Internal Control Evaluation: A Computational Model of the Review Process

Rayman D. Meservy  
*Brigham Young University, rayman_Meservy@byu.edu*

Andrew D. Bailey Jr

Paul E. Johnson

Follow this and additional works at: https://scholarsarchive.byu.edu/facpub

Part of the Accounting Commons, and the Management Information Systems Commons

**BYU ScholarsArchive Citation**

https://scholarsarchive.byu.edu/facpub/3239

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Internal Control Evaluation: A Computational Model of the Review Process

Rayman D. Meservy, Andrew D. Bailey, Jr. and Paul E. Johnson

ABSTRACT

This study investigated the strategies by which experienced auditors evaluate systems of internal accounting controls. The research method included: (1) observations, using concurrent protocols, of a small sample of practicing auditors performing the internal control evaluation task; (2) extensive interviews with one of the practicing auditors; (3) formalization of auditor processes as a computational model; and (4) validation of the model. The simulation model was implemented as an expert system and tuned to one auditor. The model output consists of a trace of the model processing including: (1) recommendations for specific controls to be compliance tested; and (2) a list of management letter issues. The model was tested using new cases and cross-validated against the performance of additional auditors performing the task. A separate group of auditors rated the performance of the computational model and the additional auditors performing the validation task.

RESEARCHERS, as well as teachers and practitioners, are interested in the processes that auditors use when making judgments and decisions. While expertise in such fields as medicine, physics, and chess has been studied intensely during the past 20 years (e.g., Klenmuntz [1968], Einhorn [1980], Elstein et al. [1978], and Johnson et al. [1982]), comparatively little research has been done in the fields of business and management.

The study and evaluation of internal accounting controls is a problem involving the expertise of well-trained auditors and is a requirement of every audit performed by CPAs. This research examines auditor expertise as it relates to the audit task of evaluating internal accounting controls.

This paper first focuses on expertise and the auditing task of reviewing and evaluating internal controls. The applicability of Decision Support Systems (DSS), Artificial Intelligence (AI), and Expert Systems (ES) to auditing tasks is then discussed. The current research covers the complete process of developing an expert system, but focuses most heavily on the cognitive science issues. The resulting computational model is intended to emulate the expert, i.e., the focus of our work is on understanding the expert's decision processes and not on building a better problem solver. The various methods of assessing the auditor's judgment processes used to build the knowledge base are presented. An expert system shell was used to build a computational model of the resulting process. Validation of the resulting computational model using several approaches is discussed.

EXTANT AUDITING DSS/AI/ES/ APPLICATIONS

There are many views about what should be included as DSSs and their relationship to AI and ES. While no definitive statements of definitions and characteristics for DSS exist, in this paper we use the term Decision Support Systems or DSS in a broad sense to refer to any interactive computer application that helps a decision maker by providing access to large data banks or by implementing a decision model, or both.

Artificial Intelligence (AI) has been defined by Barr and Feigenbaum [1981] as "the part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate with intelligence in human behavior — understanding language, learning, reasoning, solving problems, and so on."

Expert systems have been defined by Stefik et al. [1982] as "problem-solving programs that solve substantial problems generally conceded as being difficult and requiring expertise. They are called knowledge-based because their performance depends critically on the use of facts and heuristics used by experts." Feigenbaum [1978] says, "we must hypothesize from our experience to-date that the problem solving power exhibited in an intelligent agent's performance is primarily a consequence of the specialist's knowledge employed by the agent, and only very secondarily related to the generality and power of the inference method employed. Our agents must be knowledge rich, even if they are methods poor." Thus, expert systems attempt to capture specific knowledge from an acknowledged expert concerning a specific problem domain and to replicate the decision inference process used by this expert. In this paper, expert systems are used in...
A narrow sense to refer to interactive computer applications that help a decision maker by simulating the specific knowledge and inference processes used by experts in their limited domain of expertise.

The development of organized methods of collecting, organizing, and scoring data collected on audits was a long-established practice before the computer. However, the advent of inexpensive computing has opened new horizons. As a result, many of the manual aids previously used by auditors have been or are being converted to computer support systems. These include the systems prepared by the major public accounting firms, as well as private software houses catering to the public and internal auditing communities. Most of these systems concentrate on data collection from client systems, organization analysis based on descriptive and normative models, statistical sampling methods, and workpaper control. In most cases, they lack the characteristics necessary to be a DSS as defined earlier. The term Decision Aid (DA) might better describe these systems.

In the last several years, auditing firms and academics have become very active in extending the capabilities of Decision Aids by adding menu-driven, interactive components to support the various audit functions. Peat, Marwick, Mitchell & Co.'s SEACAS, the computerized version of SEADOC [1980], and Arthur Young & Co.'s recently announced ASQ, Auditing Smarter and Quicker, are good examples of this trend. Balachandran and Zolters [1981] provide an example of possible future directions in this area.

Non-Expert DSS are part of a new wave of technological innovation in the field of auditing. Auditors accept both the technology and the potential of such systems. Within the next several years many such systems will come on line.

In the last few years, numerous projects have been started that attempt to apply expert systems techniques to the development of DSS for auditors. These include the development of TAXMAN [McCarty, 1977] and TAX ADVISOR [Michaelson, 1982] to provide legal tax advice. TAX ADVISOR is based on the EMYCIN [van Melle et al., 1981] shell. AUDITOR [Dungan and Chandler, 1980] is an expert system for the evaluation of the adequacy of the client's allowance for bad debts. Other systems include those by Braun and Chandler [1982] to aid auditors in Analytic Review, by Hansen and Messier [1982], a model for evaluating EDP Controls, and by Bailey et al. [1985] for internal control evaluation.

There is currently an acceleration of research and development activities for DSS in auditing. The above brief discussion of systems in auditing represents only a small portion of the current interest and activity. Further, this discussion ignores the many studies necessary to the final implementation of such systems. Studies such as that of Biggs [1979] on the going concern judgment process, Mustcher [1984] on the issues of "subject" to opinions, and many others facilitate the future development of DSS for auditors.

**EXPERTISE AND THE AUDIT TASK**

Expertise has been defined as "knowledge about a particular domain, understanding of domain problems, and skill about solving some of these problems" [Hayes-Roth et al., 1983]. Davis has proposed that the nature of expertise includes the ability to: (1) solve the problem; (2) explain the result; (3) learn; (4) restructure knowledge; (5) break rules; (6) determine relevance; and (7) degrade gracefully [Davis, 1982].

An expert's knowledge consists of both public and private information. Public knowledge includes the facts, theories, and definitions as found in the texts and journals referenced by those studying in the domain. However, experts also possess private information that is not found in any of the public literature. Much of this private knowledge is in the form of rules of thumb which we will refer to as heuristics. Heuristics allow experts to "make educated guesses when necessary, to recognize promising approaches to problems, and to deal effectively with errors or incomplete data" [Hayes-Roth, 1983]. Knowledge engineers, who are concerned with the acquisition and representation of knowledge, concentrate much of their effort on the elucidation and reproduction of such "rules of expertise." Human expertise in problem solving is largely the recognition and use of heuristics. Feigenbaum emphasizes that "experience has taught us that much of this knowledge is private to the expert, not because he is unwilling to share publicly how he performs, but because he is unable. He knows more than he is aware of knowing" [Feigenbaum, 1978].

We consider auditors to be experts in performing certain tasks. The objective of this study was to determine the processes that auditors use in a specific audit task, formalize and implement those processes as a computational model, and then test the model.

**AUDITING INTERNAL ACCOUNTING CONTROLS**

The American Institute of Certified Public Accountants (AICPA) [1979] defines internal controls as "the plan of organization and all the coordinate methods and measures adopted within a business to safeguard its assets, check the accuracy and reliability of its accounting data, promote operational efficiency, and encourage adherence to prescribed managerial policies . . ." (Section 320.09) [AICPA, 1972]. Mair, Wood, and Davis, in their book Computer Control and Audit, make the statement: "Controls act upon things that can go wrong which, in turn, leads to the reduction of exposure" [Mair et al., 1978]. Although all public accounting firms evaluate controls and general guidelines have been suggested by several different researchers [Mautz & Winjum, 1981], auditors still have difficulty evaluating the quality of internal control systems.

The general objective in studying accounting internal controls is to satisfy the auditor's second standard of field work: "There is to be a proper study and evaluation of the existing internal control as a basis for reliance thereon and for the determination of the resultant extent of the tests to which auditing procedures are to be restricted" [AICPA, 1972]. Thus, the primary purpose is to determine whether the accounting controls are strong enough to be relied upon to produce reliable financial information. If the internal controls are determined to be strong, then the scope of other audit procedures may be more restricted than when the internal controls are determined to be weak. A second objective is to provide the auditor with a basis for constructive suggestions on how to improve the client's internal accounting controls [AICPA, 1972].

**CONCEPTUALIZATION OF THE TASK**

The strengths and weaknesses of an internal accounting control system are evaluated by determining control objectives, identifying controls and faults from a description of the system, and then combining the controls and faults into an overall evaluation of the sufficiency with which each control objective has been met. (See Figures 1a and 1b.)

Controls and faults are conceptual objects that can be identified by particular recognizable patterns of data embedded within the statements describing the accounting information system. Controls prevent, correct, or detect system expo-
These associations permit chains of deductive reasoning to be constructed connecting the structural design of the accounting system and the likely functional strengths and weaknesses of the system. This is typically called prototypic or recognition-based reasoning.

The result of the internal control evaluation task consists of: (1) a suggested list of controls for the compliance testing phase; and (2) a list of control weaknesses. The list of control weaknesses indicates significant problems discovered during the evaluation.

*Not necessarily a tree (overlap)

A Control Objective (CO) is a high-level abstraction of what preventive devices should be used. Control Objectives may be broken down into Control Subgoals (COS). Control Procedures (CP) look for an implementation (a matching pattern) and are the lowest level of control knowledge.
process and the resulting exposures that could occur. The auditor uses this list in establishing subsequent compensating audit steps and for interaction with management.  

The specific weaknesses identified are combined with the controls to determine sufficiency for each control objective, which in turn results in the auditor expanding some of the substantive tests performed later in the audit.

Subjects  
Practicing CPAs in middle management of a large auditing firm served as subjects in the project. Managers in the local office of Peat, Marwick, Mitchell & Co., an international public accounting firm, participated in both the model building and the cross-validation. The model was built largely with the help of one auditor, who also assisted in the model validation. Six other auditors assisted in validating the model. (See Figures 2a and 2b.) All the subjects were considered expert at the task by their superiors.

Task Difficulty  
There is no unique set of acceptable controls that is considered normative; rather, accounting systems can be configured using a wide variety of acceptable combinations of controls. As a result, the evaluation of internal accounting controls is a difficult task. In addition, experts are generally unable to describe each step in the evaluation process, making it difficult to train new experts. The process is normally taught by providing post rationalizations of the analysis process and by having novices solve numerous case problems. Behavioral studies clearly indicate that the post explanations often do not match the process actually followed by the expert [Nesbett & Wilson, 1977]. The evaluation of internal controls has typically relied on such decision support aids as flowcharts and questionnaires concerning the client's accounting systems. The weaknesses inherent in these traditional techniques have been recognized by accountants and accounting firms for some time. The last several years have witnessed the introduction of a number of new approaches intended to regularize the data collection and evaluation process. Previously referenced research suggests a greater likelihood of success on the collection side than on the evaluation side. Nevertheless, public accounting firms hope that these new approaches will lead to greater consensus in evaluating internal controls among their field auditors.

The Task  
Subjects were asked to perform the task of reviewing and evaluating actual audit workpapers used in audits by Peat, Marwick, Mitchell & Co. The task was limited to the purchases, payables, and cash distribution transaction cycle as found in manufacturing, wholesale, and retail industries. Specifically, the subjects were asked: (1) to identify control weaknesses or problems, and (2) to make recommendations for specific controls to be compliance tested. The workpapers consisted of SEADOC [Peat, Marwick, Mitchell & Co., 1980] flowcharts and descriptions of client firms, as prepared by the in-charge auditor. The task was also veridical in that our subjects were managers, who ordinarily do not prepare such documents, but generally do review and evaluate internal controls based on such SEADOC documents prepared by others. Although our study places the manager in a normal and familiar role, the subject would normally be involved in an ongoing relationship with the client. For study purposes, our subjects evaluating internal controls were provided with documentation for clients with an ongoing relationship with the firm but in which no such relationship existed for them personally.

FIGURE 2a  
Plan of Work
METHODOLOGY:
DECISION PROCESSES

The methods discussed below were employed to help understand the expert auditor's decision processes in analyzing internal controls and to validate that understanding. For this study our focus was on understanding, not on building a better problem solver. The methodology is summarized in two phases, discovery and verification. (See Figures 2a, 2b, and 3 and discussions below.) The notable difference in using these tools to understand the expert process as opposed to a problem solver orientation is the extensive amount of work required in the discovery phase. The discovery phase was composed of two steps, model development and model implementation and tuning. The objective of the discovery phase was the development of a theory of task performance for our expert.

FIGURE 2b
Plan of Work

Meservy, Bailey and Johnson

FIGURE 3
Research Overview

SIMPLIFIED REPRESENTATIONS OF RESEARCH APPROACHES

COMMON RESEARCH IN ACCOUNTING

DISCOVERY PHASE

VERIFICATION PHASE

POSTULATE A THEORY THAT CAN BE TESTED
(OFTEN SUGGESTED THROUGH LITERATURE REVIEW AND DISCUSSIONS)

EXPERIMENTS

ANALYSIS

MODEL BUILDING RESEARCH

DISCOVERY PHASE

VERIFICATION PHASE

MODEL BUILDING

MODEL TUNING

MODEL EVALUATIONS

EXPERIMENTS

ANALYSIS

KNOWLEDGE ACQUISITION

KNOWLEDGE REPRESENTATION

PROBLEM REPRES.

IMPLEMENTATION

(The model must be verified through experimental research similar to any other theory)
THE DISCOVERY PHASE: MODEL DEVELOPMENT

Preliminary representations of expertise, including key concepts and relationships, were developed from interviews and experimental task data using experts both as collaborators and as subject-informants. These descriptions consisted of problem-solving steps and heuristics that represent auditor judgment in: (a) identifying internal accounting control objectives; (b) identifying controls and faults in the accounting system; and (c) evaluating the system controls, weaknesses, and sufficiency of documentation.

KNOWLEDGE ACQUISITION

Research at the University of Minnesota by Johnson [1983b] has isolated general principles of eliciting expert knowledge. This research adapted these approaches and used them to elicit knowledge from our expert auditors. The multi-method approach is summarized in three parts: observational, descriptive, and intuitive methods.

Observational Methods for Knowledge Acquisition

The observational approach adopted for this study consisted of collecting "thinking-aloud protocols" which were used to identify problem-solving mechanisms being used by experts. These protocols provided information about the organization of the expert's knowledge base, the knowledge it contains, and the control structures used to apply that knowledge. A major difficulty with observational methods is, however, that the very techniques used to determine the reasoning processes may distort those processes [Nisbett and Wilson, 1977; Johnson, 1983a].

Previous experience in cognitive modeling indicates that the decision-making behavior of experts cannot be adequately understood by analyzing problem-solving processes alone. Several higher levels of analysis are important to the modeling of expertise. In this study, auditor decision processes were analyzed in terms of: (1) episodes, (2) views or frames of reference, and (3) problem-solving processes. The categories range from the more general decision- and goal-seeking strategy of the subjects, to very specific types of processes which allow the auditor to progress from one state of knowledge about the audit to another.

Macro analysis of the protocols first identified the higher-order episodes used by the auditor in the decision-making process. Episodes involve proposing tentative goals and/or subgoals appropriate for the task, and then doing the analysis necessary to either substantiate or disprove the goals. The macro goal categories for the task are: (1) Decide on the likely inherent risk category of client and the most appropriate overall audit approach. Such firm categorization is based on understanding the macro environment within which the firm operates, firm size, growth, industry, and general management characteristics. (2) Decide if there are significant processing controls which can be relied upon and the appropriate compliance tests. (3) Choose which boundary controls for accounts payable to rely upon and the appropriate compliance tests. (4) Choose which controls over disbursements can be relied upon and the appropriate compliance tests. (5) Evaluate the effectiveness of general computer controls and other appropriate firm environment factors. (6) Draw conclusions on overall audit approach, controls to be relied upon, and appropriate audit procedures.

The episodic categories are presented in Figure 4 by dividing and thus standardizing each transcribed auditor protocol into 100 equal units and then classifying each unit in the appropriate category. Figure 4 presents

* Figure 4 is an example. A similar figure was prepared for each model and subject protocol on each case.
the flow of the protocol from left to right. Each asterisk represents a unit of time during the protocol. As can be seen from Figure 4, the episodes are sustained goal-seeking categories. However, these episodic goals are somewhat interdependent and must sometimes be suspended until other episodic goals have been reached. Consider the specific processing controls dependent upon the computer as an example. Before conclusions about the specific tests of computer processing controls can be reached, the auditor must first determine the effectiveness of the general computer controls and the extent to which they can be relied upon. Figure 4 allows us to relate how the episodes unfold and their relationship with other activities scored in the protocols.

Auditors also appear to have major frames of reference which allow them to organize and evaluate various aspects of the data cues. The views or frames of reference identified from the auditor protocols were: (1) processing, (2) segregation of duties, (3) electronic data processing (EDP) factors, and (4) the adequacy of the working papers.

The data cues are analyzed by the auditor from each of the above perspectives, then integrated before deciding which controls may be effective, which controls should be compliance tested, what type of compliance tests should be performed, and how large a sample would be sufficient. Our primary subject reported “flipping from one view to another” as he analyzed the data. He also reported significant interactions among these various views of the data. This behavior is evident in his protocols. Figure 4 illustrates, by reading down the Figure for a given time unit, that there may be mention of more than one reference frame in any of the protocol units. Also note that the primary frame of reference is processing, which enjoys the most sustained attention. It should be noted that the working papers themselves are organized primarily from a processing frame of reference.

Decision operators provide the links between individual knowledge states at the micro level. Scoring the protocol for problem-solving processes allows the determination of which operators are used by auditors in their evaluation of internal controls. Figure 4 also provides a list of cognitive operators used during scoring.

Using observational methods, the macro analysis of the protocol provided the episodes and views around which the model was built. The micro analysis, scoring of the decision processes, helped most in the model-building process when scored within the bounds of the specific goals and objectives that direct the search and confirmation processes.

The initial protocol phase of research was followed by a refinement phase in which experts were asked to comment on the episodes and views developed from the protocols and were queried for the appropriate decision rules.

Descriptive Methods for Knowledge Acquisition

Descriptive methods of assessing expertise are used to formalize portions of the expert’s knowledge by transforming that knowledge into an explicit representation, often in the form of a heuristic or rule. Most existing methods of assessing expertise rely heavily on descriptive methods. The major limitation of the descriptive method is that the more competent an expert becomes, the less able he/she is to describe his/her problem-solving knowledge [Johnson, 1983a]. One method of accomplishing this end is through interviews with auditors in which they attempt to characterize their knowledge and skill in the given task situation. The researcher then formulates the response as a production rule. The type of questions asked in this process included the following:

What objectives h do you think of when you see problem data about c?
What evidence (controls or sub-objectives) makes you more certain that objective h is satisfied?
What makes you conclude that objective h cannot be satisfied? [see Moen, 1984]

A second, more direct method is through the creation of a precise “language” in which the expert is asked to describe his/her expertise. The structure of the language is determined by the architecture of the knowledge base. In this case the auditor described his/her lines of reasoning by means of production rules, i.e., if-then statements.

Intuitive Methods of Knowledge Acquisition

Intuitive methods for capturing knowledge exist in two forms. In one case, a knowledge researcher interacts with the auditor and the literature of the field in order to become familiar with its major problem-solving methods, which are then checked against the opinion of other auditors and eventually incorporated into the computer program. A second intuitive method of knowledge involves a researcher who is an expert in the area attempting to describe the basis for his/her own knowledge and skill. Intuition as a means of recovering one’s own knowledge is subjective and may not be adequate [Johnson, 1983a]. Intuitive methods were used by researchers in the model-building process only when necessary to guide the use of observational and descriptive methods in deriving the production rules.

In addition to the above methods, we referred to textbooks in an attempt to provide a base of reference for the rules [Johnson and Jaenicke, 1980]. Although these books did provide a reference point for basic internal control terminology and issues, only those rules elicited directly from the experts were used for model building.

This is consistent with our objective of studying auditor decision processes rather than building a new problem solver. In addition, subsequent discussions with the experts indicate that while the textbook rules are generally correct, they do not coincide with the auditor’s decision processes.

The products of the knowledge acquisition portion of the research included a representation of auditor expertise in: (a) the identification of specific internal accounting control objectives; (b) the evaluation and review processes, identifying which controls should be further tested for reliance thereon; and (c) the type of processes used in recognizing controls and weaknesses.

Knowledge Representation Structures

Production rules were chosen to represent the expert knowledge of an auditor. No claim is made for the universal applicability of such a representation; rather, we rely on the demonstrated utility of rule-based systems in representing problems with characteristics similar to those encountered in analyzing internal controls. Rule-based representations (also referred to as Situation — Action rules or If — Then rules) allow easy modification and explanation, both considered essential for building and then tuning such a computational model. Basically, each rule captures a “chunk” of the domain knowledge, meaningful in and of itself to the domain specialist [Feigenbaum, 1978].

The rules are normally associated with “lines of reasoning” and “episodes” that are comprehensible to the domain expert [Feigenbaum, 1978]. Lines of reasoning incorporate the system analysis method and frames of reference employed by the subject. Episodes involve proposing a tentative goal and/or subgoal (hypotheses) and then trying to either substantiate or disprove that goal. Such a generate-and-test frame-
work has been identified in behavioral studies by Biggs and Mock [1983] and others when studying audit settings. The formulation and maintenance of lines of reasoning and episodes often requires the integration of many different “chunks” of knowledge. It is important that the computational model be able to explain its use of knowledge to the domain expert for both refinement and validation purposes. The computational theory, including the choice of hierarchical structures and the partitioning of the knowledge base into sub-components, is discussed elsewhere [Meservy, 1985].

MODEL IMPLEMENTATION AND TUNING

The preliminary model was implemented as a computational model by using the Galen modeling tool developed at the University of Minnesota [Thompson et al., 1983]. Galen’s architecture reflects its development in modeling problem-solving processes. Galen’s inference engine has the ability to partition the knowledge base, to search for a hierarchical set of goals, to apply forward and backward chaining, and to build and interact with LISP representation of the audit working papers. While demonstrating the generality of Galen, the adaptation of an already proven tool enhanced development productivity. The rules developed above and the Galen inference engine were combined to form the computational model.

The Model

The implemented computational model with its associated rules included approximately 300 If statements. Note that some of the If statements in Galen may represent six or more production rules in other expert systems. The actual implementation of the Galen-based system involved organizing the formalized knowledge into rule teams, along the line of reasoning categories identified from the protocol analysis.

Of the approximately 300 rules in the model, normally only ten to 25 percent of the rules fire in any given evaluation. These percentages held for the three model traces which were used in the cross-validation test.

Model Tuning

After the acquired knowledge was mapped into Galen’s representational framework, the system was “tuned.” Tuning involved running several prototype internal accounting control information systems descriptions through the evaluation process and, in collaboration with the expert, checking the lines of reasoning and episodes for reasonableness and making adjustments in the rules. Some of the more important aspects of expertise were discovered in this process, i.e., knowing when to discontinue the current line of reasoning or begin another. Aspects of expertise incorporated during the tuning phase included rules about the use of other rules, known as meta rules, which were used by the expert in guiding his/her thinking about each case.

THE VERIFICATION PHASE

It is important in simulation studies that the model not only be built and tuned, but also be tested and verified. Typically, researchers have not been able to formulate single critical experiments which validate such models. Furthermore, due to the small sample size, statistical evaluations of experimental results are generally not available and researchers are constrained to rely on graphical techniques. The approach used in this study included several empirical tests, each of which addressed different types of data and different aspects of the model’s behavior.

Verification experiments focused upon comparisons between the control evaluation strategies of the model and the processes employed by human auditors. The framework used to evaluate the model’s performance had two major features: (1) tests of the quality of model processes and of cue usage, and (2) tests of sufficiency or adequacy of model outcomes.

Because accounting information systems differ, each represents a new challenge to the model. Therefore, if the model is to be tested, the cases chosen should not represent recombinations of portions of previous cases employed, but totally new cases representing a wide range in risk, reliance, and work paper documentation. The three cases chosen represent such a range within the confines of the model limitations, discussed earlier. Copies of actual work papers obtained from the firm were used for the three cases.

The computational model was fine tuned, as previously noted, around the expertise of one individual auditor (the primary subject, SI). The computational model was initially validated against this individual using the three cases, and then it was cross-validated against three additional expert auditors (subjects 4, 5, and 6).

For each case, the primary subject and the three other auditors were asked to read aloud the pertinent data and give “thinking-aloud” protocols while reviewing and evaluating the actual work papers prepared by an in-charge auditor. As part of the task, the subjects were asked to list: (1) recommendations for specific controls to be compliance tested, and (2) weaknesses identified from the system description. At the conclusion of the session, subjects were asked to fill out a participant background questionnaire. The complete problem-solving session for each subject was tape recorded and transcribed. The computational model also received each case and made similar evaluations. In addition to making recommendations for specific controls to be compliance tested and weaknesses, the model provided a trace of all data analyzed and rules fired.

Tests of Quality of Model Processes and Cue Usage

The first type of analysis involves establishing the quality of the evaluation processes employed by the model and the model’s response to “critical cues.” To establish quality of evaluation processes, the inferences made must not only be “legal,” but must be the type of inferences that experts would make. Determining that a sample of model behavior constitutes adequate auditing behavior is not a simple matter. In games, such as chess, it is fairly easy to determine if the model is performing the requisite behavior because the rules used to determine whether a given move is “legal” are well-defined. By contrast, in environments such as internal control evaluation, it is not clear what constitutes a “legal move.” Furthermore, as in a game, though all reasoning steps are explainable by logic or rules, some “lines of reasoning” (smaller sets of steps or moves) must be made according to a criterion of quality in order for the task to be done well. In medical diagnosis, for example, typically there is a small set of cues that, if interpreted properly, would lead to a correct diagnosis. Experts may differ in their interpretations of other pieces of information, but they tend to agree more on the interpretation of these critical cues and the use of additional cues to mediate between competing hypotheses [Johnson et al. 1982].

The evaluation framework adopted in the present study requires that the model’s rules for performing the evaluation based on heuristics of expert behavior. The judgment as to whether the model is performing the task is then based on comparisons between the specific acts of model behavior on the task and the behavior of expert auditors. These comparisons focus upon: (1) the identification and use of specific
goals and objectives which direct the search and confirmation processes, and (2) the processes linking problem states to one another. The quality of the model's "thinking" process was evaluated by transcribing, scoring, and analyzing each model problem-solving trace against the problem-solving trace of expert auditors as revealed in their verbal protocols. Protocols provide a depth of understanding about judgment and decision making unavailable using other methods. However, as in other methods, the data must be reduced to a structured, objective image of the processes that auditors are using. The analysis is developed by synthesizing the results of two analytical methods: a top-down, global analysis and a bottom-up, problem state/process analysis (Bouwman, 1978).

The top-down analysis identifies single problem-solving goals from the protocols. Proposed categories are developed through functional analysis of the review task, formalized descriptions of evaluation processes generally, and model fragments found in the auditing literature; an example is the representation or mental picture of the segregation of duties within the purchasing/cash distribution function.

The bottom-up analysis focused on problem-solving states, the basic set of facts, concepts and hypotheses generated by the subject, and an associated set of reasoning processes (reading, planning, evaluating, searching for information, etc.). Figure 4 contains a list of hypothesized categories for the internal control evaluation and review task. Scoring protocols for reasoning processes permits an understanding of how auditors use past knowledge, generate new knowledge, and the type of processes that link this knowledge together. The analysis provides a "picture" of the path taken by the auditor to perform the task. It differs from the top-down analysis primarily by being more elementary and in finer detail [Malone, 1984].

A third, more general type of analysis, "lines of reasoning" and search strategy, was then determined by analyzing the "protocol graphs" and the sequence of problem-solving goals. Lines of reasoning involve the methods employed by the subjects to solve the task; an example would be the focus on receipt of goods rather than on commitments.

Tests of Adequacy of Model Outcomes

The second type of analysis, for sufficiency or adequacy of model outcomes, was performed to establish that the computational model could identify and evaluate internal accounting controls. The 15 outcomes from the three cases examined, including the lists described above, were re-typed, randomly sorted, and renumbered for each case to hide the original identity and remove inter-case identity. These outcomes were then given along with the cases to three other expert auditors (subjects 7, 8, and 9) to judge. This process is referred to here as a peer review.

The peer review required each subject acting as a reviewer to first read case A and judge the five solutions produced by the model and four subjects. The peer reviewer was then asked to judge each solution for completeness, effectiveness, and agreement with results. The possible range of each evaluation was from one through seven, where one indicated a minimal rating and seven represented extremely complete, extremely effective, or substantial agreement with conclusions. After rating all five solutions, the peer reviewer was instructed to divide the solutions into two groups based on perceived similarity. Between one and four solutions were allowed in each group. Next, the reviewer was instructed to divide the solutions into three groups, again based on similarity. This time between one and three solutions formed a group. At this point the peer evaluation was complete, and the subject proceeded with the next case. These ratings were later analyzed to determine how well the model performed in relation to the person it was modeled after and in relation to the other auditors.

Validity and Consistency

After each transcribed protocol and trace was scored, a second scorer or rater, trained in the rules for coding these protocols, again scored portions of the transcribed protocols from the model and for all subjects and cases. The coded protocols from each rater were compared, and the proportion of agreement between the lists developed by each rater for the protocols rated were computed. Cohen's K [Cohen, 1960], an inter-rater reliability coefficient, was employed to adjust for agreement due to chance. Traces generated by the computational model were scored by the same methods used to score subject protocols. The inter-rater reliability analysis, after being adjusted for agreement due to chance, ranged from 65.3 percent to 92.6 percent, with all but three falling in the .7 to .9 range.

RESULTS

This section examines the results of several empirical tests addressing different types of data and aspects of model behavior. The analysis of the experimental data is presented in two parts. The first part, Quality of Model Processes and Cue Usage, examines in detail the scored protocols and model traces, establishing the quality of the evaluation process employed by the model. In the second part, Adequacy of Model Outcomes, peer reviews of the model's and subjects' internal control evaluations are presented and discussed.

QUALITY OF MODEL PROCESSES AND CUE USAGE

The quality of evaluation processes used by the model is established by determining that the inferences made by the model are not only all "legal," but of the type experts make. The analytical approach used synthesizes the results of two analytical methods: (1) a top-down, global analysis and (2) a bottom-up, knowledge state/processes analysis. This section analyzes: (1) the hypotheses generated, (2) the processes identified, (3) the cue usage or the order and data attended to, and (4) the lines of reasoning employed. The last two are closely related in this study and will be discussed together.

Hypothesis Generation

The major hypotheses generated by the computational model were compared with the major hypotheses generated by expert auditors. The complete set of hypotheses and tentative conclusions employed by auditors in their review is so numerous that in this study we have limited ourselves to the major hypotheses generated. Whereas in medical diagnosis the physician may think of major diseases or disease categories, the auditor thinks in terms of controls, weaknesses, and problems, as well as major objectives.

Questions of interest were:

1. Are the hypotheses generated by the computational model found in the protocols of the primary auditor?
2. Are the hypotheses generated by the computational model found in the protocols of other auditors from the same firm?

Figure 5 is a Hypotheses Graph of the hypotheses generated by the subjects and the model. The graph is organized and presented by case and within each case by model and subject. Only those hypotheses which were employed during the experiment are listed on the left side of the graph. The ""+"" represents a positive decision, a ""-"" represents a weakness, problem, or question, and a ""?"" represents an assumption or basic issue about the hypothesis
category. In the context of the auditor review where team decisions are involved, it is often more important that the hypothesis has been effectively dealt with than the "+" or "-" determined. Also, because "?'s" may represent important issues for the auditor that need to be resolved, no great significance should be attached to the distinction afforded by the question mark.

It is interesting to note that in the three cases, there is only one hypothesis that was used by the model that was not used by one of the expert auditors, and that use was in the form of a possible weakness: whether a blind copy of the purchase order (PO) had been used in the receiving function. Note also the similarity between the model and the primary subject (S1 from whom the model was developed) as compared to the other subjects involved.

In case A, the model included each of the hypotheses identified by S1. Additionally, the model identified three possible management letter discussion issues not discussed by S1 in case A, but that had been discussed in earlier “tuning” cases. At the bottom of the graph is found the total number of hypotheses the subject and model had in common, the total number of hypotheses the subject included, and the agreement between the two represented by the ratio of common hypotheses to subject hypotheses. Case A represents the lowest overall average agreement between the model and the subjects as a whole, i.e., approximately 60 percent compared to 75.8 percent average agreement for case B and 81 percent average agreement for case C. While not wholly comparable, Ashton [1974] found auditor consensus to be about 70 percent, and Ashton and Brown [1980] found auditor consensus to be 67 percent on average in the internal control-related studies.

In case B, the model and S1 agree on all hypotheses, except that each generated one additional hypothesis. During the three test cases, this was the only instance in which S1 generated any hypothesis that the model did not generate. The model generated seven additional hypotheses (three in case A, one in B, and three in C) not generated by S1, five of which were management issue questions. This fact in conjunction with performance, as discussed earlier in the chapter, may indicate a bootstrapping effect, where the computer model is consistent and never forgets to analyze any possibilities. On the other hand, the model may not yet have some of the meta-rules allowing it to dismiss such questions.

Although case C was the most unique, there was less variation in the hypotheses generated here than in any other case. S4 was the only auditor that appeared to generate a substantially different set of hypotheses than the model. Note that the hypotheses sets generated by the model for each of the three cases are quite different from each other, at least compared to the similarity of the model and subjects within any given case.

### Problem-Solving Processes

The reasoning processes allow progression from one knowledge state to another. Figure 4 contains a list of hypothesized processing operators for the internal control evaluation and review task. The list includes operators for information acquisition and retrieval, planning approaches, analyzing the information acquired, and making decisions (taking action). The questions of interest are:

1. Are the reasoning processes produced by the model found in the protocols of the primary auditor?
2. Are the reasoning processes produced by the model found in the protocols of other auditors from the same firm?

The most detailed part of the analysis is the 15 protocol graphs, e.g., one graph is presented for the model and each subject’s evaluation of each case. Figure 4 is the Protocol Analysis Graph for Subject 1’s evaluation of case C. The graph presents a sequential picture of the goals and cues being attended to, the processes involved, and when the resulting decisions were made.

A vertical line through the Figure 4 graph for any unit of time provides information about what is occurring on several different levels at any instant and allows immediate access to that portion of the protocol. Given any of the auditor’s decisions, the graph allows a quick picture of
the general processes and the goals and cues that led to that decision. For example, using Figure 4, "Controls to Test," at the bottom of the Figure indicates that S1 decided to compliance-test the comparison of invoices, purchase orders, and receiving reports. By reading across "Written Results" in the Figure for "Controls to Test" over to the number "2," located in the 37th division, we observe that this occurred when S1 was 37 percent of the way through his protocol. Analyzing the 37th column of the graph under "Episodes," S1 had been examining the accounts payable boundary conditions. Further examining column 37 under "Views or Frames: Processing," commitments, receipt of goods, and accounts payable additions (invoices) have all been scored, meaning that all these categories were mentioned at that point in the protocol. Note that invoices had been mentioned frequently between columns 29 and 37 in the graph, whereas commitments and receipt of goods had been mentioned more often in earlier portions of the protocol. Certain lines of reasoning, as discussed later, may also be determined. Finally, under "Cognitive Processes," "Audit Decision" was scored as the process in use. Using the protocol graphs, direct access to that portion of the protocol that produced the second written result for "Controls to Test" is possible by reference to the percent column.

On the right side of Figure 4, the percentage occurrence of each operator is presented. These percentages, averaged for each subject and the model, have been included below as Figure 6. Note that the biggest difference in the model is the high percentage of read operators used. This difference, however, may only be the result of the scoring rules used on the trace and/or the way the model reads in the data, displaying it on the trace.

At the right of Figure 6 is a column labeled "outside." Whenever the model is outside the range of percentages found in the four expert protocol averages, the difference is marked in this column. Note that the model was outside the given range four times, but never by more than one percent. In addition, in those instances when the model was outside the given range, it was not more than one percent different from the primary subject (S1).

In general, the operators scored for in the model are quite consistent with the operators found in the expert protocols. Note the relatively high percentages of the read, evaluation, and audit decision operators scored. The next highest operators are the conjecture and assumption operators. Auditors evaluating internal accounting controls make numerous tentative evaluations or judgments, dealing with uncertainty by raising questions, building conjectures, posing assumptions, and proposing numerous tentative evaluations. The use of discrete assumed outcomes rather than probabilistic assessments of uncertainty is consistent with the findings of Biggs et al. [1985] and with Doyle's [1983] "reasoning by assumption." No probabilistic remark, such as "I'm 75 percent positive that invoices are adequately accounted for," was found in any of the protocols.

No explicit decision rule was ever mentioned by the experts while completing the internal accounting control evaluation. This result is consistent with the protocols from subjects S1, S2, and S3.

In summary, although some differences were found in the degree to which operators were used and/or graphed, the operators employed in the model were found in the primary subject's and the other experts' protocols.

**FIGURE 6**

**Reasoning Process Category Analysis**

*(in percents)*

<table>
<thead>
<tr>
<th>Subject</th>
<th>Avg.</th>
<th>Model S1</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I. Information Acquisition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td>34.9</td>
<td>46.3</td>
<td>34.3</td>
<td>47.3</td>
<td>39.3</td>
<td>18.7</td>
</tr>
<tr>
<td>Data Search</td>
<td>1.2</td>
<td>0.0</td>
<td>1.0</td>
<td>1.3</td>
<td>0.3</td>
<td>2.3</td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>0.8</td>
<td>0.0</td>
<td>0.3</td>
<td>0.0</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Total</td>
<td>37.0</td>
<td>46.3</td>
<td>35.7</td>
<td>48.7</td>
<td>41.7</td>
<td>22.0</td>
</tr>
<tr>
<td><strong>II. Plan</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plan</td>
<td>2.1</td>
<td>0.0</td>
<td>0.7</td>
<td>0.3</td>
<td>4.3</td>
<td>3.0</td>
</tr>
<tr>
<td>Total</td>
<td>2.1</td>
<td>0.0</td>
<td>0.7</td>
<td>0.3</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td><strong>III. Analytical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assumption</td>
<td>2.9</td>
<td>0.0</td>
<td>0.7</td>
<td>0.7</td>
<td>6.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Conjecture</td>
<td>6.8</td>
<td>5.7</td>
<td>8.0</td>
<td>2.0</td>
<td>6.7</td>
<td>10.3</td>
</tr>
<tr>
<td>Evaluation</td>
<td>33.0</td>
<td>32.7</td>
<td>41.0</td>
<td>27.7</td>
<td>26.3</td>
<td>37.0</td>
</tr>
<tr>
<td>Question</td>
<td>1.9</td>
<td>0.3</td>
<td>0.0</td>
<td>5.0</td>
<td>1.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Inference</td>
<td>1.0</td>
<td>2.0</td>
<td>1.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Total</td>
<td>45.3</td>
<td>40.7</td>
<td>51.3</td>
<td>36.0</td>
<td>41.3</td>
<td>52.7</td>
</tr>
<tr>
<td><strong>IV. Action</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generate Alternative</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Decision Rule</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Audit Decision</td>
<td>15.6</td>
<td>13.0</td>
<td>12.3</td>
<td>15.0</td>
<td>12.7</td>
<td>22.3</td>
</tr>
<tr>
<td>Total</td>
<td>15.6</td>
<td>13.0</td>
<td>12.3</td>
<td>15.0</td>
<td>12.7</td>
<td>22.3</td>
</tr>
<tr>
<td><strong>Final Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>
ments compared to S1 and the other experts. This may indicate too careful an analysis by the model, requiring too many rules, or it may indicate the commitment rules were improperly organized in the knowledge base.

The model also spent more time on segregation-of-duty issues in case A than any of the experts. Part of this difference may be that the auditors didn't specifically verbalize segregation issues. There appears to be considerable variance among the auditors, as well, in the commitment and segregation-of-duty categories.

For case B, the model spent more time on the accounts payable boundary conditions and less time on the processing controls than any of the expert auditors. Related to the boundary conditions, substantially more time was again spent examining commitments. Although similar conditions are noted in case C, a major difference is the time experts spent examining commitments.

With few exceptions, upon examining the relative percentages and the protocol graphs, it can be seen that the model is performing the task in a manner similar to auditors, and in the same order.
the information on the processing overview flowchart before concluding.

Regarding the episodic reasoning in case B, the model, S1, and S6 first examined general and processing controls, respectively, then spent more time on the a/p boundary controls; used less time on disbursement boundary controls; and then returned to use the majority of time on the processing controls. Of the other two auditors, S4 examined all of the information sequentially without returning, whereas S5 seemed to go back and forth several times.

While examining case C, the model, S1, and S6 used the same basic episodic strategy (Figure 4). S4 also employed the same strategy examining C as B, while S5 used a very different approach (examining the boundary controls first without even looking at the processing controls). The peers rated S5’s solution set the lowest on case C. The model, S1, and S6 (whose solutions were ranked the highest by the peer experts on all three cases) used the same basic episodic approaches.

An intermediate level of analysis, views or frames of reference, was also analyzed. Each of the views could likewise be thought of as a “line of reasoning,” along with the manner in which they interact.

An examination of the processing category in case A revealed that the model showed a definite sequential pattern between (first) commitments, (second) receipt of goods, and then (third) a/p additions, referring back to the other categories evaluated as needed. The primary subject, S1, also showed the same definite pattern. In contrast, S6 had a definite reasoning pattern of focusing on the receipt of goods, and then brought in commitments and additions as appropriate. The pattern for S4 was almost opposite that of the model and S1, while S5 produced a mixed approach more similar to S6.

For case B, the sequential evaluation of commitments, receipt of goods, and a/p additions were all evident when processing the a/p boundary conditions. Subjects 5 and 6 again had a definite focus on the receivables, and only included other information as appropriate. This is referred to as a mixed approach. The approach for S1 in case B appeared to be more mixed.

Examining case C revealed that the model and S1 (Figure 4) again show the stepped approach, with S4 stepped, but just the opposite. Subjects 5 and 6 appear to again have a more mixed approach.

Although auditors may examine the accounts payable processing stream before examining the disbursements processing stream, or first examine inputs to the processing system before examining summarization and posting types of controls, those patterns were not prominent in the limited protocols gathered. Other scoring methods may also be found that would highlight various levels and lines of reasoning not evident from existing scoring methods.

On the global level of analysis, each expert appeared to carefully and systematically examine all available cues. Although there was some directed search among the documents, it appeared to be minimal.

### ADEQUACY OF MODEL OUTCOMES

The peer evaluations were next analyzed to determine if there were substantial differences between the model and the primary subject (the modeler), or the other expert auditors. Good simulation models solve tasks in the same manner as experts, particularly similar to the expert modeled. The two questions of interest were:

1. Are there substantial differences between the outcomes of the primary auditor and the model as judged by experts?
2. Are there substantial differences between the outcomes of the model and the three auditors as judged by experts?
The three cases were analyzed by case and by subject in a blind review. Here we present only the combined results. The following figure presents ordinal relationships for the various solutions, as determined from the ratings of S7, S8, and S9. Ties are scored for the higher category.

For each of the three categories, completeness, effectiveness, and agreement with results, the model obtained five first-place ratings, three second-place ratings, and either a third-place or a fourth-place rating. The model had the most first-place ratings for the higher category. Six of the 13 first-place ratings for S6 were also ties. Six of the 11 first-place ratings for S1 were without ties. Examining the results for relatively poor performance, the outputs for the model, S1, and S6 were never judged the worst solutions.

Another type of analysis involves weighting and combining the ordinal rankings. Each first place was assigned five points, each second place received four points, each third place received three points, etc., and the results were added. The total for the model was 118, the total for both S1 and S6 was 110, and the points for S4 and S5 were 63 and 61, respectively. Caution must be used in interpreting these ordinal results. The peer reviewers were also asked to separate each case’s solution into two groups and then into three groups. The results for how often the model was placed in the same group with each of the auditors are indicated in Figures 14 and 15.

Analyzing the above two tables over all cases and all subjects, the model appears to be the most like the modeler (S1). It is interesting to note that for both the 2 and 3 groupings, the model has been placed in the same group with each subject at least once. Figures 13, 14, and 15 support the conclusions that: (1) there is no substantial difference between the outcomes of the primary auditor and the model as judged by experts; and (2) there is no substantial difference between the outcomes of the model and the three auditors as judged by experts. The model appears to be adequately performing the task of reviewing and evaluating internal accounting controls.

In conclusion, the model appears to simulate the processes of expert auditors, particularly the auditor after whom it was modeled. As indicated earlier in this section, a network of tests is necessary to validate a simulation model. Such a network of tests has been employed, but further validation is always appropriate.

**CONCLUSION AND EXTENSIONS**

In the process of building and testing the computational model, a theory was formulated of the kind that Allen Newell and Herbert Simon call “dynamic” [Newell and Simon, 1972]. Dynamic theories represent the initial state of psychological systems, such as human memory, in the form of symbol structures and the procedures for manipulating these structures (i.e., mental operations). Typical dynamic theories are formalized as computer algorithms, allowing the theory to be tested, which in turn allows future states of the system to be predicted [Johnson, 1983]. The dynamic theory encompassed in the simulation model presented here allowed the model to be tested. Furthermore, the theory can now be manipulated or perturbed, allowing new predictions to be examined.

Point predictions occur when exact predictions of how a given task will be done under variations in the knowledge base and alternative task conditions can be hypothesized [Johnson, 1983]. Suggestions for future research include systematically manipulating the model, creating and examining new hypotheses, and constructing specific task materials from which differential lines of reasoning, assumptions, and “garden paths” resulting in sub-optimal solutions can be predicted. Such predictions can be worked out either by analysis of expert thinking or by examining the behavior of the simulation model [Johnson and Hassebrook, 1982].

Another extension to the present model is to join a powerful internal control model and query language, TICOM [Bailey et al.,
1985] under the control of an expert system decision model. Such a union would allow the resulting system to query the user for appropriate modeling information to represent the client’s system as a TICOM model. Then, based on the questions of interest, decision rules could be employed to query the model, combining results for appropriate recommendations. The resulting system would allow both individual companies and auditors to make effective and efficient evaluations of internal accounting controls.¹

¹ Researchers interested in pursuing any of these issues should: (1) see Bailey et al. (1985) for information concerning software and manuals related to TICOM; (2) obtain a copy of Meservy (1985) from the Michigan Microfilm Library, as it contains the complete code for the model discussed in this paper; and (3) contact the Artificial Intelligence Research Center at the University of Minnesota to obtain information on the use of the Galen program.

REFERENCES

American Institute of Certified Public Accountants (AICPA), Statement on Auditing Standards (November 1972), AU Section 320.


Davis, R., “Expert Systems: Where Are We? And Where Do We Go From Here?”, The AI Magazine (Spring 1982), pp. 3-22.


