acceptable, if the underlying applications demonstrate characteristics similar to those of the auditing domain (see earlier argument under the *Limitations* section).

The results also suggest that developers of explanation facilities for an auditing ES might consider making justification available because it appears to be a more useful type of information in helping increase user confidence in the system. Indeed, as suggested by several partici-

pants in their written comments, there is a question concerning the extent to which presentation of justification should be separated from the system's conclusions. Some believed that justification should be included by default as part of the system's conclusions, while others preferred to have control over what information to receive. In any case, the results further suggest that users should be given the flexibility to decide whether justification will be automatically presented with the advice being offered by the system, much as they have a choice with conventional interfaces capable of providing both menu- and command-driven dialogue structures, or both tabular and graphic presentation formats.

Developing an ES explanation facility with the capability to provide justification knowledge will be resource intensive. Many issues must be addressed, including the acquisition of valid domain knowledge, its internal representation, and automatic generation and presentation of comprehensible justification as explanations. Currently, a "canned text" approach remains the most efficient implementation technique (Wexelblat, 1989). The approach, similar to what was used in this study, offers predetermined natural language responses by anticipating possible user queries. Because the system has no control over the content of the explanation, however, this approach is likely to cause serious updating and maintenance problems over time. In the long run, newer theories such as the one that treats explanation generation as "reconstructive problem-solving" promise to provide much more flexible and maintainable explanation facilities (see Wick and Thompson, 1992).

Acknowledgements

References

AAAI. *Proceedings of the AAAI Workshop on Explanation,* Eighth National Conference on Artificial Intelligence, Boston, MA, July 1990.

AAAI. Working Notes for the AAAI Spring Symposium on Producing Cooperative Explanations, Stanford University, Stanford, CA, March 1992.

Achinstein, P. *The Nature of Explanation*, Oxford University Press, New York, 1983.

Ashton, R.H., Kleinmuntz, D.N., Sullivan, J.B., and Tomassini, L.A. "Audit Decision Making," in *Research Opportunities in Auditing: The Second Decade*, A.R. Abdel-khalik and I. Solomon (eds.), American Accounting Association, Sarasota, FL, 1989, pp. 95-132.

Barr, A., Cohen, P.R., and Feigenbaum, E.A. *The Handbook of Artificial Intelligence* (IV), Addison-Wesley, Reading, MA, 1989.

Berry, D.C. and Broadbent, D.E. "On the Relationship Between Task Performance and Associated Verbalizable Knowledge," *The Quarterly Journal of Experimental Psychology* (36A), May 1984, pp. 209-231.

Berry, D.C. and Hart, A. *Expert Systems: Human Issues,* MIT Press, Cambridge, MA, 1990.

Blocher, E. and Willingham, J.J. *Analytical Review*, McGraw-Hill, New York, 1985.

Buchanan, B.G. and Shortliffe, E.H. (eds.). *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, Reading, MA, 1984.

Cavalli-Sforza, V. and Moore, J.D. "Collaborating on Arguments and Explanations," in Working Notes for the AAAI Spring Symposium on Producing Cooperative Explanations, Stanford University, Stanford, CA, March 1992.

Chandrasekaran, B., Tanner, M.C., and Josephson, J.R. "Explanation: The Role of Control Strategies and Deep Models," in *Expert Systems: The User Interface*, J.A. Hendler (ed.), Ablex, Norwood, NJ, 1988, pp. 219-247.

Clancey, W.J. "The Epistemology of a Rule-Based Expert System: A Framework for Explanations," *Artificial Intelligence* (20), May 1983, pp. 215-251.

Clancey, W.J. "Heuristic Classification," *Artificial Intelligence* (27), December 1985, pp. 289-350.

Clancey, W.J. "From GUIDON to NEOMYCIN and HERACLES in Twenty Short Lessons: ORN Final Report 1979-1985," *AI Magazine* (7:4), August 1986, pp. 40-60.

Clancey, W.J. "Notes on 'Epistemology of a Rule-Based Expert System'," *Artificial Intelligence* (59), February 1993, pp. 197-204.

Cohen, L.J. "Belief and Acceptance," *Mind* (98:391), July 1989, pp. 367-389.

Davis, R. "Retrospective on 'Diagnostic Reasoning Based on Structure and Behavior'," *Artificial Intelligence* (59), February 1993, pp. 149-158.

Dretske, F.I. *Explaining Behavior: Reasons in a World of Causes*, MIT Press, Cambridge, MA, 1988.

Duda, R.O. and Shortliffe, E.H. "Expert Systems Research," *Science* (220), April 15, 1983, pp. 261-268.

Ehninger, D. and Brockriede, W. "Toulmin on Argument: An Interpretation and Application," *Quarterly Journal of Speech* (46), February, 1960, pp. 44-53.

Ehninger, D. and Brockriede, W. *Decision by Debate* (2nd ed.), Harper & Row, New York, 1978.

Emory, C.W. *Business Research Methods* (4th ed.), Irwin, Homewood, IL, 1991.

Ericsson, K.A. and Simon, H.A. *Protocol Analysis: Verbal Report as Data*, MIT Press, Cambridge, MA, 1984.

Feigenbaum, E., McCorduck, P., and Nii, H.P. *The Rise of the Expert Company*, Vintage Books, New York, 1988.

Good, M.D., Whiteside, J.A., Wixon, D.R., and Jones, S.J. "Building a User-Derived Interface," *Communications of the ACM* (27:10), October 1984, pp. 1032-1043.

Gould, J.D., Conti, J., and Hovanyecz, T. "Composing Letters with a Simulated Listening Typewriter," *Communications of the ACM* (26:4), April 1983, pp. 295-308.

Graesser, A.C. and Murachver, T. "Symbolic Procedures of Questions Answering," in *The Psychology of Questions*, A.C. Graesser and J.B. Black (eds.), Lawrence Erlbaum, Hillsdale, NJ, 1985, pp. 15-88.

Guindon, R. (eds.). *Cognitive Science and Its Applications for Human-Computer Interaction*, Lawrence Erlbaum, Hillsdale, NJ, 1988.

Hart, A. and Wyatt, J. "Connectionist Models in Medicine: An Investigation of Their Potential," *Proceedings of Artificial Intelligence in Medicine*, AIME89, Springer Verlag, London, 1990.

Hollnagel, E. "Commentary: Issues in Knowledge-Based Decision Support," *International Journal of Man-Machine Studies* (27), November/December 1987, pp. 743-751.

Johnson, P.E., Jamal, K., and Berryman, R.G. "Audit Judgment Research," *Accounting, Organizations and Society* (14:1/2), January/February 1989, pp. 83-99.

Johnson, P.E., Jamal, K., and Berryman, R.G. "Effects of Framing on Auditor Decisions," *Organizational Behavior and Human Decision Processes* (50), October 1991, pp. 75-105.

Kintsch, W. and Vipond, D. "Reading Comprehension and Readability in Educational Practice and Psychological Theory," in *Perspectives on Memory Research*, L.G. Nilsson (ed.), Lawrence Erlbaum, Hillsdale, NJ, 1979, pp. 115-138.

Klein, D. *Decision-Analytic Intelligent Systems: Automated Explanation and Knowledge Acquisition*, Lawrence Erlbaum, Hillsdale, NJ, 1994.

Lehnert, W. *The Process of Question Answering*, Lawrence Erlbaum, Hillsdale, NJ, 1978.

Leonard-Barton, D. and Sviokla, J.J. "Putting Expert Systems to Work," *Harvard Business Review* (88:2), March-April 1988, pp. 91-98.

Mackie, J.L. *The Cement of the Universe: A Study of Causation*, Clarendon Press, Oxford, U.K., 1980.

Moore, J.D. and Swartout, W.R. "Explanation in Expert Systems: A Survey," ISI Research Report, RR-88-228, Information Sciences Institute, University of Southern California, Los Angeles, CA, 1988.

Moore, J.D. and Swartout, W.R. "A Reactive Approach to Explanation," in *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, MI, August 20-25, 1989.

Neches, R., Swartout, W.R., and Moore, J.D. "Enhanced Maintenance and Explanation of Expert Systems Through Explicit Models of Their Development," *IEEE Transactions on Software Engineering* (SE-11), November 1985, pp. 1337-1351.

Nettler, G. *Explanations*, New York, McGraw-Hill, 1970.

Nickerson, R.S., Perkins, D.M., and Smith, E.E. *The Teaching of Thinking*, Lawrence Erlbaum, Hillsdale, NJ, 1985.

Nisbett, R. and Wilson, T. "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* (84), 1977, pp. 231-259.

Norman, D.A. "Some Observations on Mental Models," in *Mental Models*, D. Gentner and A.L. Stevens (eds.), Lawrence Erlbaum, Hillsdale, NJ, 1983, pp. 7-14.

Quillici, A. "Arguing Over Plans," in *Proceedings of the AAAI Spring Symposium Series, Symposium: Argumentation and Belief*, Stanford, CA, March 26-28, 1991.

Regian, J.W. and Shute, V.J. (eds.). *Cognitive Approaches to Automated Instruction*, Lawrence Erlbaum, Hillsdale, NJ, 1992.

Slatter, P., Nomura, T., and Lunn, S. "A "Representation" for Manufacturing Sequencing Knowledge to Support Co-operative Problem Solving," in *Proceedings of Joint Ergonomics Society/ICL Conference on Human and Organizational Issues of Expert Systems*, Stratford Upon Avon, U.K., May 1988.

Stefik, M., Aikins, J., Balzer, R., Benoit, J., Birnbaum, L., Hayes-Roth, F., and Sacerdoti, E. "The Organization of Expert Systems, a Tutorial," *Artificial Intelligence* (18), March 1982, pp. 135-173.

Suermondt, H.J. *Explanation in Bayesion Belief Networks*, unpublished Ph.D. dissertation, Program in Medical Information Sciences, Stanford University, Stanford, CA, April 1992.

Swartout, W.R. "XPLAIN: A System for Creating and Explaining Expert Consulting Programs," *Artificial Intelligence* (21:3), September 1983, pp. 285-325.

Teach, R.L. and Shortliffe, E.H. "An Analysis of Physicians' Attitudes," *Computers in Biomedical Research* (14), December 1981, pp. 542-558.

Todd, P. and Benbasat, I. "Process Tracing Methods in Decision Support Systems Research: Exploring the Black Box," *MIS Quarterly* (11:4), December 1987, pp. 493-512.

Torasso, P. and Console, L. *Diagnostic Problem Solving*, Van Nostrand Reinhold, New York, 1989.

Toulmin, S. *The Use of Argument*, Cambridge University Press, Cambridge, UK. 1958.

Toulmin, S., Rieke, R., and Janik, A. *An Introduction to Reasoning* (2nd ed.), Macmillan, New York, 1984.

Wason, P.C. and Evans, J.St.B.T. "Dual Processes in Reasoning?" *Cognition* (3:2), June 1975, pp. 141-154.

Wexelblat, R.L. "On Interface Requirements for Expert Systems," *AI Magazine* (10:3), Fall 1989, pp. 66-78.

Wick, M.R. "Expert System Explanation in Retrospect: A Case Study in the Evolution of Expert System Explanation," *The Journal of Systems and Software* (19:2), October 1992, pp. 159-169.

Wick, M.R. and Thompson, W.B. "Reconstructive Expert System Explanation," *Artificial Intelligence* (54:1-2), March 1992, pp. 33-70.

Zahedi, F. *Intelligent Systems for Business: Expert Systems with Neural Networks,* Wadsworth, Belmont, CA, 1993.

About the Authors

L. Richard Ye is associate professor of the Accounting and MIS Department at the California State University, Northridge. He holds a Ph.D. in MIS from the University of Minnesota. His current research interests include knowledge-based tutoring systems and human-computer interaction.

Paul E. Johnson is Carlson Professor of Decision Sciences and a faculty member in the Department of Information and Decision Sciences in the Carlson School of Management at the University of Minnesota. He is a fellow of the American Psychological Association and The Psychological Science Society as well as a member of The Cognitive Science Society, Institute for Operations Research and the Management Sciences, The Decision Science Institute, and the Association for Computing Machinery. His current interests include sources of variability in decision making and knowledge work and the misrepresentation of information in organizations and technology.