ELSEVIER

Cognitive Science 25 (2001) 355–392

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Detecting deception: adversarial problem solving in a low base-rate world

Paul E. Johnson[a,*], Stefano Grazioli[b], Karim Jamal[c], R. Glen Berryman[d]

[a]*Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, MN 55455, USA*

[b]*Department of Management Science and Information Systems, Red McCombs School of Business, The University of Texas at Austin, USA*

[c]*Department of Accounting and MIS, School of Business, University of Alberta, Edmonton, Alberta, Canada*

[d]*Department of Management and Economics, Carlson School of Management, University of Minnesota, Minneapolis, MN 55455, USA*

Abstract

The work presented here investigates the process by which one group of individuals solves the problem of detecting deceptions created by other agents. A field experiment was conducted in which twenty-four auditors (partners in international public accounting firms) were asked to review four cases describing real companies that, unknown to the auditors, had perpetrated financial frauds. While many of the auditors failed to detect the manipulations in the cases, a small number of auditors were consistently successful. Since the detection of frauds occurs infrequently in the work of a given auditor, we explain success by the application of powerful heuristics gained from experience with deceptions in everyday life. These heuristics implement a variation of Dennett's intentional stance strategy, which is based on interpreting detected inconsistencies in the light of the Deceiver's (i.e., management's) goals and possible actions. We explain failure to detect deception by means of perturbations (bugs) in the domain knowledge of accounting needed to apply these heuristics to the specific context of financial statement fraud. We test our theory by showing that a computational model of fraud detection that employs the proposed heuristics successfully detects frauds in the cases given to the auditors. We then modify the model by introducing perturbations based on the errors made by each of the auditors in the four cases. The resulting models account for 84 of the 96 observations (i.e., 24 auditors x four cases) in our data.

* Corresponding author. Tel.: +1-612-626-0320.

E-mail address: pjohnson@csom.umn.edu (P. Johnson).

0364-0213/01/$ – see front matter
PII: S0364-0213(01)00040-4

1. Introduction

Information in organizations and in society generally is often communicated in a context of conflict of interest, and with awareness of potential decision consequences (Feldman & March, 1988). Deception is one particularly powerful form of information manipulation that occurs when an opportunistic agent induces a misrepresentation that is designed to influence the behavior of another agent (Hyman, 1989; Johnson, Grazioli & Jamal, 1993; Russow, 1986). Deception poses a problem to individuals and organizations because those who fail to detect a deceiver's malicious manipulation take action on the basis of misrepresented information and, as a result, behave in ways that are more favorable to the deceiver than otherwise warranted. The problem of detecting deceptions is solved by identifying the elements of the manipulation and by taking corrective action.

The need to detect possible manipulations arises frequently since adversarial situations with *potential* for deception are found in virtually all domains of activity characterized by a conflict of interest between cognitive agents (e.g., in business, politics, warfare, games, intelligence and counterintelligence (Bowyer, 1982; Ekman, 1992; Mawby & Thompson, 1986). Yet, the occurrence of specific deceptions must be relatively rare, especially in domains where interactions are repeated and feedback is available: if a particular form of deceit is frequent, victims learn how to circumvent it (Akerlof, 1970; Kraut, 1980).

The relatively infrequent occurrence of specific deceptions, and the adversarial nature of deception as a social interaction, constrains the nature of the knowledge that cognitive agents develop to solve the detection problem (Thagard, 1992). Moreover, the element of intentionality introduced by the presence of an adversary sets the problem of detecting deception apart from the larger class of tasks where evidence is independent of the nature of the process employed to identify and analyze it (Johnson et al., 1981). Deception works because the Deceiver, 1) intentionally exploits the weaknesses of the Target's information processing, and 2) is aware of the Target's possible detection efforts, and acts so as to thwart them.

Under these conditions it is not surprising that naive Targets have high average rates of failure at detecting specific forms of deception (e.g., Ekman, 1992). What is surprising is the fact that the rate of failure often remains high, even when the Targets have been trained at the task within which deception occurs and are motivated by high stakes (e.g., Johnson, Grazioli & Jamal, 1992). For example, failure to detect forms of deception in the financial markets entails monetary losses amounting in some cases to hundreds of million of dollars, as well as the loss of careers and reputations (e.g., Albrecht, Wernz & Williams, 1995). Yet, the detection problem is not hopelessly unsolvable: small percentages of individuals in experimental studies of deception (e.g., secret service agents, judges, psychiatrists, auditors) have been found to consistently succeed at detecting forms of deceit within tasks in their domains of expertise (e.g., Ekman 1992; Johnson, Grazioli & Jamal, 1992).

The study reported here investigates success and failures in detecting deceptions in one such task. Specifically, we examine how twenty-four practicing auditors cope with the task of detecting deception (fraud) in cases in which financial information was manipulated by management for the purpose of making companies appear more profitable than they actually were.

Using the auditors' data, we develop a theory of detecting financial deceptions based on the proposition that successful individuals adopt a variant of Dennett's intentional stance strategy (1989). This strategy consists of the application of deception-detection heuristics developed outside the context of auditing (Johnson, Grazioli & Jamal, 1993). The theory is evaluated by building a 'competence' model of deception detection that embodies these heuristics, and by testing the model against the cases given to the auditors.

We expand the theory to account for the observed behavior in our sample of auditors, inclusive of successes and failures. Because the detection of fraud is relatively infrequent (Hansen, 1996; Loebbecke, Einning & Willingham, 1989), most auditors do not receive feedback from which they can develop skills in fraud detection. As a result, learning how to apply detection heuristics in the domain of auditing is incompletely specified (underconstrained). Knowledge is generated that is compatible with past experience, but fraught with the potential for failure.

Accordingly, we explain failure at detecting deception by means of imperfections in the knowledge of the domain in which the deception occurs (i.e., accounting). Specific hypotheses about these imperfections are derived from the errors observed in the behavior of a subset (five) of the twenty-four auditors. We evaluate the theory of success and failure at detecting financial deceptions by seeding the hypothesized imperfections in the competence model of deception detection and using the modified model to predict detection outcomes for the remaining nineteen auditors.

2. The problem of fraud detection

Most companies whose securities are publicly issued and held are required to release audited financial statements annually. These statements are a means of informing various interested parties about the company's economic conditions. Investors, creditors and regulators use financial statements as evidence in making decisions such as investing in a company's stock, granting a loan, and assessing compliance with laws and regulations. Since these decisions have consequences for the company, management may opt to misstate its financial conditions for the purpose of favorably affecting the outcomes of these decisions.

The public has incurred substantial losses due to financial misrepresentations (Albrecht, Wernz & Williams, 1995). Through the legal system, society has charged professional auditors with responsibility for detecting fraudulent misrepresentations and determining whether the financial statements issued by a company are fairly presented. Financial statements are prepared for companies on the basis of the Generally Accepted Accounting Principles (GAAP), a set of principles that specify how to fairly present the financial condition and results of operations of a company.

In the study reported here, we focus on the detection of fraudulent manipulations of financial reports. Such manipulations cover intentional misstatements or omissions of amounts or disclosures in financial statements, which are intended to deceive the users of those statements. The context for the study is the auditing task called second-partner review. In this task, a partner of an accounting firm reviews draft financial statements of a client company, footnotes to the statements, and other descriptive financial materials. The objective of the review is to evaluate the 'engagement' partner's conclusion as to the appropriate audit 'opinion.' The engagement partner is the auditor who directs the audit fieldwork and develops the opinion as to the fairness of the company's financial statements.

Several audit opinions are possible. If the auditor feels that the financial statements are fairly presented, an 'unqualified' (or 'clean') opinion is issued. If the auditors feels that the statements are fairly presented, yet some areas of concern need to be highlighted, an 'unqualified opinion with additional paragraph' for items such as doubt as to continued financial viability (going concern), material uncertainties or lack of consistency, is issued.

If the auditor disagrees with the accounting methods used, or with the manner in which they are applied, and the violation has a material effect on the financial statements, the auditor renders either a 'non-GAAP' or an 'adverse' opinion. A 'non GAAP' opinion is issued when there is a departure from GAAP as to the accounting principles selected or their application, and where disclosures are inadequate. An adverse opinion is issued when the departures from GAAP are judged to be very material.

In the experiment reported here, twenty-four practicing auditors who were partners in a major international accounting firm were asked to perform a second partner review of four cases on which an unqualified opinion was issued. Subjects in the experiment had on average 19.7 years of experience as auditors, with a range from 12 to 40 years. One-half of the auditors had experience in the high technology medical products industry, the other half had experience in the retail industry.

Each of the four cases is based on the annual report produced by a publicly traded firm, abridged to a length of approximately 15 single-spaced pages. All cases describe real companies. Two cases (Pharmaceuticals, Inc. and Surgical Products, Inc.) describe medical products companies. The other two cases (America's Family Video Store and Big John's Electronic Sales, Inc.) depict retail companies. The names of the companies and other identifying information were altered to avoid recognition.

Subjects were not informed that the cases they would examine contained financial statement frauds. The frauds present in the four cases were constructed by management and were missed by the accounting firm responsible for conducting the audit of the company's financial statements. The frauds were designed to increase reported income and assets, and range from overstating the value of the inventory to improperly capitalizing costs that should be expensed. A brief description of the cases and their manipulations is presented in Appendix A.

In addition to the four fraud cases, each subject was also given a nonmanipulated ('clean') case as an experimental control designed to check for the frequency of 'false positives,' that is, cases where the subject feels that a fraud is present when there is none. Audit partners with experience in the medical products industry received a clean medical industry case (Vascular Products) as their first case, followed by the four fraud cases. Audit partners with experience in the retail industry received a clean retail industry case (Southern Retail) as their first case. The four fraud cases were given to all subjects in the same order: America's Family, Pharmaceuticals, Big John's, and Surgical Products.[1]

All subjects worked on the cases in individual sessions held in their offices. Each subject was given a set of written instructions in which they were asked to conduct a second-partner review of each case and to evaluate the appropriateness of the engagement partner's conclusion that the case presented no material issue (an unqualified audit opinion).[2] Upon completion of the five cases (four fraud and one clean case) each subject was asked to rate the appropriateness of issuing the six alternative audit opinions listed above on each case using a 7-point scale. The 'unqualified' opinion is the correct solution for the two clean cases. Non-GAAP is an acceptable solution and 'adverse' is the best solution for the four fraud cases.

Since we are interested in understanding how auditors detect deception, as opposed to modeling technical differences among different auditing opinions, we assigned the opinion chosen by the auditors to one of three outcomes: 'Misleading,' 'Unqualified+,' and 'Unqualified.' 'Misleading' is used when the auditors chose an 'adverse' or 'non-GAAP' opinion as most appropriate for the case. 'Unqualified+' is used when the auditors chose to issue a lack of consistency, going concern, or material uncertainty paragraph. 'Unqualified' is used when the auditors chose the clean opinion.[3]

The process of thinking used by the auditors was captured by instructing each subject to think-aloud (Ericsson & Simon, 1993) while performing each review task. All sessions were tape-recorded, transcribed and coded according to a scheme developed in previous research on the fraud detection process and described in Appendix B (see also Jamal, 1991; Johnson, Grazioli & Jamal, 1992).

The category of outcome for the opinion given by each subject on the four fraud cases and the clean case is presented in Table 1. Inspection of the data in Table 1 reveals that detecting financial deception is a difficult task. Two subjects failed all fraud cases, and eighteen (out of twenty four) failed at least two cases. Using a strict definition of success (i.e., giving an adverse opinion) no subject was successful on all four fraud cases and twenty subjects failed at least three cases. Seven subjects also failed to give an unqualified opinion on the clean cases. The responses of these subjects are discussed in a later section of the paper.

The data in Table 1 also show that detecting financial deception under the experimental conditions of this study is not impossible. Two subjects (S1 and S2) were successful on all fraud cases. Four subjects were successful on three out of the four fraud cases. Most subjects succeeded in at least one case. Analysis of the think-aloud protocols of the individuals who detected the deception reveals that these successes were not chance events; rather, they were the conclusions of insightful financial analyses. Table 1 also shows that industry specialization does not appear to have a clear effect. While some individuals who have experience in the industry of a case (medical products or retailing) were able to conclude an adverse opinion, so also were individuals who did not have such experience.

The variability present in the data in Table 1 poses a challenge to our ability to develop a single theory that explains the behavior of the subjects who perform the fraud detection task successfully, as well as those who fail. We tackle this problem in two steps: first, we construct a theory of successful detection of financial deception, and second, we expand the theory to account for failures to detect.

3. A theory of detecting deception in financial information

The nature of financial statement fraud poses two constraints on the knowledge that is developed to support success at detecting this form of deception: 1) infrequent occurrence of

Table 1
Audit opinions given by twenty-four auditors on four fraud and two clean cases[a]

Subject[b]	Medical fraud cases		Retail fraud cases		Clean cases	
	Pharma-ceutical	Surgical products	Big John's	America's family	Vascular products	Southern retail
S1	Misleading	Misleading	Misleading	Misleading	Misleading	
S2	Misleading	Misleading	Misleading	Misleading	Unqualified+	
S3	Misleading	Misleading	Unqualified+	Misleading	Unqualified	
S4	Misleading	Misleading	Unqualified+	Unqualified	[c]	
S5	Misleading	Unqualified	Misleading	Unqualified+	Unqualified+	
S6	Misleading	Unqualified+	Unqualified	Misleading	Unqualified	
S7	Misleading	Misleading	Misleading	Unqualified+	Unqualified	
S8	Unqualified	Misleading	Unqualified+	Misleading	Unqualified	
S9	Unqualified	Unqualified+	Misleading	Unqualified+	Unqualified+	
S10	Misleading	Unqualified+	Unqualified+	Unqualified+	Misleading	
S11	Unqualified	Misleading	Unqualified	Unqualified	Unqualified	
S12	Unqualified+	Misleading	Unqualified+	Unqualified	Unqualified	
S13	Misleading	Unqualified+	Misleading	Misleading		Unqualified
S14	Misleading	Unqualified+	Misleading	Misleading		Unqualified
S15	Misleading	Unqualified	Misleading	Misleading		Unqualified
S16	Unqualified+	Misleading	Misleading	Unqualified+		Unqualified+
S17	Unqualified	Unqualified+	Misleading	Unqualified		Unqualified
S18	Unqualified	Unqualified+	Misleading	Unqualified		Unqualified
S19	Misleading	Misleading	Unqualified+	Unqualified+		Unqualified
S20	Misleading	Unqualified+	Misleading	Unqualified		Unqualified
S21	Misleading	Unqualified	Unqualified+	Unqualified		Unqualified
S22	Misleading	Unqualified+	Unqualified+	Unqualified		Unqualified
S23	Unqualified+	Unqualified+	Unqualified+	Unqualified		Unqualified
S24	Unqualified	Unqualified	Unqualified	Unqualified		Unqualified

[a] 'Misleading' indicates that the auditor rendered either an 'Adverse' or a 'Non-GAAP' opinion. 'Unqualified+' indicates that the auditor felt that the case is fairly represented, yet there are areas of concern that need to be explicitly highlighted. 'Unqualified' indicates a case judged to be fairly represented. Specific definitions of these audit opinions can be found in the main text.

[b] Subjects S1–S12 specialize in medical companies, S13–S24 specialize in retail companies.

[c] Subject S4 judged equally likely a Misleading (Adverse), an Unqualified+, and an Unqualified Opinion on this case.

successful detection, and 2) the presence of an adversary (i.e., management). Building a theory of the knowledge that supports successful performance on infrequently occurring tasks is challenging because a low rate of occurrence rules out explanations of performance based on direct and repeated experience (e.g., Ericsson, 1996; Larkin, McDermott, Simon & Simon, 1980). Since fraud detection occurs infrequently in the professional life of a typical auditor (Loebbecke, Einning & Willingham, 1989), we must propose a theory of success based on knowledge that is partially independent from specific experience.

The presence of an adversary distinguishes fraud detection from tasks in which a cognitive agent generates and tests hypotheses about a malfunction in some system of interest (e.g., Clancey, 1988; Johnson, Kochevar & Zualkernan, 1992). While most diagnostic tasks (e.g., medical diagnosis, electronic troubleshooting) entail determining faults generated by unintentional events (e.g., a heart condition, or a faulty component), fraud detection is characterized by the presence of an adversary (management) who intentionally causes the misrepresentation of a company's economic conditions that is, a financial deception (Johnson, Grazioli & Jamal, 1993; Thagard, 1992). Frauds are difficult to detect because they are manipulated so as to appear to be the result of a fair financial reporting process.

We assume that cognitive agents (auditors included) have frequent and varied experience with the phenomena of deception, both as a Deceiver and a Target-that is, victim-of deception (e.g., Ceci, Leichtman & Putnick, 1992; Ekman, 1992; Grover, 1993). As a way to cope with potentially deceiving adversaries, cognitive agents develop knowledge for detecting deceptions (DePaulo, Stone & Lassiter, 1985; Vasek, 1986). This knowledge is particular to the task of detecting deceptions, but is not expressed in terms of specific domain content (Cosmides, 1985, 1989; Cosmides & Tooby, 1994, 1995).

From this assumption and a general strategy for solving problems that involves interactions with other agents (the intentional stance–Dennett, 1989), we describe heuristics that can be used to solve the deception detection problem. Drawing on work by Bowyer (1982) and Mawby & Mitchell (1986), these heuristics are identified by first studying the ways in which a deception can be created by an adversary, and then deriving means by which these deceptions can be detected.

To succeed, the Deceiver must have knowledge of how to bring about failure in the process that is used by the Target to interpret information and act in the world. We assume that this knowledge includes an ability to envision the process of the Target's thinking (Dennett, 1989; Wellman, 1990). The Deceiver's theory of the Target's mind includes three basic processes (e.g., Neisser, 1976): 1) the Target seeks (scans or accesses) information in the environment, 2) the Target categorizes or interprets this information, and 3) as a result of this categorization or interpretation the Target takes action that includes further sampling of information.

We identify two types of failure in the Target's thinking that are brought about by actions of a Deceiver. We refer to the actions of the deceiver as 'deception tactics' and describe them by reference to a deception 'core,' which is the item that the Deceiver either intends to hide from the Target, or whose fictitious properties the Deceiver intends to simulate (Johnson, Grazioli & Jamal, 1993).

In the first type of failure the Deceiver acts to disrupt or influence the Target's process of seeking information. This is done in one of three ways: 1) the Deceiver attempts to block or remove attributes of an object, event or state of affairs (the deception core), 2) the Deceiver attempts to diffuse or confuse attributes of the core that it wants the Target to miss, and 3) the Deceiver attempts to focus the Target's attention away from the core. We summarize these three actions of the Deceiver by the terms *masking, dazzling* and *decoying,* respectively (from Bowyer, 1982).

In the second type of failure, the Deceiver acts to influence the Target's process of categorizing and interpreting information. We suppose that this is also done in three ways: 1) the Deceiver frames or labels the core so that the Target misclassifies it, 2) the Deceiver modifies the core so that it is mistaken for something else, and 3) the Deceiver provides weak evidence for the core so that the Target will dismiss it. We label these actions *repackaging, mimicking* and *double play,* respectively (cfr. Bowyer 1982). The six deception tactics are summarized in Table 2, with examples from the domain of accounting.

Table 2
Deception tactics (adapted from Johnson et al. 1993)

Deception tactics	Examples of tactics for creating a financial deception
Masking	Failing to record or disclose an expense or a liability
Dazzling	Disclosing information in the footnotes to the statements rather than in showing it in their body
Decoying	Emphasizing legal issues (blind alleys) that after a close examination turn out to be immaterial or handled appropriately
Repackaging	Changing the descriptions or labels that characterize economic entities or reframing issues to maliciously justify the use of favorable accounting procedures
Mimicking	Creating fictitious transactions or transactions without substance
Double play	Improperly applying General Accepted Accounting Principles, where an item is not individually material

Just as the Deceiver solves the problem of constructing a deception by using knowledge of how the Target thinks and acts, we assume that the Target solves the problem of detecting manipulations by using knowledge of how the Deceiver thinks and acts when constructing a deception.[4] This includes an ability to ascribe goals and actions to the Deceiver (Dennett, 1989; Trope, 1986; Wellmann, 1990). Specifically, we assume the process used by the Target to infer the behavior of the Deceiver has three components: 1) the Target identifies inconsistencies between observations in the environment and expectations for these observations, 2) the Target determines that the inconsistencies are functional to the goals of the Deceiver, and 3) the Target identifies potential actions of the Deceiver associated with one or more of the six manipulations of the environment described above, and judges that the Deceiver is able to manipulate the environment so as to create the observed inconsistencies.

We describe the Target's knowledge for detecting a deception by a set of domain-independent heuristics termed 'detection tactics.' In what follows, we derive properties of these tactics and propose models of them, which we represent in the form of schemata (Bartlett, 1961; Fiske & Taylor, 1991; Shank & Abelson, 1977). Alternative representations for deception events are possible: Thagard (1992) uses networks of propositions, Worden (1996) uses scripts, and also discusses pros and cons of alternative forms of representation.

From the perspective of the Target, we assume a detection tactic exists to counter each of the six deception tactics employed by the Deceiver. Accordingly, we identify a schema for each detection tactic. Each schema contains conditions that identify a situation as potentially deceptive and initiate the hypothesis that a specific deception tactic (e.g., masking) has occurred (Johnson, Grazioli & Jamal, 1993).

As an example, the schema for detecting a masking deception has three conditions: 1) an expectation about the presence of an item in the environment is violated (i.e., something is unexpectedly missing); 2) the determination that this violation is functional to one of the goals that can be ascribed to the other agent (the deceiver); and 3) the determination that another agent has the capability of deleting the item from the environment of the Target. When a given situation satisfies these conditions, the hypothesis of deception is generated and actions that aim to restore a correct representation of the deception core are taken.

To be useful, the schemata developed for each detection tactic must be applied to specific situations, which requires verification that the conditions for a given schema apply. This cannot be done without *mediating knowledge* of the domain in which the detection of deception occurs. Such knowledge maps the conditions of the schemata into entities and relationships in a particular task situation (e.g., second partner review of a set of financial statements).

We treat the detection tactics as specifications for knowledge from the domain of accounting that is required to detect financial deceptions. As applied to auditing, each detection tactic requires domain-specific knowledge that identifies: 1) inconsistencies, 2) possible adversarial goals of the Deceiver (management), and whether a given inconsistency contributes to the achievement of these goals and, 3) manipulations that can generate the inconsistency. In the next section we introduce a model of fraud detection that contains domain knowledge for identifying inconsistent information in financial statements and generating interpretive hypotheses about these inconsistencies based on management goals and possible manipulations of the accounting process.[5]

The proposition that fraud detection is based on knowledge of how to detect deceptions in the world satisfies the two constraints identified at the beginning of this section. First, it addresses the low-base-rate constraint because it argues that this knowledge develops by abstraction from numerous instances of deception to which the auditor has been exposed in everyday life. Second, it fully appreciates the adversarial nature of detecting the deceptions created by another agent (management) as it turns this constraint into a source of power by using the goals and actions ascribed to the agent as a basis for detecting the actions that such an agent may have taken in order to perpetrate the deception.

4. Competence model of fraud detection

To gain precision in our explanation of success in the task of fraud detection, we construct a model of the knowledge that is sufficient for success in the class of detection tasks faced by the subjects in our research. This *competence model* is an idealization of human performance at detecting financial statement fraud in the task of second-partner review. It is based on models described in previous work on cognition in audit settings (Bouwman, 1983; Johnson, Grazioli & Jamal, 1992; Johnson, Jamal & Berryman, 1991; Peters, 1990), interviews with a domain informant, and analysis of cases in the experimental materials given to subjects in the present study.

The competence model implements a method for detecting deceptions that is consistent with the theory introduced in the previous section. This method is based on the generation and testing of hypotheses that explain inconsistencies between expectations and information given in a set of financial statements (cues). Fig. 1 shows the method's structure. The input to the method is composed of a set of financial statements. The method's output is an indication of the extent to which the financial statements items are a fair representation of the company's economic condition.

The method decomposes fraud detection into four distinct processes. The first process–activation–generates expectations for the values of cues in the financial statements (e.g., the inventory balance). The expectations are compared to the observed values. The magnitude of the discrepancy between observed and expected determines whether the cue is labeled as

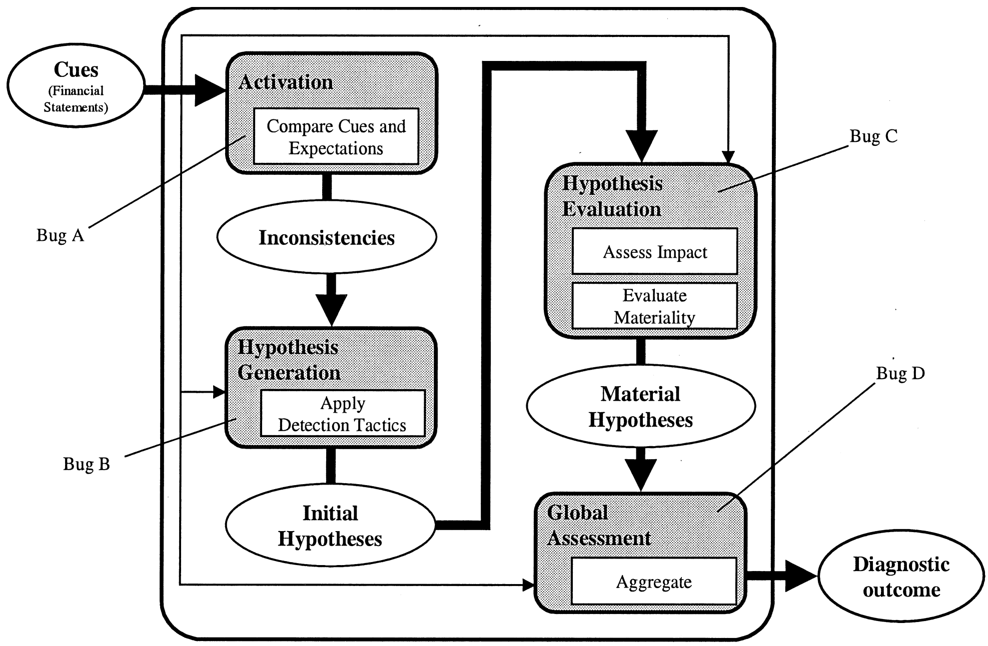

Fig. 1. Fraud Detection Method

inconsistent. A second process–hypothesis generation–proposes hypotheses to explain the inconsistencies resulting from 1) intentional manipulations, 2) errors in accounting, or 3) insufficient disclosure of financial information. The third process-hypothesis evaluation-process assesses hypotheses on the basis of their materiality and the fourth process-global assessment-process combines all accepted hypotheses and produces a final outcome, which is an indication of the extent to which the financial statements items are a fair representation of the company's economic condition. A detailed description of each process appears in Appendix C.

The competence model shown in Fig. 1 was implemented as a computer program using a classical production system architecture (e.g., Giarrantano & Riley, 1989; Newell & Simon, 1972) with a forward chaining strategy.[6] The input to the program is a set of 24 data items from the financial statements that include present and past balances and information from the notes to the statements. The output of the program is one of three conclusions about the information in the statements: 'misleading,' 'unqualified+,' and 'unqualified.'

The model was given the six cases shown in Table 1. Based on previous analysis of the cases (Appendix A), historical documents, and the opinion of the collaborating expert, we identified two major manipulations in each case. Surgical Products was manipulated by masking the R&D costs and by using the repackaging strategy to modify the accounting policies concerning molds and dies. Pharmaceutical was manipulated by altering (repackaging) the inventory and mailing lists balance. Big John's inventory and pre-opening costs were manipulated (repackaging). American Family was manipulated by repackaging the inventory and using a double-play tactic in goodwill amortization. In addition to the main manipulations, we also identified use of the repackaging tactic for the inventory of Surgical

Table 3
Process trace of the competence model on four fraud and two non-fraud cases[a]

Cue	Subprocess in the model	Cases					
		Pharma-ceuticals (fraud)	Surgical products (fraud)	Big John's (fraud)	America's family (fraud)	Vascular product (clean)	Southern retail (clean)
Molds & dies	Activation		HIT				
	Hp. generation		repackaging				
	Hp. evaluation		HIT				
Change in accounting estimates	Activation		HIT				
	Hp. generation		double play				
	Hp. evaluation		MISS				
Goodwill amortization	Activation				HIT		
	Hp. generation				masking		
	Hp. evaluation				HIT		
Inventory	Activation	HIT	HIT	HIT	HIT		
	Hp. generation	repackaging	repackaging	repackaging	repackaging		
	Hp. evaluation	HIT	HIT	HIT	HIT		
Mailing list	Activation	HIT					
	Hp. generation	repackaging					
	Hp. evaluation	HIT					
Pre-opening cost	Activation			HIT			
	Hp. generation			repackaging			
	Hp. evaluation			HIT			
R & D	Activation	False alarm	HIT			False alarm	
	Hp. generation	masking[b]	masking			masking [b]	
	Hp. evaluation	corr. reject.	HIT			corr. reject.	
Sales return	Activation		HIT				
	Hp. generation		Error Hp.[c]				
	Hp. evaluation		Miss				
Overall	Global assessment	Misleading	Misleading	Misleading	Misleading	Unqualified	Unqualified

[a] HIT = In Activation: the model correctly identified the cue as inconsistent; in Hypothesis Evaluation: the model accepted a correct hypothesis.

MISS = In Activation: the model failed to identify a manipulated cue as inconsistent; in Hypothesis Evaluation: the model accepted an incorrect hypothesis.

False Alarm = In Activation: the model generated an inconsistency when there was none.

Corr. Reject. = In Hypothesis Evaluation: the model rejected an incorrect hypothesis.

Repackaging, Masking, etc. are the hypothesis generated by the model to explain inconsistencies identified during the activation phase. Misleading and Unqualified are the final outcomes selected by the model for each case.

[b] The model generated an incorrect masking hypothesis. No masking manipulation or error was present for this cue.

[c] The model generated an incorrect hypothesis of error in the accounting process. The sales returns in the case were the object of a repackaging manipulation.

Products, mimicking for their sales returns, and double-play for their change in accounting estimates.

An abbreviated process trace of the program on the four fraud cases and the two nonfraud cases is presented in Table 3. Following the decomposition of the method shown in Fig. 1, the Table presents the intermediate and final processing results of the activation, hypothesis generation, hypothesis evaluation, and global assessment processes. Each row in the Table represents a cue, and each column represents a specific case. The cells contain the responses of the model to each cue. The last row contains the final outcomes selected by the model as a response to the results of the intermediate processing of the cues in each case.

The first line of each cell in Table 3 indicates whether the model correctly recognized (a hit), failed to recognize (a miss), or incorrectly recognized (false alarm) each manipulated cue. Table 3 shows that the model's success on generation of inconsistencies was 85% (11/13): the model identified the eight manipulations in the four cases, plus three more that were singled out from analysis of the cases. The model had no misses and two false alarms, in which it incorrectly identified two cues as inconsistent.

The second line in each cell of Table 3 indicates the initial hypothesis the model selected to explain each identified inconsistency. The model incorrectly generated the initial hypothesis that the R&D balances in the financial statements of Pharmaceuticals and Vascular Products were symptoms of masking manipulations (false alarms). The model selected these hypotheses because the R&D expenditures at these companies were not keeping up with the rate of growth of revenue. The model expected that high-tech firms should have significant R&D expenditures, and that these expenditures should be growing with the revenue. When this expectation was violated, the model generated the hypothesis that part of these expenditures had not been reported (masking).

The model also generated the incorrect initial hypothesis that the sales returns balance in the Surgical Products case was the result of an accounting error. The hypothesis was generated because the reported sales returns-a cost to the company-appeared to increase more than sales. The sales return balance is one of the manipulations created by the management of the company.[7] No deception hypothesis was generated by the model because higher sales returns are not functional to the goal of overstating income ascribed to management.

The last line in each cell shows the results of the evaluation of the hypotheses generated by the model to explain the inconsistencies. The model correctly rejected the two false alarms proposed in hypotheses generation. It also incorrectly accepted the hypothesis of an error in accounting generated to explain the sales return and dropped the correct deception hypothesis generated to explain the change in accounting estimates in Surgical Products because the hypothesis did not pass the materiality test described in the previous section. Overall, the model successfully evaluated 85% of the generated hypotheses (11/13, by coincidence the same success ratio as above).

The final outcomes resulting from the global assessment of the confirmed hypotheses are presented in the last row of Table 3. All six final outcomes are correct. In the Surgical Products case, the model chose the more serious 'misleading' outcome, suggested by the confirmed hypotheses of deception, over an 'unqualified+' outcome suggested by the confirmed hypothesis of error (the sales return).[8]

Overall, the results shown in Table 3 indicate that the mechanisms described by the theory are sufficient to detect the manipulations contained in the four fraud cases. In addition, the model derived from the theory was able to recognize the two clean cases as ones that did not contain financial deceptions. Based on these results we conclude that the proposed theory is a plausible account of successful detection of the deceptions used to misrepresent financial information; however, the theory is clearly inadequate to explain the behavior of the auditors in our experiment, most of who failed the fraud detection task.

5. Explaining auditor behavior

To explain why auditors fail at detecting the deceptions created by management, we draw on the early repair theory developed by Seely-Brown and VanLehn (1979), the more recent bug theory proposed by VanLehn (1990), and analysis of cognitive misrepresentations by Millikan (1993) and Cummins (1989). We begin with the argument that the knowledge that leads to failure is the result of an agent's adaptation to a task environment. We suppose that the knowledge that underlies failure is not radically different from the knowledge that supports success, so that it is possible to isolate local imperfections that may be responsible for errors and failure of performance.

We make two assumptions about the nature of the knowledge held by cognitive agents. The first assumption is that task base-rate affects the power of an agent's knowledge. A high base-rate task is one that is performed frequently. In general, feedback on solutions in such tasks offers opportunities to revise (and improve) knowledge that initially leads to failure. By contrast, in low base-rate tasks feedback occurs infrequently so that agents have little opportunity to refine their knowledge to avoid future errors (Druckman & Bork, 1991). Since the second partner review of a set of nonmanipulated financial statements is a frequently performed task, we expect the knowledge that supports it to be stable and powerful. At the same time, we expect the knowledge that supports the review of deceptive statements to be less polished and less effective.

Our second assumption is that the cognitive architecture that supports the development of problem-solving knowledge can be functionally divided into two components (VanLehn, 1990): an acquisition function and an interpretation function. The acquisition function takes as input samples of problems and their solutions, and generates as output knowledge that supports solutions to similar problems. The interpretation function takes as input the knowledge produced from the acquisition function, plus a problem instance, and generates as output a solution.

The acquisition function is activated by impasses in problem solving (Anderson, 1989; Laird, Newell & Rosenbloom, 1987; VanLehn, 1990). Impasses are situations in which the available knowledge does not apply completely to the task at hand. Under these circumstances, the cognitive agent considers an impasse to be a problem and tries to solve it. The process of overcoming an impasse is called repair. Repairs are changes to the agent's operative knowledge. Repairs can be based on generalization, integration (VanLehn, 1990), and analogy (Anderson, 1989). They also include new or modified assumptions about the applicability of available solutions to the problems to be solved.

There is no a-priori guarantee that a given repair creates operative knowledge that leads to a correct solution to a given problem (Ben-Zeev, 1995; VanLehn, 1990). Before receiving feedback on the quality of a solution, the agent does not know if a proposed repair solves the problem correctly or not. If the problem is not solved, another repair is needed. If the problem is solved, however, an intriguing situation may occur: the repaired knowledge successfully solves the problem at hand, but-unknown to the agent-fails to solve other problems that are similar in form to the one that spurred the repair. In this situation we say that a ‘bug’ has been introduced in the agent’s knowledge (VanLehn, 1990).

The concept of knowledge bug as inappropriately repaired operative knowledge is the focus of our explanation of error in the behavior of individuals who fail to detect deceptions. Error is a term used to qualify a given behavior with respect to some normative criterion that specifies rules for that behavior. In a given situation, whether a particular behavior is an error depends on a norm specified by the observer. To the extent that we isolate specific elements of the processes and representations that are responsible for error behaviors, we call these elements knowledge bugs.

While the knowledge that supports high base-rate tasks is likely to be correctly repaired because of the extensive feedback from task performance, the knowledge that supports low base-rate tasks is more likely to contain bugs. In the domain of accounting, for example, this means that some repairs of the knowledge used to support second-partner review of nonfraudulent companies fail when they are applied to the review of fraudulent companies.

The knowledge bugs that affect low base-rate tasks such as fraud detection share four characteristics:

Similarity: Since repairs are based on the existing knowledge, and since this knowledge is acquired on the basis of feedback from a high base-rate task (i.e., second-partner review of nonfraudulent companies), the buggy knowledge that affects the solution to the low base-rate task (fraud detection) is similar to the knowledge that supports frequent success.

Simplicity: Since bugs are repaired elements of knowledge that the agent could not refine due to a lack of feedback (because of the low base-rate of occurrence), they tend to be relatively simple distortions or transformations of the knowledge that supports success.

Noninterference: Since bugs that interfere with the knowledge that supports high-base rate tasks are quickly repaired, it follows that the remaining bugs do not interfere with this knowledge.

Plausibility: Since a bug is a bona-fide repair to an impasse, it should support limited success, or at least not be implausible.

To understand differences between the subjects who fail and those who succeed at the task of detecting financial statement frauds, we propose specific hypotheses about imperfections (bugs) in the knowledge that supports fraud detection. We then evaluate these hypotheses by implementing them in a *performance model* of fraud detection, and by comparing the behavior of this model with the behavior of subjects in Table 1.

6. Performance model of fraud detection

A performance model of fraud detection is a way to account for the behavior of subjects engaged in a particular set of tasks. As such, it should diverge from a *competence model* and fail to detect the frauds contained in the case materials when the subjects it models fail to do so. The performance model is obtained by seeding the hypothesized imperfections in the competence model described earlier. The resulting model exhibits error behaviors and (sometimes) ultimate detection failure.

A known difficulty in proposing hypotheses about sources of errors is that typically many of these sources are compatible with the same error behavior (Reason, 1990). Unconstrained attempts to generate error hypotheses for complex tasks may quickly become an endless list of possible malfunctions, difficult to test and evaluate (VanLehn, 1990).

To overcome this difficulty, we introduce three sources of constraint to select candidate hypotheses. The first source of constraint is the competence model of the task proposed above. This model contains detailed descriptions of processes and representations that support success in the detection of deception. The second source of constraint is the four theoretical properties of knowledge bugs that occur in low base-rate tasks (listed above). The third source of constraint for the knowledge bugs is the actual errors committed by each auditor, as described next.

The sample of 24 subjects whose data on the fraud detection task are shown in Table 1 was divided into two subsets: a modeling subset (5 subjects) and a testing subset (19 subjects). The modeling subset was used to identify errors and develop hypotheses to explain them. The second subset was put aside for testing purposes. Two of the five subjects in the modeling subset (S1 and S2) were selected because they detected fraud in all four cases, which makes them the best performers in our sample. One subject (S24) was selected because he failed to detect the manipulations and consistently rendered an unqualified opinion in each of the four cases. This subject was the worst performer in the sample. The other two subjects in the modeling subset (S7 and S12) were randomly selected among those occupying an intermediate position in performance. One subject was successful on three cases; the other was successful on one case.

An error analysis (Johnson, Grazioli, Jamal & Zualkernan, 1992) was performed based on the method for fraud detection shown in Fig. 1. In this analysis, the verbal protocols of the five subjects in the modeling subset were coded to determine which of the cues in each of the four fraud cases were considered inconsistent, what hypotheses were proposed and accepted for explaining them, what final opinion was rendered and on what basis (see Appendix B). The resulting process traces (one trace per case per subject, for a total of $5 \times 4 = 20$ traces) were compared with the process trace of the competence model (Table 3) and, treated as a specification for normatively correct behavior. On this basis, we identified four categories of error (Error A, B, C, and D-as shown in Fig. 1):

Error A (activation error)-This type of error is defined as missing a cue, or dismissing it early, that is, after attending to it, but without generating any inconsistency.

Error B (hypothesis generation error)-This type of error is defined as interpreting an inconsistency as an error in accounting or as an insufficient disclosure issue.

Error C (evaluation error)-This type of error is defined as failing to evaluate an inconsistency as material.

Error D (assessment error)-This type of error is defined as failing to take into account a hypothesis while selecting a final opinion.

Results of the error analysis applied to the five subjects in the testing set are presented in Table 4. For the five subjects in the modeling subset there were 49 instances of the four types

Table 4
Errors in the behavior of five subjects from Table 1

Subject	Case: Pharmaceutical Products									Case: Surgical Products								
	Types of Errors in processing the mailing lists cue[b]				Types of Errors in processing the inventory cue				Outcome[a]	Types of Errors in processing the molds and dies cue				Types of Errors in processing the R&D cue				Outcome[a]
	A	B	C	D	A	B	C	D		A	B	C	D	A	B	C	D	
S2									Misleading									Misleading
S1		X				X			Misleading		X				X			Misleading
S7						X			Misleading		X							Misleading
S12		X		X		X		X	Unqualif.+		X				X			Misleading
S24		X	X	X		X	X	X	Unqualified		X	X	X	X				Unqualified

Subject	Case: Big John's									Case: America's Family Video								
	Types of Errors in processing the pre-opening costs cue				Types of Errors in processing the inventory cue				Outcome[a]	Types of Errors in processing the inventory cue				Types of Errors in processing goodwill amortization				Outcome[a]
	A	B	C	D	A	B	C	D		A	B	C	D	A	B	C	D	
S2						X			Misleading									Misleading
S1	X					X			Misleading						X			Misleading
S7						X			Misleading		X	X	X		X	X		Unqualif.+
S12		X				X	X		Unqualif.+		X	X			X	X	X	Unqualified
S24		X	X	X		X	X	X	Unqualified		X	X		X				Unqualified

[a] Unqualif.+ standards for 'unqualified with additional paragraph.'
[b] Error A (activation error) = missing a cue, or dismissing it early. Error B (generation error) = interpreting an inconsistency as an error in accounting or as an insufficient disclosure issue. Error C (evaluation error) = failing to evaluate an hypothesis as material. Error D (assessment error) = failing to take into account an hypothesis while selecting a final opinion.

of errors across the four cases. Error B was the most frequent (26 occurrences), Error A the least frequent (3 occurrences).

Table 4 also shows that no subject in the modeling subset, including the successful ones, was flawless across all cases. S1 made one error; S24 made 19. Not surprisingly, task performance appears to decline as the number of errors increases. However, the data indicate that the presence of a small number of errors does not necessarily degrade performance, suggesting the presence of redundancies in the problem solving knowledge used by the auditors.

Examination of the data shown in Table 4 also indicates that, in general, subjects did not manifest similar error patterns across cases and that cases did not induce similar error patterns across subjects. This finding, which is similar to the 'migration' phenomenon reported by VanLehn (1990), and the 'mal-rules' instability observed by Payne and Squibb (1990), is consistent with the conjecture that subjects are not well adapted to the task of fraud detection.

Overall, errors appear to be related to outcomes in non-obvious ways. The observation of success in the presence of error, as well as failure resulting from different error patterns, suggests that errors interact in determining subject performance (Johnson, Grazioli & Jamal, 1992). To explain failure at detecting fraud we need to take into account how these local errors affect the processing of information about a case. We do this by first identifying knowledge bugs for the observed errors, and then by modeling their interaction.

7. Knowledge bugs

We begin by hypothesizing four specific perturbations (bugs) in the knowledge of accounting embodied in the competence model of fraud detection that lead to errors in performance, and hence failure to detect the financial frauds in the case materials.

7a. Bug A. Error A is defined as the inability to identify a manipulated cue as inconsistent. There were three occurrences of Error A in the behavior of the subset of subjects shown in Table 4. These errors were made by two subjects: S1 on Big John's and S24 on America's Family and Surgical Products. Examination of the competence model suggests that Error A may be generated by either a poor procedure that generates the expectation for the cue, or by use of a poor criterion to establish that the difference between cue and expectation requires further investigation.

Since the criterion is found in specific professional guidelines that are likely to be known by the subjects, we speculate that Error A is generated by a poor procedure that generates the expectation for the cue. Accordingly, Bug A is implemented in the model by forcing the expectation of a cue to be equal to its actual value, so that when the value and the expectation are compared, the model does not produce an inconsistency (because their difference is less than the criterion).

7b. Bug B. Error B refers to the generation of incorrect hypotheses when interpreting inconsistencies. The subjects in the modeling subsample committed the hypothesis generation Error B 26 times (Table 4). In 19 occurrences this failure resulted in the hypothesis of an error in accounting. For the remaining seven instances an insufficient disclosure hypothesis was generated.

VanLehn (1990) suggests that a frequent cause of bugs is inappropriate generalizations created to repair impasses. In our research, interpreting a fraudulent manipulation as an error in the valuation process, or as a disclosure issue, is an inappropriate generalization. This generalization is plausible because error and disclosure problems have a higher frequency than fraud (Koehler, 1996), and at times the line between an intentional and unintentional misrepresentation is quite fine. As a result, our model does not distinguish between errors in accounting that are functional to increasing operating income and intentional misrepresentation designed to achieve the same result.

According to the theory proposed in the previous sections, failure to generate a deception hypothesis is the result of a bug in the knowledge that mediates between specific situations in the case under consideration and the domain-independent deception tactics. More specifically, the model assumes that the tactics are not applicable if any one of the three structural

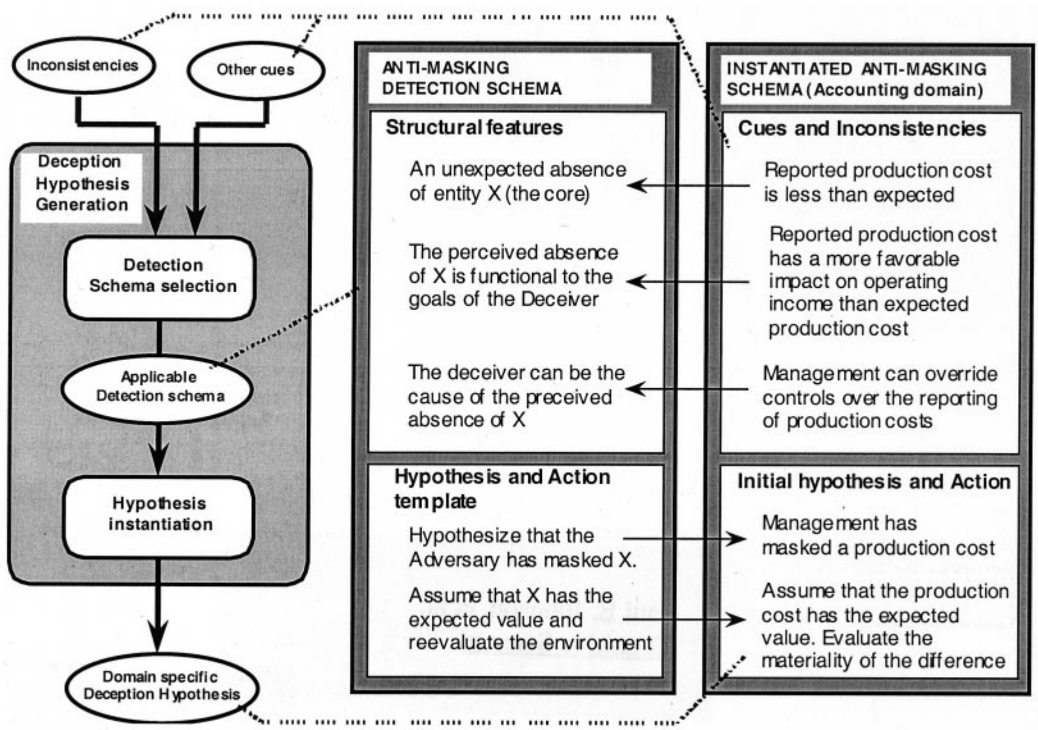

Fig. 2. Instantiation of a Detection Tactic.

conditions (see Fig. 2 in appendix C) is not satisfied: existence of an inconsistency, functionality, or feasibility.

Setting aside the possibility of failure in generating an inconsistency (described above under the heading of bug A), we are left with two possible explanations for Error B. The first is that functionality is not established, perhaps because an adversarial goal is not ascribed to management. The second possibility is that feasibility is not established, perhaps because of a failure in retrieving a possible manipulation of the accounting process that explains an inconsistency. Either possibility will result in a failure to generate the hypothesis that the inconsistency is the result of intentional manipulation. Since we assume that ascribing adversarial goals to management is a higher base-rate activity than retrieving a specific manipulation of the accounting process from memory, we operationalize the failure to apply the tactics by deleting the manipulation from the set of available explanations for the inconsistency.

We assume that when the deception hypothesis is not generated, auditors default on one of the other available explanations for detected inconsistencies: either error in accounting or insufficient disclosure (which one depends on the specific inconsistency as described in Appendix C). The generation of either hypothesis is an over-generalization because the specific information that discriminates between accounting errors and fraud (i.e., the intentionality and the manipulation) is ignored. Accordingly, the hypothesis generation process in the model generates error or disclosure hypotheses to explain inconsistencies whenever a deception hypothesis is not available.

7c. Bug C. Error C refers to the rejection of a correct hypothesis about an identified inconsistency. The five subjects in our subsample committed this error eleven times, sometimes twice in the same case. We suppose that the choice of a criterion to evaluate an inconsistency and its corresponding explanatory hypothesis is the impasse that originates an inappropriate repair.

Accordingly, Bug C is operationalized by forcing the model to use an inappropriate basis for testing the materiality (i.e., magnitude) of an inconsistency: rather than the operating income, the model uses sales revenue or total assets. The model uses sales revenues when the inconsistent cue belongs to the income statement, and total assets when the inconsistent cue belongs to the balance sheet. Since total asset and sales revenues are much larger than the operating income, the impact of each inconsistent cue is comparatively smaller. This causes an underestimation of the significance of the inconsistency and often a dismissal of the corresponding hypothesis.

7d. Bug D. Error D refers to the failure to take into account an accepted hypothesis while selecting a final opinion. We argue that error D is caused by a failure to combine the detected inconsistencies, so that an inconsistency is evaluated in isolation, that is, not in combination with others. S12, S24 and S7 committed nine of these errors in the four fraud cases. Bug D is operationalized as excluding one or more inconsistencies from the aggregation test (Johnson, Grazioli & Jamal, 1992).

Thus far, we have argued that explaining detection performance entails: 1) identifying the knowledge that supports the problem solving task of interest at a level of detail such that a competence model of the solution process can be built, and 2) hypothesizing imperfections that affect this knowledge, based on a theory of failure at the task and the error behavior observed from experimental subjects. In the next section we test this argument by seeding the bugs into the competence model of the solution process and by verifying that the resulting performance model displays responses (inclusive of successes and failures) that match the behaviors observed in the human subjects.

8. Theory evaluation

The proposed theory of success and failure at detecting deception is evaluated by comparing the problem solving outcomes of the 19 auditors in the testing (hold-out) subsample with the predictions from the performance model. The performance model takes as input a case, plus a list of the bugs manifested by a given subject in response to the two major manipulations for that case, and predicts the subject's outcome.[9] The performance model was developed to achieve complete (100%) accuracy in fitting the 20 outcomes (misleading, unqualified+, and unqualified) of the five subjects in the modeling subsample for the four fraud cases. The 19 subjects in the testing subset each completed four fraud cases, resulting in a total of 76 outcomes to be predicted.

9. Error analysis of subjects in the testing sample

The 19 subjects in the testing subset engaged in thinking aloud while working on each of the four fraud cases. The verbal protocols that summarized each subject thinking aloud

Table 5
Errors made by 24 subjects on four fraud cases[a]

Subject (2)	CASE							
	Pharmaceuticals		Surgical products		Big John's		America's family	
	Mailing lists	Inventory	Molds & dies	R&D	Pre-opening costs	Inventory	Inventory	Goodwill amortization
S1*	B	B	B	B	A	B		B
S2*						B		
S3		B			B	B, D		B
S4		B			B	B, C	B, D	B, D
S5		B	B, D	A		B	B	B, D
S6		B	B	B	B, C	B, C, D		A
S7*		B	B			B	B, C, D	B, C
S8	B, C, D	B, C, D			B	B, D		B
S9	B, C, D	B, C, D	B	A		B		B, D
S10		B, C, D	B	B	A	B, C	B, D	B, D
S11		B, D		B	B, D	A	B, D	B, D
S12*	B, D	B, D	B	B	B	B, C	B, C	B, C, D
S13		B	B	B		B		A
S14		B	B	B		B		
S15		A		A		B	B	B
S16	B, C	B, C				B	B	B, D
S17	B, D	B, D	B, C, D	A		B	B, C	B, C, D
S18			B	A		A	B, C	B, C, D
S19					B	B, D	B	A
S20		B	B			B	B, D	A
S21			B, D	A	B, C	B	B, C, D	B, C, D
S22		B	B	B	B	B, D		A
S23	B, D	B, D	B, C, D		B	B	B, C	A
S24*	B, C, D	B, C, D	B, C, D	A	B, C, D	B, C, D	B, C	A

* These five subjects were used in the model development phase.
[a] Letters in cells indicate types of errors. A = activation error; B = hypothesis generation error; C = hypothesis evaluation error; D = assessment error.

activity were analyzed for the presence of errors following the procedures discussed in Appendix B. Table 5 summarizes the errors observed in the responses of each of these 19 subjects (plus, for completeness, the 5 subjects shown in Table 4).

As shown in Table 5, approximately half the subjects committed at least five errors (and less than ten). The other one-half committed more than ten errors. The greatest number of errors (19) was committed by S24. No subject was error free across cases, although some subjects completed some cases without committing any of the four types of errors identified in our error analysis. Subject S2, who performed successfully in all four cases, manifested only one error (Error B in Big John's–See Table 5). Five subjects concluded at least one case successfully and without errors.

Each type of error (i.e., A, B, C, and D) occurred in each of the four cases. Activation Error A was the least frequent: 19 instances were observed, out of 192 opportunities (24 subjects x 4 cases x 2 cues). The comparatively low frequency of Error A is consistent with our theory. The knowledge that is employed by the subjects to identify inconsistencies for the

Table 6
Prediction accuracy of the performance model

Samples	Fraud cases				Accuracy across cases
	Pharmeuticals (medical)	Surgical products (medical)	Big John's (retail)	America's family (retail)	
Testing subsample: 76 observations (19 subjects on four cases)	84%	79%	95%	79%	84%
Total sample: 96 observations (24 subjects on four cases)	87%	83%	96%	83%	87%

purpose of the detecting fraud is the same knowledge that it is used to identify inconsistencies for the purpose of assigning other, more frequent outcomes. As a result, subjects have a better chance to refine it.

Hypothesis generation Error B was the most common. The 24 subjects had 173 opportunities to commit Error B in the four cases (192 instances, minus 19 cues that were not identified as inconsistent and therefore did not require an interpretation). In 114 instances (66% of the time) subjects made Error B. This relatively high frequency is consistent with the explanation of errors as overgeneralization of hypotheses selected to interpret identified inconsistencies. All subjects made at least one Error B; some made as many as six out of a possible eight opportunities (two cues x four cases).

Thirty-two evaluation Errors C and 43 global assessment Errors D were observed in our sample. The intermediate frequency of these errors is further evidence that the subjects are less than perfectly adapted to the fraud detection task.[10] The total number of errors observed across all subjects and cases was 208.

10. Comparison of model and subject outcomes

Results of comparison between the outcomes of the performance model and the outcomes selected by each subject on each of the four fraud cases are summarized in Table 6, which shows that the model correctly predicted 84% of the outcomes in the testing sample (misleading, unqualified+, and unqualified). Overall, the model accurately predicted 87% of the outcomes in Table 1. The measured accuracy of the model (Cohen's Kappa) after correcting for chance agreement is 0.83 (Cohen, 1960). Table 7 details the outcomes for subjects on each of the four fraud cases compared with outcomes of the model. In what follows we examine the mismatches between the subjects and the model in each case.

10.1. Pharmaceuticals

On this case the model failed to correctly predict the outcome of three out of the 19 subjects in the testing subsample. As shown in Table 7, the model predicted an unqualified opinion for S16, who instead rendered an unqualified+ opinion. The model based its prediction on the fact that S16 failed to correctly evaluate the materiality of the (wrong)

Table 7
Comparison of subject and performance model responses on four fraud cases*,[a]

Subjects	Cases			
	Pharmaceuticals	Surgical products	Big John's	America's family
S1*	M	M	M	M
S2*	M	M	M	M
S3	M	M	U+	M
S4	M	M	U+	U/U+
S5	M	U	M	U+
S6	M	U+	U	M
S7*	M	M	M	U+
S8	U	M	U+	M
S9	U	U+	M	U+/M
S10	M	U+	U+/U	U+
S11	U/M	M	U	U/U+
S12*	U+	M	U+	U
S13	M	U+	M	M
S14	M	U+	M	M
S15	M	U/M	M	M
S16	U+/U	M	M	U+
S17	U	U+/U	M	U
S18	U/M	U+	M	U
S19	M	M	U+	U+
S20	M	U+/M	M	U
S21	M	U	U+	U
S22	M	U+	U+	U/M
S23	U+	U+/M	U+	U
S24*	U	U	U	U

* These five subjects were used in the model development phase.

[a] M = misleading; U+ = Unqualified opinion with additional paragraph; U = Unqualified opinion. Model responses match subject responses in all instances except those with a double entry (the letter above the line is the subject response; the letter below the line is the model response).

hypotheses generated to explain the cues in the case (Error C in Table 5). Other subjects, in the same situation, rendered an unqualified opinion, as did the model.

The model also incorrectly predicted success for two subjects (S18 and S11) who actually failed to detect the fraud. The model predicted success (detection) for S18 because this subject did not manifest any of the four types of error in his protocol for the case. Despite the absence of errors, S18 rendered an unqualified opinion (Table 7).[11] The model predicted success for S11 because of his flawless performance on the mailing list cue. The sufficiency of correct processing of this cue for achieving success is demonstrated by the process trace of S15, who succeeded in evaluating Pharmaceuticals despite missing the other cue (Table 5). S11 manifested errors B and D on the inventory cue, which suggests possible interference between correct (mailing list) and incorrect (inventory) processing.

10.2. Surgical products

The model failed to predict the outcome for four subjects in this case. S17 issued an 'additional paragraph' to his opinion, as opposed to the model's 'unqualified' prediction. The model's prediction was based on the fact that S17 completely missed the R&D cue and committed three errors (B, C and D) on the molds and dies cue. By contrast, S5, who missed the R&D cue and made fewer errors (only B and D), concluded his review with an unqualified opinion. The last three subjects gave unqualified (S15) and unqualified+ (S20, S23) opinions despite the fact that their protocols did not reveal errors in processing at least one of the two main cues for the case.

10.3. Big John's

The model failed to predict the outcome for one subject (S10). S10 missed the inventory cue and failed to appreciate the materiality of the other main cue. On this basis, the model predicted an unqualified opinion, while the subject chose an unqualified+ opinion.

10.4. America's family

As shown in Table 7, the model failed to predict the outcome of four subjects (S4, S9, S11, and S22) on this case. S4 and S11 rendered an unqualified opinion, not the predicted unqualified+. In making this prediction, the model is consistent with the behavior of S10, who, after committing the same errors as S4 and S11, rendered an unqualified opinion (see Tables 5 and 7).

The model also predicted success for S9 and S22, based on their performance on the inventory cue. The inventory manipulation in this case is so substantial that an errorless handling of this issue is considered sufficient reason to predict successful detection. S13, for instance, succeeded on the basis of the inventory cue after missing the other cue (goodwill). Contrary to the prediction, both S9 and S22 failed to detect the fraud.

Previous work (Johnson, Grazioli & Jamal, 1992; Johnson, Grazioli, Jamal & Zualkernan, 1992) has suggested that some auditors occasionally render a "misleading" final outcome without generating the hypothesis that a financial deception has occurred. Such subjects typically generate hypotheses of error in accounting to explain inconsistencies that are, in fact, symptoms of fraud. This type of response occurred four times in our 96 tasks (i.e., 24 subjects x 4 cases): Subject S1 in the Pharmaceutical and Surgical Products cases, subject S15 in the American Family case, and subject S12 in the Surgical Products case. The verbal protocols of these subjects reveal that they combined hypotheses of error to conclude that the financial statements of the examined company were misleading. This phenomenon suggests redundancies in subjects' knowledge that may protect them from falling prey to a deception despite not being able to generate the hypothesis that deceptive manipulations have occurred. As described in Appendix C, such knowledge is included in our competence model, although it is not covered by the proposed theory.

Analysis of model and subject behavior on the four fraud cases shows that prediction failures are fairly evenly spread across type of subjects and cases. In eight out of the ten instances where subject and model outcomes are discrepant, the model chose a more severe outcome than did the subject. In most cases, modifications of the model's predictions for the purpose of reducing the identified mismatches would reduce the fit of the model outcomes with the opinions of other subjects in the sample. In the remaining cases, such modifications would comprise ad hoc rather than principled alterations of the theory on which the model is based (Lewandowsky, 1993; McCloskey, 1991).

Overall, the ability of the model to predict success and failure in fraud detection on the cases shown in Table 1 is quite remarkable. It is difficult to detect regularities in surface features of the data presented in Table 1. Our results are consistent with the proposition that such regularities in fact lie in the deeper knowledge for detecting the manipulations (deception) created by another agent, plus perturbations in the knowledge of the domain on which the manipulations are based.

10.5. Subjects who failed on the clean cases

All of the failures we observed in the performance of subjects on the fraud cases shown in Table 1 were 'miss' failures (i.e., failures to identify a financial deception when one is present). The theory of failure presented here does not consider 'false alarms' (i.e., identifying a financial deception when none is present). Accordingly, the bugs we proposed to model errors made by subjects share the characteristic of leading the model astray (i.e., from the path to a correct detection). Such bugs do not explain why a subject took that path when it was not warranted. We speculate that just as auditor's knowledge of accounting contains perturbations (bugs) that distort behavior on fraud cases, there are other distortions that are responsible for failure on clean cases. We did not model these failures in the present study.

Most subjects (17/24) correctly classified the clean cases as free from problems. However, seven subjects chose a different opinion (see Table 1). Of these, S16, S5, and S9 chose to add a paragraph to their unqualified opinion due to the presence of an unresolved litigation. S2 chose to add a paragraph due to the lack of consistency in the treatment of pension expense. S1 and S10 incorrectly conclude that the Vascular Product case had serious auditing problems; both based their conclusions on a variety of cues they considered anomalous. S1 and S10 incorrectly issued a 'non-GAAP' opinion based on inventory, intangibles, litigation, revenue recognition, taxes, and accounts receivable. In fact, these cues do not represent departures from GAAP.

The fact that subjects failed to detect manipulations in fraud cases three times as often as they erroneously detected manipulations in clean cases is consistent with the argument that auditors have better knowledge for evaluating clean cases, since they process more of them. However, this knowledge is also the basis for unwarranted generalizations that occasionally lead to failures.

11. Knowledge bugs or lack of information: an alternative explanation

In previous sections we have argued that auditors fail to detect financial deception because their domain knowledge contains bugs. An alternative explanation–rooted in the theory of decision-making under uncertainty (e.g., Fox & Tversky, 1998)-is that subjects interpret the inconsistencies in a given case based on uncertain choice among alternative interpretations (i.e., an inconsistency can be either evidence for an error or an intentional manipulation).

This explanation does not require a bug in auditors' knowledge: lacking sufficient information, the auditors make a judgment call, which in many cases turns out to be incorrect, hence the failure to detect fraud.

To assess the plausibility of the uncertainty-based explanation for our results, we examined the auditors' ratings of the appropriateness of their opinions. As described above, after completing evaluation of all cases, each auditor was asked to assign a rating as to the appropriateness of each opinion on each case. For each auditor and each case, we identified the 'dominant' opinion, that is, the opinion the auditor felt to be most appropriate, as well as its runner-up (next most appropriate opinion).

According to both explanations, the auditors' knowledge is well adapted to the task of reviewing clean cases, with which they have frequent and repeated experience. When they encounter a clean case, they should classify it appropriately and confidently. The two explanations differ, however, in predicting what would happen when the auditors are faced with the inconsistencies that characterize a fraudulent case.

According to the uncertainty-based explanation, auditors become uncertain about alternatives, look for additional information, and as a result, should express uncertainty regarding the appropriateness of the opinion they render. Accordingly, we would expect to observe that ratings of the auditors' dominant opinion on the fraud cases would be lower than the ratings of their dominant opinion on the clean case, and similar to the ratings for its runner-up (the most likely alternative). By contrast, the knowledge-bug explanation predicts that auditors' ratings of their dominant opinion for the fraud cases will be 1) no different from ratings of their dominant opinions on the clean case, and 2) significantly higher than ratings for its runner up in all cases.

In general (across cases) the auditors rated (on a 1–7 scale) their dominant opinion as highly appropriate (mean rating of 5.64, std. dev. = 1.03). A multivariate ANOVA using a within-subjects repeated-measures (across cases) design and three multivariate tests (Pillai's trace, Hotelling's trace and Wilk's lambda) consistently rejected the hypothesis that the case affected auditors' rating of the appropriateness of their dominant opinion ($p = 0.221$). Paired-sample contrasts testing the difference between ratings assigned to each fraud case and the clean case were also not significant.

A non-parametric test was additionally used to evaluate the hypothesis that ratings for the dominant opinion on each case were sampled from different distributions. A Friedman test rejected the hypothesis that the case affected the ratings ($p = 0.353$), which reinforces the conclusions reached by the multivariate ANOVA.

Across cases and subjects, the average difference between the auditors' rating of their dominant opinion and its runner up was 2.3 (std. dev. = 1.71). Case-by-case, paired-sample t tests of the differences between judgment of the dominant opinion and its runner-up revealed that on average the difference between the dominant opinion and its runner-up was positive and highly significant ($p < 0.000$).

We interpret these results as strong support for the knowledge-bug hypothesis and, correspondingly, as disconfirmation of the alternative explanation that auditor choice under uncertainty based on lack of information is responsible for the findings of our research.

12. Conclusion

We have investigated how a group of highly trained individuals succeed and fail at solving the problem of detecting deceptions in a specific adversarial setting. To explain success we have proposed that these individuals apply a variation of Dennett's intentional stance. In our case, this strategy is based on a set of heuristics-the detection tactics-that interpret inconsistencies between expectations and observations in the light of the Deceiver's goals and possible actions. We have argued that such tactics are: 1) developed from experience with deception across a variety of domains, and 2) available for use in specific circumstances such as the auditing of a company's financial statements.

Our explanation of failure to detect financial deceptions integrates previous empirical work on deception (Johnson, Grazioli & Jamal, 1993), a theory of information processing errors (VanLehn, 1990), and a process model of the detection of financial statements fraud. We have argued that the detection knowledge used by subjects in our research incorporates specific knowledge of the domain where the deception occurs. However, this domain-specific knowledge contains perturbations that distort the processing of information used to identify the manipulations that have taken place. These perturbations are the result of incorrectly applying what has been learned in the context of fairly presented information to problems in which an adversarial agent has manipulated information with the intent to deceive.

Specific hypotheses about the nature of these perturbations (knowledge bugs) were derived from a comparison of the behavior of subjects with the behavior of a model of successful detection. The comparison allowed us to identify the localized errors made by the subjects in the detection task. The distribution of these errors was consistent with the proposed theory. We also discovered that success is compatible with making errors. No subject-even the successful ones-was flawless across all cases. Our work shows that errors interact in non-obvious ways to create complex patterns of behavior that are unstable, yet susceptible to principled analysis.

The theory we propose was evaluated by constructing and testing a model of performance in a real-world task that includes the detection of a specific form of deception (financial statement fraud). The model, which contains the hypothesized perturbations, was used to predict the behavior of a sample of individuals whose organizational responsibilities include the detection of financial deception. The model matched 87% of the subjects' outcomes (84 out of 96 observations).

Despite its success in being able to predict the behavior (success and failure) of the 24 subjects in our study, there are significant limitations to the theory we have constructed. Chief among these is the fact that we did not address the question of what initiates the use of detection knowledge in the presence of specific cues. We have assumed that detection tactics are routinely summoned by subjects as part of the process of explaining an inconsistency. When the conditions of a given tactic are satisfied, the hypothesis that a deception has occurred is generated.

There may, however, be specific cues that initiate use of the detection tactics. Cosmides & Tooby (1992; 1994) and others (e.g., Gigerenzer & Hug, 1992; Platt & Griggs, 1993), have argued that prompts to look for cheaters (including Deceivers) in the context of a social contract can have a major effect on subjects' performance. The presence of such a prompt may generate a greater success rate for the task of fraud detection than we have observed.

We have also not modeled the process by which bugs are manifest in the behavior of a given subject for a particular case, or how bugs develop in a given subject's experience. The learning history of each auditor is populated with a variety of experiences, varying from textbooks, case materials and training conducted by firms, to individual experience with clients and discussions with colleagues. In all of this there are opportunities for the (over) generalizations described here to occur, despite considerable motivation to perform effectively in situations where a deception may be present. Studies of the learning histories of subjects at various stages of professional development are needed to support our argument for the origin of perturbations of domain knowledge that we have called knowledge bugs.

Finally, recent work on the concept of mindreading and theory of mind (Baron-Cohen, 1995; Cosmides & Tooby, 1994; Cosmides & Tooby, 1995; Leslie, 1987; Wellman, 1990; Whiten & Perner, 1991) suggests that there may be a more general formulation for the process we have described by means of the deception detection tactics. Although we have postulated a process (ascription) through which auditors assign goals to agents who have manipulated the environment they must evaluate, we have not extended the argument for this process to the prediction of the behavior of agents outside the context of our audit cases. Such a step would strengthen our argument that the successful auditor detects deception using knowledge developed in other domains (i.e., not accounting). Studies in which auditors are given a variety of detection tasks would provide important convergent evidence for the arguments presented here.

An interesting practical consequence of our work is that fixing individual bugs may not be the most effective means to improve auditor performance. Auditors who succeed consistently in fraud detection are not necessarily free of imperfections in their knowledge. The fact that successful performance does not require flawless knowledge suggests that the operative knowledge of the individual auditor contains redundancies that make it robust to the presence of small distortions.

From a scientific perspective, the work presented here provides insight into a complex, real-world, task in which success does not seem to depend directly on specific experience (Johnson, Zualkernan & Tukey, 1993; Zualkernan & Johnson, 1992). Although in our data many who attempt the detection task fail, some persistently succeed. We have argued that even those who fail may have the knowledge needed for success. However, such knowledge is obscured by imperfections derived from the context in which given instances of the task occur. More intensive study of tasks such as the one investigated here may provide insight into powerful forms of knowledge that are at present poorly understood, yet underlie performance that is more common in everyday life and the world of work than we have previously supposed.

Notes

1. The strength of this treatment administration schedule is that it keeps order effects constant, so that process comparisons can be made across subjects. Its weakness is that possible order effects and learning effects are confounded. A statistical analysis was performed to determine the presence of case order effects on auditor perfor-

mance. Both the Scheffe and the Tukey tests of the differences among group means rejected the hypothesis that the performance (success or failure) of the 24 auditors in the sample is different across the four cases (alpha = 0.05). In addition, a Pearson coefficient of correlation between performance and order of presentation was non-significant and small and negative in value. At the level of the overall sample there is no evidence for order effects. This conclusion does not, however, preclude order effects for individual auditors.

2. Strictly speaking, the information in the financial statements is not sufficient to accuse a company of fraud in accordance with the law and accounting standards. However, that is not the purpose of the second partner review task, which rather is to control for the quality of statements before they are issued by an auditing firm. When the second partner disagrees with the opinion of the engagement partner, the auditing firm is likely to withhold the statements until the issues are resolved. Issue resolution may involve re-examining the auditing working papers (the result of the auditing field work). The working papers were not provided as part of our experimental materials.
3. To verify the proposed aggregation of subject responses into three categories (i.e., misleading, unqualified+ and unqualified) we performed a cluster analysis (Norusis, 1988) of the ratings issued by the subjects on the appropriateness of each of the six different audit opinions. If we constrain the number of clusters to three, the cluster analysis generates exactly the three classes we propose.
4. The Target may also simply recognize a given deception as an instance of one she has seen in the past. However, the low frequency of occurrence of financial statement fraud make this detection-by-recognition unlikely.
5. In the theory of fraud detection proposed here we identify classes of potential manipulations using a model of the accounting process found in accounting textbooks (e.g., Nikolai & Brazley, 1992), plus knowledge of the manipulations that occurred in the specific case materials given to subjects. The model describes the accounting process as consisting of three steps: 1) classification, 2) valuation, and 3) disclosure of economically significant events and transactions (items). According to the model, item classification consists of assigning each item to a meaningful accounting category. Once the item category has been assigned, accounting procedures to associate a value with the item can be selected. At times, the selection of these procedures allows for a considerable degree of freedom. The actual vauation process generates a value for each item. The details of this computation, the procedures used and the final value may then be disclosed in the financial statements. An example of an economically significant event from the cases comprising our experimental materials is the aging and wearing of tools due to use and time. Such tools need to be classified as assets and depreciated over their useful lives. On the basis of this classification, rules are selected to compute depreciation. The computation is then carried out, and its results disclosed (e.g., Nikolai & Bazley, 1992). Asset depreciation may be affected by a double play deception tactic, which specifies manipulations that maliciously declare the deception core so that the Target explains it away. In the specific case of asset depreciation, overtly declaring an increase in the useful lives of assets is a plausible manipulation the decreases depreciation expense and increases reported income.

6. The model was implemented using the CLIPS environment (Giarrantano & Riley, 1989); it contains 72 production rules. Its working memory contains frames (sets of property-value pairs) that represent the cues, their properties and values.
7. Surgical Product booked inventory held on consignment by others as sales. When the unsold goods were returned to the company, they were treated as sales returns.
8. When more than one opinion is supported by the data of a case, the model concludes the more severe opinion (e.g., when both misrepresented and unqualified+ are supported, the program evaluates a misrepresentation).
9. A methodological alternative is to develop models of auditor types (as in Johnson, Grazioli, Jamal, & Zualkernan, 1992). These models would take as input the four fraud cases and for each case produce an outcome (misleading, unqualified+, or unqualified). However, examination of subjects responses in Table 1 reveals that 21 different models of this sort would be required to fully account for the behavior of the 24 subjects. Only three subjects, in fact, have perfectly overlapping sets of (recategorized) outcomes for the four cases. Based on these considerations we decided to model subject/case interactions. This approach modifies (parameterizes) the competence model as a function of the bugs inferred from the behavior of each subject in processing the two major manipulations of each case. Since the behavior with respect to an individual cue can be characterized in nine different ways (e.g., no errors; error A alone; error B and error D, etc.), it follows that the behavior of a subject with respect to two cues is characterized in 81 different ways, corresponding to the different combination of errors (9×9) made with respect to each cue. Each of these 81 different combinations corresponds to a specific set of bugs, which are inserted in the competence model. Although the total number of possible models (81) is large, this approach turns out to be more parsimonious since we have, in fact, an efficient measure of the knowledge manifested by each subject on each case. This measure is the set of errors committed by a given subject on each case (e.g., errors B, C, D for one cue, and error A for the other cue). Hypotheses about specific bugs allow us to convert errors into hypotheses about the subjects knowledge, and in turn, into specific changes in the competence model. Once we have this knowledge and a given case, we are able to predict individual subject success and failure at the fraud deception task (Table 1) with remarkable precision.
10. To evaluate the stability of the knowledge that auditors bring to bear to the task (irrespective of how correct or incorrect the knowledge is) we developed a measure of error-pattern variability across cues and cases. For each subject, we first determined a modal pattern of errors across the eight critical cues (e.g., no A, B, no C, D-see Table 5). We then counted the differences between the modal pattern of errors and each of the eight patterns manifested by the subject on each of the eight critical cues in the cases. A subject that exhibited the same pattern of errors across all cues received a count of zero; a subject that exhibited maximum variability in the patterns of errors received a count of 16, which is the combinatorial maximum for four errors and eight cues. The measure we computed is the ratio of the count of the observed differences over its combinatorial maximum. We assume that a ratio equal to zero indicates high stability of a subjects knowledge, and a ratio equal to 1 indicates high

instability. For instance, the ratio of subject S2 (1/16) suggests high knowledge stability when compared to the ratios of subjects S17 and S21, who had the highest variability in our sample (10/16). On average, subjects had a ratio of 6.4/16. Correlation analysis concluded that detection success and error-pattern variability are negatively correlated (Pearlson coeff. $= -0.64$, $p < 0.001$). In theoretical terms, unsuccessful individuals tend to repair their impasses in unstable ways.

11. S18 is the only subject that did not commit any of the errors tracked in this study and selected an unqualified opinion. Six more subjects appeared to perform at a level sufficient to succeed (Table 7), but did not. This behavior might be the result of an unwillingness by these auditors to criticize the opinion of a colleague, or the unwillingness to charge a client with a serious misrepresentations of their financial statements. False alarms are costly; loss of the client is likely.
12. A detailed analysis of the cases comprising our experimental material revealed evidence of other deception tactics (i.e., decoying, inventing, dazzling and mimicking; Johnson, Grazioli & Jamal, 1993). These manipulations did not affect the main cues in the cases.
13. By contrast, the criterion for generating inconsistencies is 5% of the reported operating income. Because of this smaller threshold, some detected inconsistencies led to the generation of initial hypotheses that are later discarded by the hypothesis evaluation process, which applies a 10% criterion. This feature of the model simulates the behavior of the auditors that appeared to consider some cues as inconsistent, yet fell short of generating explicit hypotheses about them.
14. When the auditor knows that the financial statements are misleading because they are not prepared in conformity with GAAP, she must issue either a Non-GAAP opinion or an adverse opinion, depending on the materiality of the item in question. If the departure from GAAP has a material effect on the financial statements, the auditor is required to issue a (Non-GAAP) qualified report. Auditing guidelines suggest that a misstatement of 5 to 10% of net income before taxes should be considered material. In the current paper, we use 10% of net income before tax (operating income) as a base for determining materiality of a hypothesis, and 5% as a criterion for identifying inconsistencies. If the departure from GAAP is "highly material" such that the financial statements are pervasively misstated, then the auditor is required to issue an adverse opinion. Auditing standards do not explicitly define what makes an item highly material. The concept "highly material" suggests a much higher magnitude of misstatement than 10% of net income before tax. The concept of pervasively misstated suggests more than one individual misstatement. We operationalize the concept of "highly material" as being any item or combination of items that exceed 40% of net income before taxes (operating income).

Appendix A: case summaries

The manipulations in each fraud case were identified on the basis of articles in the financial press and published reports of the Securities and Exchanges Commission (Accounting and Auditing Enforcement Releases). We call the effects of these manipulations on the cases ‘cues,’ because such cues can be used to generate the hypothesis that a fraudulent manipulation has taken place.

In this research we focus on the processing of the two most significant cues in each case. This focus was determined based on the results of a preliminary analysis of the relationship between information processing errors and fraud detection performance (Johnson, Grazioli, Jamal & Zualkernan, 1992b). The analysis suggested that success and failure in processing the two major manipulations in each case was sufficient to explain more than 80% of the variability in the behavior of subjects (auditors engaged in the task of fraud detection). A description of the cases and the cues follows.

Pharmaceuticals, Inc. is based on a company called Generics Corporation of America. The company was the subject of an SEC enforcement release. The SEC objected to two substantive issues in this case. First, the company had capitalized an internally generated asset called a mailing list, even though the generally accepted accounting principles (GAAP) require such items to be expensed. Second, the company failed to write off obsolete inventory. The company’s drug inventory was difficult to sell or not legal to sell because it was outdated.

Surgical Products, Inc. is based on a company called U. S. Surgical Corporation. This company was also the subject of an SEC enforcement release. The SEC identified two critical issues: operating expenditures were improperly capitalized as molds and dies, and R&D expenditures were improperly capitalized. The SEC enforcement release raised several additional issues: goods shipped with right of return were recorded as sales, and inventory values were inflated by valuation based on inappropriate prices.

Big John’s Electronic Sales, Inc. is based on a company called Crazy Eddy Inc., which was also the subject of an SEC enforcement release. The case contains two major issues. First, pre-opening costs were capitalized and then amortized over a five-year period-a very questionable practice. Second, inventory was overstated due to recording of fictitious items.

America’s Family Video Store is based on a company called Blockbuster Entertainment. The case has two major issues. The first issue involves questions regarding the propriety of a change in depreciation policy for videocassette rental inventory. The second issue involves amortization of Goodwill over a long period of time (40 years), which may not be realistic for a video cassette rental business. There was a debate among several financial analysts regarding the propriety of these accounting methods in the financial press (Wall Street Journal, Forbes, and Fortune). The company’s stock price declined substantially as a result of this controversy, which was also a source of embarrassment for the company’s independent auditors.

Appendix B: coding process

The procedure used to code subjects’ protocols was conducted by individuals knowledgeable in the domain of auditing and familiar with the case materials. The first step in this procedure consisted of separating protocols into episodes defined by the processing of one or a small number of cues in the task materials (e.g., inventory, or R&D). Episodes were then classified into one of ten coding categories (C1 through C10) based on episode content (Jamal, 1991).

The first two coding categories (C1 and C2) reflected either the intentional violation of Generally Accepted Accounting Principles-GAAP (i.e., principles that specify how to fairly represent a company (C1)) or violation of GAAP without explicit comments about intentionality (C2). The next three categories (C3 through C5) were based on verbalizations that expressed the possibility of a mistake in applying GAAP to a specific transaction or event (C3), the presence of uncertainty about the values of certain transactions and events (C4), or the presence of uncertainty about the ability of a company to continue its normal operations as a going concern (C5).

The last five categories (C6 through C10) were based on verbalizations about the presentation or disclosure of specific items in the financial statements. C6 was used if the subject verbalized that presentation or disclosure of financial information lacked consistency in the application of accounting policies. C7 was used if the subject suggested that disclosure was insufficient. The other three categories were used when the subject identified or attempted to characterize deficiencies found in the financial statements based on the generation of a representation of the company (C8), the identification of an item that violated the subject's expectations (C9), or operating conditions used to explain violation of an expectation (C10).

Coding categories were augmented by recording two additional pieces of information for each subject: 1) whether the subject evaluated the magnitude (materiality) of an identified violated expectation, and 2) whether the subject combined available evidence to reach a final outcome.

Reliability of the coding scheme was assessed by repeating the coding with a second individual (knowledgeable in the domain of auditing) and calculating an appropriate statistic of interrater agreement. Due to the large volume of protocols, a stratified random sample of protocols was selected for purposes of the reliability analysis. The sample contained at least one protocol for each of the 24 subjects in the sample.

The average agreement between coders was 90% for cue identification, 74% for cue interpretation, and 84% for subject's final conclusions. Interrater reliability coefficients (Cohen's K-Cohen, 1960) ranged from 0.71 to 0.80 ($p < 0.0001$). Differences in scoring were resolved by discussion between the two coders. The results of the reliability analysis were within the range commonly found in the field (Ericsson & Simon, 1993), and similar to what has been obtained in investigations of problem solving in auditing and other professional domains (e.g., Johnson, Duran, Hassebrock, Moller, Prietula, Feltovich & Swanson, 1981; Johnson, Grazioli, Jamal & Zualkernan, 1992).

The final step in the protocol coding procedure generated error codes (Error A, B, C, and D) for each subject on each case. For each protocol, the episode involving the two major manipulations in each case (listed in Appendix A) were identified and evaluated using the following rules:

Rule A: If a cue was not coded according to the procedures described above, it was inferred that the subject did not heed it (Error A-activation error).

Rule B: Coding categories C3 through C10 were considered evidence for hypothesis generation errors (Error B). More specifically, coding categories C3 through C5 were considered evidence that the subject generated an "error in accounting" hypothesis. Categories C6 through C10 were considered evidence that the subject generated an "insufficient disclosure" hypothesis. Both hypotheses are incorrect.

Rule C: If a cue was coded as non-material according to the procedures described above, a hypothesis evaluation error was inferred (Error C).

Rule D: If the subject did not combine an identified cue with other available evidence, a global assessment error was inferred (Error D). This was done regardless of the correctness of the specific outcome reached by the subject.

The error coding shares the same reliability of the protocol coding, since the four rules do not require coder's judgment.

Appendix C

The competence model implements the proposed theory and is rooted in previous work (e.g., Johnson et al., 1992). The model is composed of four processes: activation, hypothesis generation, hypothesis evaluation, and global assessment.

Activation: The process of activation generates expectations for *cues* in a set of financial statements (e.g., inventory), compares each cue with its expectation and, based on this comparison, generates *inconsistencies* (e.g., 'the inventory level is too high'). Inconsistencies are symptoms of possible anomalies in the accounting process that produces the financial statements. Cues are categories of information that could be used as symptoms of manipulation in the cases shown in Table 1 (see Appendix B for a list). Cues were identified with the help of a domain expert familiar with the cases.

For each cue the activation process generates a numerical expectation. These expectations are represented by simple linear functions of past values of the cue, elaboration of other cues (e.g., computation of rate-of-change), and industry constants. As an example, the expected value of inventory is computed by multiplying the past value of inventory times the rate of growth in sales.

Each cue and its expectation are compared. Whenever this difference is larger than a specified criterion value, the cue value is labeled as inconsistent. The criterion value assumed here (5% of reported net income) is derived from previous work on studies of auditor behavior and from professional guidelines. Additional examples, as well an analysis of the procedures for generating expectations and comparing them to the actual values can be found in Johnson, Grazioli and Jamal (1992).

Hypothesis generation: Cues inconsistent with expectations need to be explained. Hypothesis generation is the process (see Fig. 1) that generates initial hypotheses to explain inconsistencies detected in the financial statements for each of the four fraud cases.

Inconsistencies are explained by unexpected variations in the economic conditions underlying the company's reported financial data (e.g., an unexpected surge in market prices), or by anomalies in the accounting process that produced these data. Here we focus on the second type of explanation. We assume that inconsistencies admit three alternative types of explanation: 1) manipulations of the accounting process, 2) errors in accounting, and 3) insufficient disclosure. Correspondingly, the output of the hypothesis generation process is the hypothesis that, 1) a financial deception has been attempted (deception hypothesis), 2) an error in accounting has occurred (accounting error hypothesis), or 3) disclosure of financial data are insufficient (insufficient disclosure hypothesis). The hypothesis generation process assigns one of these three hypotheses to each detected inconsistency.

Hypotheses about deception are generated by the detection tactics. Each detection tactic is represented as a schema containing conditions that are matched against features of detected inconsistencies, plus features of other cues in the financial statements under examination (see Fig. 2). As described above, each schema includes three conditions: existence of an inconsistency, functionality, and feasibility. The deception hypothesis is generated only when all conditions of a schema are satisfied.

Hypothesis generation begins by applying each detection schema to each inconsistency. If the conditions in the schema are satisfied, a deception hypothesis is generated for that inconsistency. If no schema applies, then either an insufficient disclosure hypothesis or an accounting error hypothesis is generated. An insufficient disclosure hypothesis is generated when the inconsistency cue is related to a note to the financial statements (e.g., a change in accounting policies); otherwise, an accounting error hypothesis is generated as a default.

The existence of an inconsistency is the condition that triggers the schema instantiation process. The second condition (functionality) is satisfied when the triggering inconsistency contributes to the achievement of one of the goals ascribed to management. To establish which goal to ascribe, the model examines the classification of inconsistent cues in the financial statements. If a given inconsistent cue is a revenue (an expense) and is larger (smaller) than expected, then the possible goal of overstating income is assigned to management. If it is an asset (a liability) and is larger (smaller) than expected, then the possible goal of overstating assets or owner's equity is generated.

The third condition (feasibility) is satisfied when a plausible manipulation that explains the inconsistency is available. We derive a set of specific manipulations from analysis of the deception tactics, existing literature (e.g., Schilit, 1993), and knowledge of the manipulations contained in the cases shown in Table 1. For the cases in Table 1, this set of tactics consists of double play, repackaging and masking manipulations (see Table 2).[12]

Manipulations are associated with specific cues on the basis of categorization of cues in the financial statements (e.g., as an asset, or expense). Overt changes in accounting policies reported in notes to the financial statements are assumed to be symptoms of double-play manipulations, arguably because management disclosed the manipulation in an attempt to induce the auditor to dismiss it. Inconsistent asset balances are assumed to be symptoms of repackaging manipulations, in which the asset is intentionally misclassified to obtain a more favorable accounting treatment. Finally, inconsistent expense balances are assumed to be symptoms of masking manipulations, in which management either failed to record the expense, or introduced a malicious error in its computation.

The output of the hypothesis generation process is an initial (i.e., subject to evaluation) set of hypotheses. These hypotheses are the input to the next subprocess, which evaluates them.

Hypothesis evaluation: The initial hypotheses generated by the model need to be confirmed. The process of hypothesis evaluation produces confirmed hypotheses on the basis of the materiality of the inconsistency associated with each hypothesis. The type of hypothesis suggests the nature of the test for materiality for each inconsistency. Error and disclosure hypotheses are evaluated on the basis of the magnitude (a dollar value) of the difference between the value of the inconsistent cue and its expectation.

Detection tactics contain guidance about what to do to test the hypothesis that a deception has been attempted (see Fig. 2). The antimasking tactic, which responds to manipulations based on erasing the core of the deception (say, a cost) from the auditor's environment, prescribes re-evaluating the environment under the provisional assumption that the core is present (the cost was sustained and not reported). The antidouble play tactic prescribes not explaining away weak symptoms of anomaly just because they are explicit, or because they are individually weak. The antirepackaging tactic prescribes recategorizing the items that may have been miscategorized under a worst-case scenario assumption.

Accordingly, a measure of the magnitude of the inconsistency (its dollar value) is computed following the above prescriptions. For the anti-double play hypothesis the entire dollar amount of the inconsistency is considered anomalous. For anti-masking and anti-repackaging only the difference between the expectation and the actual value is considered anomalous.

After these dollar values are assigned, the initial hypotheses are individually evaluated against a criterion for materiality. From auditing textbook guidelines and previous research (Arens & Loebbecke, 1993; Johnson, Grazioli & Jamal, 1992), we adopt a criterion for materiality for individual hypotheses equal to 10% of net operating income.[13] Hypotheses whose associated value is higher than the criterion are labeled 'material' and considered accepted.

Global assessment: Material hypotheses about individual inconsistencies need to be composed in a final assessment. The global assessment process proposed here aggregates and evaluates confirmed hypotheses to produce an outcome that expresses the extent to which the financial statements of the company are a fair representation of its underlying economic condition. The output of the global assessment process is one of the three possible outcomes: 'misleading,' 'unqualified with additional paragraph' (shortened as 'unqualified+') and 'unqualified.'

The outcome for a case is selected either on the basis of an individual hypothesis, or on the basis of aggregate (i.e., combined) hypotheses (Johnson, Grazioli & Jamal, 1992). Outcome selection from individual hypotheses is straightforward. If an individual deception hypothesis is material, the global assessment process selects a 'misleading' outcome. If business continuity is in doubt, consistency is lacking and/or uncertainties are material, the process selects an 'unqualified+' outcome. If no hypotheses about anomalies are accepted, an 'unqualified' outcome is rendered.

Aggregate evaluation is performed by first cumulating for each type of hypothesis (deception, error, and disclosure), the dollar value associated with that type and then testing these values against a criterion for aggregate materiality. As in previous models, we assumed 40% of the operating income as a criterion (Johnson, Grazioli & Jamal, 1992). If any aggregate hypothesis set is material, the global assessment process selects a misleading outcome.[14] In case that the data in a case support two alternative outcomes, the model selects the most serious (worst case scenario).

References

Akerlof, G. A. (1970). The market for lemons: qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics, 84,* 488–500.

Albrecht, W. S., Wernz, G. W., Williams, T. L. (1995). *Fraud: Bringing light to the dark side of business.* NY, NY: Richard Irwin.

American Institute of Certified Public Accountants. (1995). *AICPA Professional Standards (AU316 & AU317).* Chicago, IL: Commerce Clearing House.

Anderson, J. R. (1989). The analogical origin of errors in problem solving. In D. and K. Kotovsky (Eds.) Klahr D., Kotovsky K. (Eds.), *Complex information processing: The impact of Herbert. A. Simon* (pp. 343–371). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Arens, A. A., & Loebbecke, J. K. (1991). *Auditing: An integrated approach.* Englewood Cliffs, NJ: Prentice-Hall.

Ericsson, K. A. (1996). The acquisition of expert performance: an introduction to some of the issues. In K. A. Ericsson (Ed.) *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games,* (pp. 1–50). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bartlett, F. C. (1961). *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press.

Baron-Cohen, S. (1995). *Mindblindness.* Cambridge, MA: MIT Press.

Ben-Zeer, T. (1995). The nature and origin of rational errors in arithmetic thinking: Induction from example and prior knowledge. *Cognitive Science, 19,* 341–376.

Bouwman, M. J. (1983). Human diagnostic reasoning by computer: An illustration from financial analysis. *Management Science, 29,* 653–672.

Bowyer, J. B. (1982). *Cheating.* New York, NY: St. Martins Press.

Ceci, S. J., Leichtman, M. D., & Putnick, M. E. (1992). *Cognitive and social factors in early deception.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Clancey, W. J. (1988). Acquiring, representing and evaluating a competence model of diagnosis. In M. T. H. Chi, R. Glaser & M. J. Farr (Eds.), *The Nature of expertise,* (pp. 343–418). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cosmides, L., & Tooby, J. (1995). From function to structure: The role of evolutionary biology and computational theories in cognitive neuroscience. In Michael S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1199–1210). Cambridge, MA: The MIT Press.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31,* 187–276.

Cosmides, L. (1985). *Deduction or Darwinian algorithms? An explanation of the "elusive" content effect of the Wason selection task.* Doctoral Dissertation, Department of Psychology and Social Relations, Harvard University.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In Lawrence A. Hirschfeld & Susan A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition* (pp. 85–116). Cambridge, MA: Cambridge University Press.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides & J. Tooby (Eds.), *The adapted mind* (pp. 163–228). New York, NY: Oxford University Press.

Cummins, R. (1989). *Meaning and Mental Representation.* Cambridge, MA: MIT Press.

Dennett, D. C. (1989). *The intentional stance.* Cambridge, MA: MIT Press.

DePaulo, B. M., Stone, J. I., & Lassiter, G. D. (1985). Deceiving and detecting deceit. In Barry R. Schlenker (Ed.), *The self and social life,* New York, NY: McGraw-Hill Book Company.

Druckman, D., & Bork, R. A. (1991). *In the mind's eye: Enhancing human performance.* Washington, DC: National Academy Press.

Ekman, P. (1992). *Telling Lies: Clues to deceit in the market place, politics and marriage.* W. W. Norton and Company.

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

Feldman, M. S., & March, J. G. (1988). Information in organizations as signal and symbol. In James G. March (Ed.), *Decisions and Organizations,* (pp. 409–428). Cambridge, MA: Basil Blackwell.

Fiske, S. T., & Taylor, S. E. (1991). *Social Cognition.* McGraw Hill.

Fox, C. R. Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science, 44,* 879–895.

Giarrantano, J., & Riley, G. (1989). *Expert systems: Principles and programming.* Boston, MA: PWS-Kent Publishing Company.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43,* 127–171.

Grover, S. L. (1993). Lying, deceit, and subterfuge: A model of dishonesty in the workplace. *Organization Science, 4,* 478–495.

Hansen, J. V., McDonald, J. B., Messier, W. F., & Bell, T. B. (1996). A generalized qualitative response model and the analysis of management fraud. *Management Science, 42,* 1022–1031.

Jamal, K. (1991). *Detecting framing effects in audit judgment.* Unpublished Doctoral Dissertation, University of Minnesota.

Jamal, K., Johnson, P. E., & Berryman, R. G. (1995). Detecting framing effects in financial statements. *Contemporary Accounting Research, 12,* 85–105.

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

Johnson, P. E., Jamal, K., & Berryman, R. G. (1991). Effects of framing on auditor decisions. *Organizational Behavior and Human Decision Process, 50,* 75–105.

Johnson, P. E., Grazioli, S., & Jamal, K. (1992). *Fraud detection: Sources of error in a low base-rate world.* Paper presented at the Tenth European Conference on Artificial Intelligence (ECAI92) Research Workshop on Expert Judgment, Human Error and Intelligent Systems. Vienna, Austria.

Johnson, P. E., Grazioli, S., Jamal, K., & Zualkernan, I. A. (1992). Success and failure in expert reasoning. *Organizational Behavior and Human Decision Processes, 53,* 173–203.

Johnson, P. E., Kochevar, L. K., & Zualkernan, I. A. (1992). Expertise and Fit: Aspects of cognition. In Herbert L. Pick, Jr., Paulus Van den Broek & David C. Knill (Eds.), *Cognition: Conceptual and methodological issues.* Washington, DC: American Psychological Association (pp. 305–331).

Johnson, P. E., Grazioli, S., & Jamal, K. (1993). Fraud detection: Intentionality and deception in cognition. *Accounting. Organizations and Society, 18,* 467–488.

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

Koehler, J. J. (1996). The base-rate fallacy reconsidered: descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19,* 1–53.

Kraut, R. (1980). Humans as Lie Detectors: Some second thoughts. *Journal of Communication, 30,* 209–216.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence, 33,* 1–64.

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

Leslie, A. M. (1987). Pretense and representation: The origins of "Theory of Mind." *Psychological Review, 94,* 412–426.

Lewandowsky, Stephan. (1993). The rewards and hazards of computer simulations. *Psychological Science, 4,* 236–243.

Loebbecke, J. K., Einning, M. M., & Willingham, J. J. (1989). Auditors experience with material regularities: Frequency, nature and detectability. *Auditing: A Journal of Practice and Theory, 9,* 1–28.

Mawby, R., & Mitchell, R. (1986). Feints and ruses: An analysis of deception in sports. In R. W. Mitchell & N. S. Thompson (Eds.), *Deception: Perspectives on human and non-human deceit* (pp. 313–322). Albany, NY: SUNY Press.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science, 2,* 387–395.

Millikan, R. G. (1993). *White queen psychology and other essays for Alice.* Cambridge, MA: The MIT Press.

Mitchell, R. W., & Thompson, N. S. (1976). *Deception: Perspectives on human and non-human deceit.* Albany, NY: SUNY Press.

Neisser, U. (1976). *Cognition and reality.* San Francisco, CA: W. H. Freeman and Company.

Newell, A. (1990). Foundations of cognitive science. *Unified theories of cognition* (pp. 42–110). Cambridge, MA: Harvard University Press.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W.G. Chase (ed.), *Visual Information Processing.* New York: Academic.

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

Nikolai, L. A., & Bazley, J. D. (1992). *Intermediate accounting (Fifth Edition).* Cincinnati, OH: Southwestern Publishing Company.

Norusis, M. (1988). *SPSS/PC+ Advanced statistics V.2.0.* SPSS Inc. Chicago, Illinois.

Payne, S. J., & Squibb, H. R. (1990). Algebra mal-rules and cognitive accounts of error. *Cognitive Science, 14,* 445–481.

Peters, J. M. (1990). A cognitive computational model of risk hypothesis generation. *Journal of Accounting Research, 28,* 83–109.

Platt, R. D., & Griggs, R. A. (1993). Darwinian algorithms and the Wason selection task: A factorial analysis of social contract selection task problems. *Cognition, 48,* 163–192.

Reason, J. T. (1990). *Human Error.* Cambridge, UK: Cambridge University Press.

Russow, L. M. (1986). Deception: A philosophical perspective. In R. W. Mitchell & N. S. Thompson (Eds.), *Deception: Perspectives on human and non-human deceit* (pp. 3–40). Albany, NY: SUNY Press.

Schilit, H. M. (1993). *Financial Shenanigans: How to detect accounting gimmicks & fraud in financial reports.* New York, NY: McGraw Hill.

Seely-Brown, J., & VanLehn, K. (1979). *Towards a generative theory of "bugs".* Xerox Palo Alto Research Center, Cognitive and Instructional Series-2 (SSL-80–6).

Shank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thagard, P. (1992). Adversarial problem solving: Modeling an opponent using explanatory coherence. *Cognitive Science, 16,* 123–149.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93,* 239–257.

Turner, J. L. (1980). Classifying acts of fraud. In R. K. Elliott & J. J. Willingham (Eds.), *Management fraud: Detection and deterrence.* New York, NY: Petrocelli Books.

VanLehn, K. (1990). *Mind Bugs: The origins of procedural misconceptions.* Cambridge, MA: The MIT Press.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In Michael I. Posner (Ed.), *Foundations of Cognitive Science,* (pp. 257–579). Cambridge, MA: MIT Press.

Vasek, M. E., (1986). Lying as a skill: The development of deception in children. In R. W. Mitchell & N. S. Thompson (Eds.), *Deception: Perspectives on human and non-human deceit,* (pp. 271–292). Albany, NY: SUNY Press.

Wellman, H. M. (1990). *The child's theory of mind.* Cambridge, MA: MIT Press.

Whiten, A., & Perner, J. (1991). Fundamental issues in mindreading. In A. Whiten (Ed.) *Natural theory of mind* (pp. 1–17). Cambridge, MA: Basil Blackwell.

Zaulkernan, Imran, A., & Johnson, Paul, E. (1992). Metaphors of reasoning as problem-solving tools. *Metaphor and Symbolic Activity, 7,* 157–184.