DOCUMENT RESUME

ED 350 306 TM 018 129

| | |
|---|---|
| AUTHOR | Graesser, Arthur C. |
| TITLE | Questioning Mechanisms during Complex Learning. |
| SPONS AGENCY | Office of Naval Research, Arlington, Va. |
| PUB DATE | [92] |
| CONTRACT | N00014-90-J-1492 |
| NOTE | 60p.; Text is in small print. |
| PUB TYPE | Reports - Research/Technical (143) |
| EDRS PRICE | MF01/PC03 Plus Postage. |
| DESCRIPTORS | Algebra; *College Students; Computer Assisted Instruction; Difficulty Level; Feedback; Grade 7; Higher Education; Junior High Schools; *Junior High School Students; Knowledge Level; *Learning Processes; Models; *Questioning Techniques; Student Attitudes; *Tutoring |
| IDENTIFIERS | *Point and Query Interface; Questions; QUEST Program |

ABSTRACT

The psychological mechanisms that underlie human question asking and answering during comprehension and complex learning were studied. The transcripts of 83 tutoring sessions on research methods for college students and 22 algebra tutoring sessions for seventh graders were collected and analyzed. It was estimated that student questions were about 100 times as frequent in a tutoring session as in regular classroom settings. Analyzed dimensions and categories of questions were correlated with the students' depth of understanding of the material. Students to some extent took an active role in self-regulating their knowledge through their questions, but they needed training in improving questioning skills. In an auxiliary study on question asking, the "Point and Query" interface, a computer interface based on the QUEST model of questioning, improved the speed and quality of questioning for 32 college students. Another study found that questions were stimulated when there was a contradiction, when anomalous information was inserted, and when critical information was deleted. There is a list of 132 references. (SLD)

# QUESTIONING MECHANISMS DURING COMPLEX LEARNING

Arthur C. Graesser, Principal Investigator

Department of Psychology, Department of Mathematical Sciences, and the Institute for Intelligent Systems

Mailing address: Arthur C. Graesser
Department of Psychology
Memphis State University
Memphis, TN 38152
(901) 678-2742
graesserac@memstvx1.bitnet

Reproduction in whole or part is permitted for any purpose of the United States Government.

This research was sponsored by the Cognitive Science Program of the Office of Naval Research, under Contract Number N00014-90-J-1492.

APPROVED FOR PUBLIC RELEASE:
DISTRIBUTION UNLIMITED

## ABSTRACT

This research investigated the psychological mechanisms that underlie human question asking and answering during comprehension and complex learning. Questioning mechanisms are fundamental components of human cognition and must be integrated in contemporary models of complex learning, curiousity, creativity, conversation, and intelligence. A scientific understanding of human question asking and answering also provides critical insights on how to design dialogue facilities in intelligent tutoring systems, expert systems, and human-computer interfaces.

The primary studies on this contract investigated question asking and answering during tutoring. We collected and analyzed the transcripts of 83 tutoring sessions on research methods (college students) and 22 tutoring sessions on basic algebra (7th graders). We estimated that student questions were approximately 100 times as frequent in a tutoring session as a typical classroom setting; this in part might explain why learning is substantially better in tutoring than classroom settings. We analyzed the knowledge states, strategies, and interaction patterns of students and tutors during questioning. The questions were classified on several dimensions: degree of specification, content of information requested, and the psychological mechanism that generated a question. These dimensions and categories were correlated with the students' depth of understanding the material. We found that students to some extent took an active role in self-regulating their knowledge by identifying their knowledge deficits and asking questions that repair such deficits. However, students need substantial training in improving their question asking skills. Most of the students' answers to deep questions asked by the tutor (e.g., why, why-not, how, what-if) were poor in quality, so the tutor helped answer these questions in the form of a collaborative process that took several conversational turns. We analyzed the structure of these interactions, the feedback supplied by the tutors, and the cognitive strategies that generated answers to the questions.

There were two auxiliary studies on question asking. In one project, we designed a human-computer interface that facilitates the speed and quality of questioning, called the "Point and Query" (P&Q) interface. The student points to a word or picture element on the computer screen and then to a question about that element from a menu of relevant questions. The set of relevant questions and the answers to the questions was based on a psychological model of questioning called QUEST. The frequency of student questions on the P&Q software was approximately 800 times that in a classroom setting. In the second project, we investigated the stimulus conditions that trigger questions when students comprehend text and attempt to solve mathematics problems. Questions were triggered when there is a contradiction, when anomalous information is inserted, and when critical information is deleted.

The purpose of this project was to investigate the psychological mechanisms that underlie question asking and answering during complex learning and comprehension. Question asking and answering are fundamental cognitive activities that play a critical role in complex learning (Brown, 1988; Collins, 1985, 1988; Dillon, 1987; Miyake & Norman, 1979; Palinscar & Brown, 1984), memory (Norman, 1973; Pressley, Goodchild, Fleet, Zajchowski, & Evans, 1989), creativity (Sternberg, 1987), dialogue (Clark & Schaefer, 1989; Goffman, 1974; Turner & Cullingford, 1989), intelligence (Schank, 1986), and other components of the cognitive system (Lauer, Peacock, & Graesser, 1992). On the practical side, a scientific understanding of human question asking and answering should provide insights on how to design dialogue facilities in intelligent tutoring systems, expert systems, and human-computer interfaces.

This final report has four major sections. The first section reviews the relevant research in cognitive science on question asking and answering. This includes a brief overview of the previous ONR grant on human question answering and a theoretical scheme for analyzing questions. The second section reports the results of a project on tutoring that we conducted on the present grant. We collected and analyzed the transcripts of tutoring sessions on research methods, statistics, and mathematics. The third section describes a new human-computer interface that we have designed which facilitates the speed and quality of questioning (called the "Point & Query" interface). The user points to a word or picture element on the computer screen and then to a question about that element from a menu of relevant questions. We have analyzed the questions that college students ask when they explore information about woodwind instruments with the Point & Query interface. The fourth section reports a series of experiments that identify the stimulus conditions which trigger questions when students comprehend text or attempt to solve mathematics problems. We examined the extent to which questions are generated when there is either a contradiction, an insertion of anomalous information, or the deletion of critical information.

## I. REVIEW OF RELEVANT RESEARCH ON QUESTION ASKING AND ANSWERING

### Theoretical schemes for analyzing questions

Researchers in several fields have proposed schemes for classifying questions and schemes for decomposing questions into subconstituents. This section reviews the schemes that are most relevant to a psychological theory of questioning.

Presupposition and Focus. Every question can be decomposed into presupposed information and the focal information being queried. For example, in the question "When did Frank drop out of college?" one presupposition is that Frank dropped out of college whereas the focus addresses the time of that event. The presupposed information is in the common ground, i.e., the mutual knowledge that the questioner believes is shared by the questioner and answerer (Clark & Schaefer, 1989; Kass & Finin, 1988). It is the "given" information from the perspective of the given-new contract in discourse processing theories (Clark & Haviland, 1977; Gernsbacher, 1990; Halliday, 1967; Needham, 1990). In a detailed and complete analysis, there would be several propositions presupposed in the example question, including: (1) Frank exists, (2) a particular college exists, (3) Frank went to the college, (4) Frank dropped out of the college, and (5) the questioner believes that both the questioner and answerer know 1-5. In contrast to the presupposed information, the focus of the question draws the answerer's attention to the information that the questioner needs and hopes the answerer will supply. The answer includes new information that is outside of the common ground, at least when genuine information-seeking questions are asked.

Some questions have incorrect or problematic presuppositions. Suppose, for example, that Frank never drank booze but the questioner asked "Did Frank stop drinking booze?". A cooperative answerer would correct the erroneous presupposition (e.g., "Frank never drank booze") rather than merely answering the question YES or NO (Kaplan, 1983). It is misleading to give a YES or NO answer to this question even though this verification question technically invites a YES or NO answer. A YES answer means that Frank once drank and subsequently stopped whereas a NO answer means that Frank continues to drink; for both of these answers, it is presupposed that Frank drank booze (Green, 1989; Grishman, 1986; Kempson, 1979). Cooperative answerers are expected to correct erroneous presuppositions rather than to supply a misleading YES/NO answer. A crafty lawyer can trick a witness into accepting an erroneous presupposition by insisting that the witness supply a YES or NO answer to this type of leading question (Loftus, 1975).

Listeners do not always carefully scrutinize the validity of the presuppositions of questions. A striking example of this is the Moses illusion (Reder & Cleeremans, 1990). When asked "How many animals of each kind did Moses take on the ark?", most people answer "two" in spite of the fact that they know that Noah rather than Moses took the animals on the ark. Listeners normally assume that the speaker is being cooperative and is presupposing only correct information (Grice, 1975), so the answerer does not expend much effort evaluating whether the presuppositions behind a question are true. In contrast, the answerer does notice incorrect information in the *focus* of the question, e.g., "Was it Moses who took two animals of each kind on the ark?" (Reder & Cleeremans, 1990).

*Assumptions behind Information-seeking Questions*. Some questions are genuine information-seeking questions in the sense that the questioner is missing information and believes that the answerer can supply it. Van der Meij (1987) identified several assumptions that must be met before an utterance constitutes a genuine information-seeking question:

1. The questioner does not know the information asked for with the question.
2. The questioner believes that the presuppositions of the question are true.
3. The questioner believes that an answer exists.
4. The questioner wants to know the answer.
5. The questioner can assess whether a reply constitutes an answer.
6. The questioner believes the answerer knows the answer.
7. The questioner believes that the answerer will not give the answer in absence of the question.
8. The questioner believes that the answerer will supply the answer.
9. The questioner poses the question only if the benefits exceed the costs, e.g., the benefits of knowing the answer must exceed the costs of asking the question.

A question is not an information-seeking question to the extent that these assumptions are not met. For example, instead of being information-seeking questions, some interrogative utterances are indirect requests for the listener to do something on behalf of the speaker (Clark, 1979; Francik & Clark, 1985; Gibbs & Mueller, 1988; Gordon & Lakoff, 1971; Searle, 1969). When a speaker says "Could you pass the salt?" at a dinner conversation, the speaker wants the listener to perform an action rather than formulating a reply that addresses the listener's salt passing abilities. This utterance fails to meet most of the nine assumptions listed above. Similarly, gripes (e.g., "Why don't you listen to me?") are interrogative expressions that would fail to meet many of the assumptions of a genuine information-seeking question. It should be noted that speech acts are normally defined according to the assumptions shared by speech participants rather than by syntactic or semantic regularities alone (Allen, 1987; Bach & Harnish, 1979; Gibbs & Mueller, 1988; Hudson, 1975; Searle, 1969).

Given this theoretical context, there is the pressing issue of what constitutes a question. It is important to acknowledge that even an information-seeking "question", or what we call an *inquiry*, is not always expressed in an interrogative syntactic form, i.e., an utterance with a question mark (?).

What is your address? (interrogative mood)
Tell me what your address is. (imperative mood)
I need to know your address. (declarative mood)

The above three utterances are inquiries but only the first utterance is an interrogative expression. Moreover, it is not the case that all interrogative expressions are inquiries, as illustrated below.

What is your address? (inquiry)
Could you pass the salt? (request, directive)
Why don't you ever listen to me? (gripe)

Therefore, there is hardly a direct mapping between the syntactic mood of an utterance and its pragmatic speech act category (Bach & Harnish, 1979; Hudson, 1975; Searle, 1969). For the purposes of this report, we define a question as either an inquiry, an interrogative expression, or both.

Categorization of Questions. Graesser, Person, and Huber (1992) developed an analytical scheme for classifying questions, which is presented in Table 1. The question categories are defined primarily on the basis of the content of the infor:ation sought rather than on the question stems (i.e., why, where, who, etc.). Causal antecedent questions, for example, tap the previous events and enabling states that caused some event to occur. A causal antecedent question can be articulated linguistically with a variety of stems: why did the event occur, how did the event occur, what caused the event to occur, what enabled the event to occur, and so on. Verification questions invite brief replies of YES, NO, or MAYBE. Most of the question categories have an interrogative syntactic form. The two exceptions are the assertion and request/di·ective categories, which are inquiries expressed in a declarative or imperative mood.

The question categorization scheme proposed by Graesser, Person, and Huber is grounded both in theory and in empirical research. The theoretical foundations include models of question answering in artificial intelligence (Allen, 1987; Lehnert, 1978; Schank & Abelson, 1977; Souther, Acker, Lester, & Porter, 1989) and speech act classifications in discourse processing (D'Andrade & Wish, 1985). The classification scheme is empirically adequate in two senses. First, the scheme is exhaustive because it could accomodate thousands of questions that were asked in the context of tutoring and classroom interactions (Graesser, Person, & Huber, in press), as will be discussed in Section II. Second, the scheme is reliable because trained judges could classify the questions with a high degree of interjudge reliability.

The question categories vary in the length of the expected answers. Questions that invite short answers, such as verification questions and concept completion questions, place few demands on the answerer because a satisfactory answer is only a word or phrase. The answers to "long-answer" questions typically span several sentences. One way to induce a listener to talk is to ask a long-answer question, e.g., causal antecedent, goal-orientation, instrumental-procedural, etc.

Some questions are hybrids of two or more question categories. Verification questions are frequently combined with another category. For example, the question "Did Frank drop out of school because of drinking?" is a hybrid between a verification question and a causal antecedent question. This hybrid question gives the option to the answerer as to whether to answer the short-answer verification question, the long-answer causal antecedent question, or both. The fact that there are hybrid questions should not be construed as a weakness in the classification scheme. Most adequate classification schemes in the social sciences are polythetic rather than monothetic (Stokal, 1974). Each observation can be assigned to one and only one category in a monothetic classification scheme whereas an observation can be assigned to multiple categories in a polythetic classification.

Goals of speech participants. An adequate theory of questioning must keep track of the goals of the speech participants (Allen, 1983, 1987; Appelt, 1984; Bruce, 1982; Clark, 1979; Cohen, Perrault, & Allen, 1982; Francik & Clark, 1985; Graesser, Roberts, & Hackett-Renner, 1990; Kaplan, 1983; Kass & Finin, 1988). Suppose that a passenger rushes through an airport, approaches a flight attendant, and asks "When does Northwest 422 leave?". A cooperative reply would give both time and location information (such as "1:33 at gate B21") even though the literal question specifies only time information. The unsolicited location information is included in the answer because the answerer appropriately analyzed the goals of the questioner. The customer obviously was in a hurry and needed to make the flight on time; the customer had the goal of being at the correct gate in addition to the goal of being at the gate on time. In this example, the location information does not address the literal question but it does properly address the questioner's goals.

The Gricean maxims (Grice, 1975) can be viewed as goals for effective communication. The goals associated with the maxim of "quality" are to be truthful and to avoid making claims that cannot be supported with evidence. The goals associated with the maxim of "manner" are to avoid obscurity, to avoid ambiquity, to be brief, and to be orderly. Similarly, there are goals associated with the maxims of "quantity" and "relation". Hovy (1988) has identified the goals that are associated with speech acts and the pragmatic components of conversation. The importance of tracking goals is compatible with the view that a theory of questioning is embedded in a more general theory of conversation and discourse context (Carlsen, 1991; Clark & Schaefer, 1989; Cohen, et al., 1982).

Question generation mechanisms. Graesser, Person, and Huber (1992, in press) identified four clusters of mechanisms that generate questions in naturalistic conversation. Some of these mechanisms are familiar to researchers investigating question asking (Kass, 1992; Ram, 1990; Reisbeck, 1988; Schank, 1986) whereas others were discovered when Graesser, Person, and Huber analyzed transcripts of tutoring sessions and classroom interactions, as will be discussed in Section II.

(1) Questions that address knowledge deficits. The speaker asks a question when he identifies a deficit in his knowledge base and wants to correct the deficit. These information-seeking questions occur in the following conditions:

(A) The questioner encounters an obstacle in a plan or problem. For example, a passenger cannot find his gate so he asks a flight attendant "Where is gate B45?".

(B) A contradiction is detected. A person observes that a television is displaying a program when the set is not plugged into an electrical outlet, so the person holds the plug and asks "How does this television work?".

(C) An unusual or anomalous event is observed. A business person hears about a 110-point increase in the Dow Jones average and asks "Why is there a sudden increase in the stock market?".

(D) There is an obvious gap in the questioner's knowledge base. A child hears her parents use the rare word "aardvark" and asks "What does aardvark mean?".

(E) The questioner needs to make a decision among a set of alternatives that are equally attractive. For example, a customer in a restaurant cannot decide between the trout and the chicken dish so he asks the waitor how each is prepared.

(2) Questions that monitor common ground. These questions monitor the common ground between questioner and answerer. The speech participants need to establish, negotiate, and update their mutual knowledge in order to achieve successful communication (Clark & Schaefer, 1989). Questions are generated in order to inquire whether the listener knows anything about a topic

(e.g., "Do you know about woodwind instruments?"), to verify that a belief is correct ("Isn't a flute a woodwind instrument?"), and to gauge how well the listener is understanding ("Do you follow?"). "Tag" questions are in this category, e.g., "A flute is a woodwind instrument, isn't it?".

(3) Questions that coordinate social action. These questions are needed for multiple agents to collaborate in group activities and for single agents to get other agents to do things. These include the following five types of speech acts: Indirect requests (e.g., Would you do X?), indirect advice (Why don't you do X?), permission (Can I do X?), offers (Can I do X for you?), and negotiations (If I do X, will you do Y?).

(4) Questions th.t control conversation and attention. These questions impose control over the course of conversation and the attention of the speech participants. These include rhetorical questions, greetings, gripes, replies to summons, and questions that change the flow of conversation. The mechanisms in this cluster 4 manage conversation whereas those in cluster 3 manage the actions of agents.

A particular question might be inspired by multiple mechanisms of question generation. For example, when a hostess asks a timid guest the question "Did you read The Prince of Tides?", the question monitors common-ground (cluster 2) and changes the flow of conversation (cluster 4).

Degree of specification. Questions substantially vary on the degree to which the linguistic content specifies the information being sought (Bamber, 1990; Graesser, Person, & Huber, 1992, in press). Questions with high specification have words or phrases that refer to elements of the desired information and the relevant "given" information. Questions with low specification have few words and phrases; the dialogue context is needed for the answerer to fill in the missing information. The examples below illustrate how a question can be posed with high, medium, versus low specification.

What are the variables in the factorial design in Experiment 2? (high specification)
What are the variables? (medium specification)
What about these? (low specification)
Huh? (very low specification)

A question is frequently misinterpreted when the question has low specification and the answerer does not understand the dialogue context.

QUEST: A model of human question answering

In our first grant (Contract Number N00014-88-K-0110), we developed and tested a model of human question answering called QUEST (Graesser & Franklin, 1990; Graesser, Gordon, & Brainerd, 1992). QUEST accounts for the answers that adults produce when they answer different categories of open-class questions, such as why, how, when, what-if. QUEST identifies the information sources for questions; the primary information sources are associated with the content words of questions (i.e., nouns, adjectives, main verbs). Each information source is organized in the form of a conceptual graph structure that contains nodes and relational arcs. Example types of structures are goal/plan hierarchies, causal networks, taxonomic hierarchies, and spatial region hierarchies. Question answering procedures operate systematically on these conceptual graph structures during the course of producing answers. QUEST's knowledge representations and computational procedures are quite similar to some models of question answering in artificial intelligence and computational linguistics (Allen, 1987; Collins, Warnock, Aiello, & Miller, 1975; Dahlgren, 1988; Lehnert, 1978; McKeown, 1985; Souther, Acker, Lester, & Porter, 1989; Webber, 1988; Woods, 1977).

It is convenient to segregate QUEST into four highly interactive components. First, QUEST translates the question into a logical form and assigns it to one of several question categories (as specified in Table 1). Second, QUEST identifies the information sources that are relevant to the question. Third, convergence mechanisms compute the subset of nodes in the information sources that serve as relevant answers to a particular question. These convergence mechanisms narrow the node space from hundreds of nodes in the information sources to less than 10 answers to a particular question. Fourth, QUEST considers pragmatic features of the communicative interaction, such as the goals and common ground of the speech participants.

An important property of QUEST consists of the convergence mechanisms that narrow down the node space from dozens/hundreds of nodes to a handful of nodes which serve as good answers to a question. An arc search procedure restricts its search to particular paths of relational arcs, depending on the question category; nodes on legal paths are better answers than nodes on illegal paths. Each question category would have its own unique arc search procedure (or set of procedures in the case of some categories). Answer quality also decreases as a function of its structural distance, that is, the number of arcs between the queried node and the answer node. A constraint satisfaction component prunes out potential answers that are conceptually incompatible with the queried node (e.g., direct contradictions, time-frame incompatibilities). Both the arc search procedures and structural distance are tractable computationally so they were implemented in a computer program written in LISP.

QUEST was tested in the context of expository texts on scientific mechanisms, narrative texts, and generic concepts (Graesser & Hemphill, 1991; Graesser, Lang, & Roberts, 1991). The model succesfully predicted (a) the likelihood of generating particular answers to questions and (b) goodness-of-answer judgments for particular question-answer pairs. The convergence mechanisms accounted for a substantial percentage of the variance in the data (40% to 75%, depending on the materials and the dependent measure). The arc search procedure was consistently the most robust predictor of question answering so we devoted considerable effort to identifying the arc search procedures of a broad diversity of questions. QUEST can also account for the answers produced in conversational contexts that have more complex pragmatic constraints, such as telephone surveys, televised interviews, and business transactions (Graesser, Roberts, & Hackett-Renner, 1990).

## Questions and complex learning

Question generation has had a somewhat controversial status in cognitive science and education. At one extreme, there is the very optimistic vision that learners are active, self-motivated, inquisitive, creative individuals who ask deep thought-provoking questions and who insist on good answers. Ideal learners are very sensitive to deficits in their knowledge and they initiate self-regulatory strategies that correct the deficits (such as question asking and answering). Although very few researchers would regard this vision as plausible for most learners in most learning environments, researchers frequently advocate educational settings that engage students in active learning and problem solving (Papert, 1980; Piaget, 1952) or that directly train students how to acquire self-regulatory learning strategies (Bransford, Arbitman-Smith, Stein, & Vye, 1985; Collins, 1985; Palinscar & Brown, 1984; Pressley & Levin, 1983; Pressley, et al., 1989). It has frequently been reported that good students are able to monitor and correct their comprehension failures (Brown, Bransford, Ferrara, & Campione, 1983; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Flavell, 1978; Zimmerman, 1989). It has also been reported that there can be substantial improvements in the comprehension, learning, and memory of technical material by training students to ask good questions (Davey & McBride, 1986; Gavelek & Raphael, 1985; King, 1989; Palinscar & Brown, 1984; Singer & Donlan, 1982) or by training them to answer good questions (Pressley, Symons, McDaniel, Snyder, & Turnure, 1988).

Some models of cognition have emphasized the important role of question generation in the cognitive system. According to these models, question generation is a fundamental component in such diverse cognitive processes as the comprehension of text and social behavior (Collins, Brown, & Larkin, 1980; Hilton, 1990; Olson, Duffy, & Mack, 1985), the learning of complex material (Collins, 1988; Miyake & Norman, 1979; Palinscar & Brown, 1984; Schank, 1986), problem solving (Klahr & Dunbar, 1988; Reisbeck, 1988) and creativity (Sternberg, 1987). According to Schank's (1986) SWALE model, for example, learning occurs when an individual observes an anomalous event and generates questions that lead to an explanation of the event. As a consequence, an important portion of long-term memory consists of a large inventory of explained anomalous events. In view of the importance of question asking as a cognitive activity, cognitive scientists have recently developed computer interfaces that make it extremely easy for the user to ask questions (Graesser, Langston, & Lang, in press; Lang, Dumais, Graesser, & Kilman, 1992; Schank, Ferguson, Birnbaum, Barger, & Greising, 1991; Sebrechts & Swartz, 1991), as will be discussed in Section III.

The other end of the continuum presents a more pessimistic picture regarding the status of question generation in cognition and education. It is well documented that student-generated questions in the classroom are both infrequent and unsophisticated (Dillon, 1987, 1988; Gall, 1970; Good, Slovings, Harel, & Emerson, 1987; Kerry, 1987; Lindfors, 1980; van der Meij, 1988). Whereas approximately 94% of the questions in a classroom are asked by the teacher, only 6% are asked by students. The percentage of student questions increases modestly to 18% in tutoring environments that allegedly cater to the learning of individual students (Graesser, Person, & Huber, in press), as will be discussed in Section II. The student questions are normally shallow questions that address the content and interpretation of explicit material, rather than high-level questions that involve inferences, application, synthesis, and evaluation (Flammer, 1981). When teachers attempt to increase student questions by positive reinforcement schedules, there are significant increases in shallow questions but not deep questions (Neber, 1987). Studies have reported a zero or negative correlation between number of student questions and the achievement level of students (Fishbein, Eckert, Lauver, van Leeuwen, & Langmeyer, 1990; Flammer, 1981); these results are incompatible with the claim that good students ask more questions. Individuals may need to master a significant amount of the material before questions come to mind, particularly the high-level sophisticated questions (Miyake & Norman, 1979).

One possible reason that student questions are rare is that students frequently fail to identify their own knowledge deficits. It is well documented that students frequently miss contradictions and inconsistencies in scientific text, mathematical word problems, and other types of material (Baker, 1979; Burbules & Linn, 1988; Epstein, Glenberg, & Bradley, 1984; Glenberg, Wilkinson, & Epstein, 1982; Markman, 1979; Otero & Campanario, 1990). Students frequently have problems in detecting contradictory data, in identifying missing data that is necessary for a solution, and in discriminating superfluous from necessary data (Dillon, 1988). If students have trouble identifying such deficits in their knowledge, there would be an inadequate cognitive foundation for asking questions.

Aside from cognitive deficits, there are social reasons for the low incidence of student questions. There are numerous potential costs to posing questions in a classroom setting (van der Meij, 1987, 1988). The student reveals ignorance and loses status when a bad question is asked. Even when the student asks a good question, the student imposes on a teacher who does not want to be interrupted. Teachers frequently have trouble understanding the students' questions when the students have low domain-specific knowledge and have difficulty setting up the context for the question (Coombs & Alty, 1980); as a consequence, the teachers end up answering the wrong questions, creating misconceptions, or simply dismissing the questions. In some cases, students do not view the classroom as a help context or the teacner as a competent information source. Quite clearly, there are numerous social barriers to asking questions in addition to the cognitive barriers.

Unfortunately, most teachers are not particularly good role models for generating good questions. Less than 4% of the teacher-generated questions are high-level questions, i.e., those that require inferences, the application of an idea to a new domain of knowledge, the synthesis of a new idea from multiple information sources, or the critical evaluation of a claim (Kerry, 1987; Dillon, 1988). Teachers rarely use sophisticated Socratic methods by asking carefully planned sequences of thought-provoking questions that expose the student's misconceptions and contradictions (Collins, 1985, 1988; Stevens, Collins, & Goldin, 1982). The mastery of effective questioning skills apparently requires substantial training, both for teachers and for students.

An accurate account of question generation probably lies somewhere between the optimistic and pessimistic extremes. We would expect questions to be more frequent to the extent that individuals can cognitively detect deficits in their knowledge and to the extent that the social context removes barriers to asking questions. At this point, researchers need to document and to explain the precise conditions that enhance the incidence of questions, the quality of questions, and patterns of questioning. This was indeed one of the major objectives of this grant.

## II. Questioning during Tutoring

The primary project on this grant investigated question asking and answering during tutoring. We collected and analyzed the transcripts of 83 tutoring sessions on research methods (college students) and 22 tutoring sessions on basic algebra (7th graders). The theoretical schemes for analyzing questions (see Section I) were used to guide these analyses.

It is reasonably well documented that learning is better in tutoring sessions than in classroom settings (Bloom, 1984; Cohen, Kulik, & Kulik, 1982). Unfortunately, however, there are very few studies which have carefully examined the process of tutoring (McArthur, Stasz, & Zmuidzinas, 1990; Putnam, 1987) so it is unclear why there is such an advantage. Perhaps students ask more questions in tutoring sessions and thereby correct their knowledge deficits. As discussed in Section 1, students ask very few questions in classroom settings in part because of social barriers. It is possible that tutoring provides a social setting that fosters a more active inquisitive student and ultimately better learning.

There is at least one alternative explanation of the advantage of tutoring over classroom learning. Tutoring exposes patterns of reasoning and problem solving that the classroom setting cannot readily furnish. Much of the reasoning and problem solving is exposed when deep-level questions are asked and answered (i.e., why, why-not, how, what-if). Unfortunately, these deep-level questions are rarely asked by the teachers in classroom settings, as discussed in Section I. Perhaps these deep-level questions are more prevalent in a tutoring setting. A major objective of this research, therefore, was to analyze the role of questions during the tutoring process.

### Methods of collecting tutoring data on research methods

Students and tutors. Tutoring protocols were collected from 27 undergraduate students enrolled in a scientific research methods class at Memphis State University. The students completed the tutoring sessions in order to fulfill a course requirement. Therefore, we had tutoring protocols on a representative sample of college students taking the class, as opposed to a restricted sample of students who were having difficulties with the material. Six psychology graduate students were selected as tutors. Each tutor had performed very well in an undergraduate research methodology class (course grade = A) and had either completed or was currently enrolled in a graduate level methodology course. Each tutor was paid $500 for tutoring students in 18 tutoring sessions.

*Learning materials*. The course instructor selected six topics that are normally troublesome for students in the course. Each topic had related subtopics that would be covered in the tutoring session. An index card was prepared for each of the six topics; subtopics were listed below the major topic, as specified below.

VARIABLES: operational definitions, types of scales, values of variables
GRAPHS: frequency distributions, plotting means, histograms
STATISTICS: decision matrix, Type I and II errors, t-tests, probabilities
HYPOTHESIS TESTING: formulating a hypothesis, practical constraints, control groups, design, statistical analyses
FACTORIAL DESIGNS: independent variables, values on independent variables, dependent variables, statistics, main effects, cells, interactions
INTERACTIONS: independent variables, main effects, types of interactions, statistical significance

The students were exposed to the material covered on a topic before they participated in a tutoring session. This was accomplished in two ways. First, each topic was covered in a classroom lecture by the instructor before that topic was covered in a tutoring session. Second, both the tutors and the students were required to read specific pages in a research methods text, entitled *Methods in Behavioral Research* (Cozby, 1989), before the tutoring session. A mean of 14 pages was read prior to a tutoring session.

The tutoring sessions spanned an eight week period. The topics covered during the first three weeks were variables, graphs, and statistics, with one topic covered per week. A two-week break followed the first three tutoring sessions. The remaining three topics were covered during the subsequent three weeks. Each tutoring session lasted approximately one hour.

*Equipment and setting*. The room used for the tutoring session was equipped with a video camera, a television set, a marker board, colored markers, and the textbook by Cozby. The television screen was covered during the entire session. The camera was positioned so that the student and the entire marker board was in sight. Therefore, the transcripts of the tutoring sessions included both spoken utterances and messages on the marker board.

*Procedure*. When the students entered the tutoring room, they were instructed to sit in view of the camera and to read the topic card aloud. The students were assigned to either a *structured tutoring* condition (which was handled by 3 of the 6 tutors) or to a *normal tutoring* condition (which was handled by the other three tutors). The three normal tutors were not given a specific format to follow, but they were told to resist the temptation of simply lecturing to the students. The three structured tutors had been trained to follow a specific format that was designed to elicit a maximum number of student questions. First, at the beginning of the session, they asked the student to generate three questions that were related to the topic. Second, the tutor pumped the student for questions throughout the entire session by asking "Do you have any questions?". Third, the structured tutors gave the students problems to solve that were relevant to the topic; we anticipated that questions might be more prevalent while students are solving a concrete problem than when they are merely comprehending material. The structured tutoring condition provided an estimate of the upper bound of student questions under ideal conditions.

Each student participated in four tutoring sessions. A counterbalancing scheme was designed so that (a) each student had two structured tutoring sessions and two normal tutoring sessions and (b) a student never had the same tutor twice. Within the first three weeks (and three topics), a student had one topic assigned to the normal tutoring, one to structured tutoring, and the other to no tutoring; the same counterbalancing occurred for the final three weeks (and topics). Each tutor had three students assigned to each of the six tutoring topics. Given that each student participated in 4 sessions and there were 27 students, a total of 108 tutoring sessions were recorded.

Transcription and coding of tutoring sessions. Although 108 tutoring sessions were videotaped, only 83 were eventually transcribed in writing and analyzed (44 normal and 39 structured tutoring sessions). The other 25 protocols could not be transcribed due to audio problems or to extreme video problems that made it difficult to decipher the messages on the marker board. The transcribers were trained on how to transcribe the protocols. They were instructed to transcribe the entire tutoring session verbatim, including all "ums," "ahs," word fragments, broken sentences and pauses. They were told to sketch any messages on the marker board in as much detail as possible. Each written transcription was verified for accuracy before it was coded and analyzed.

Trained judges coded the questions in the transcripts on a number of dimensions that were described in Section I. Two judges achieved a high reliability score in deciding whether or not a speech act was a question/inquiry (Chronbach's alpha = .96 or higher). Another pair of judges categorized each question as to whether it had a high, medium, versus low degree of specification (achieving a spot sample reliability score of .94). A pair of judges were trained to classify the questions on the question categories in Table 1 (achieving a spot sample reliability score of .96 or higher). This question category analysis could be used as a monothetic or a polythetic classification scheme (Stokal, 1974). In the monothetic scheme, the categories are mutually exclusive, so any given speech act was assigned to only one category. When a question was an amalgamation of two or more categories, a highest priority category was determined. For example, the most frequent amalgamation was the verification question category and some other category, e.g., "Is the mean of the sample 4.5?" is an amalgamation of a verification question and a quantification question. Verification questions received lower priority than the other question categories. In the polythetic scheme, each question could be assigned to one, two, or three categories. Finally, two judges were trained to assign each question to one of the four question generation mechanisms: correction of knowledge deficit, monitoring common ground, social coordination of action, and control of conversation. These judges achieved a satisfactory reliability score (.81 or higher).

Results and discussion

In reporting the results of the tutoring study, we will begin with analyses of the questions, including the number of questions and the proportion of questions in various categories. In statistical tests that compare scores between students and tutors, the unit of analysis (i.e., case) was the tutor-student dyad rather than an individual student or tutor. Any variables that examined differences between sets of question categories (e.g., high, medium versus low degree of specification) was treated as a within-subjects variable. Moreover, follow-up analyses were performed using nonparametric statistics, such as sign tests, chi-square tests, Mann-Whitney ANOVA by ranks. After reporting analyses of the questions, this section reports analyses of the answers and more complex patterns of dialogue in the tutoring sessions. Because there are a very large number of quantitative comparisons, we will simply report whether means are significantly different rather than reporting the values of the statistical indices.

Number of tutor and student questions. The mean number of student questions was significantly higher in the structured tutoring condition than in the normal tutoring condition, 44.9 versus 21.1 questions, respectively. The 21.1 frequency is considerably higher than the estimate of .2 student questions per hour in classroom settings (see Section I). We in fact verified this rate by taperecording 12 hours of classroom lectures on research methods, focusing on those hours that covered the same topics as the tutoring sessions; the rate of student questions was .17 questions per hour per student. Indeed, there are approximately 100 times as many questions in normal tutoring sessions as in classrooms; the ratio is 200 to 1 when tutors structure the session to maximize student questions.

The mean number of tutor questions was not significantly different in structured versus normal tutoring, 92.6 versus 95.2, respectively. When considering both tutor and student questions together, 18% of the questions were student questions in normal tutoring and 33% were student questions in structured tutoring. The comparable percentage in a classroom setting is 6% student questions, so once again the incidence of student questioning is more prevalent in these tutoring sessions.

These data document that the tutoring environment supports an inquisitive learner to a greater extent than does the classroom. Students have the opportunity to take an active control over their own learning and to correct their idiosyncratic knowledge deficits. Social barriers do not severely dampen inquisitiveness, as it obviously does in classrooms. Perhaps this explains why learning is superior in tutoring than in classroom environments (Bloom, 1984; Cohen et al., 1982).

Degree of question specification. Only 3% of the student questions and tutor questions had a high degree of specification, i.e., adequate references to arguments, operands, and essential contextual information. Most questions had a medium degree of specification, with percentages of 58% for tutors and 67% for students; the percentages of questions that had low specification were 39% for tutors and 30% for students.

Low specification questions produce a significant amount of misunderstandings and counter-clarification questions on the part of the listener. A counter-clarification question is asked when a listener does not understand the original speaker's question. For example, if the tutor asks "What are the levels?", a student's counter-clarification question might be "Do you mean the levels on the independent variable?". The probability that a question elicited a counter-clarification question was .00, .06, and .17 for questions that were high, medium, versus low in specification. Therefore, dialogue context is frequently not sufficient for the listener to reconstruct the intended meaning of the question. The prevalence of misunderstandings in dialogue has been documented in several contexts, including doctor-patient interactions, lawyer-witness interactions, and student-teacher interactions (Blum-Kulka & Weizmann, 1988; Coombs & Alty, 1980; Edwards & Mercer, 1989; Labov & Fanshel, 1977; Valdez, 1986).

Given that questions rarely have a high degree of specification (see also Bamber, 1990), human-computer interfaces need to accomodate the fact that questions are quite fragmentary and insufficiently articulated. For example, the RABBITT system's principle of "retrieval by reformulation" (Williams, 1984) provides a dialogue between the user and the computer which incrementally converges on a single question. Alternatively, users could select a question from a menu of questions, as in the Point & Query system (Graesser, Langston, & Lang, in press), the ASK TOM system (Schank, et al., 1991), and in other systems (Sebrechts & Swartz, 1991).

Question categories. Table 2 presents the percentage of questions in each of the question categories in Table 1. Data are segregated for tutors and students, in structured versus normal tutoring. The most prevalent category was the verification questions (28%). Other categories with comparatively high percentages were concept completion (14%), interpretational (10%), and instrumental/procedural questions (13%).

Table 2 segregates short-answer and long-answer questions. A short-question invites a very brief reply (i.e., a word or short phrase) whereas a long-answer question invites a lengthy reply of several speech acts. Long-answer questions place the burden on the listener to supply information and manage the dialogue. The tutors had a significantly smaller percentage of long-answer questions than did the students, 41% versus 51%. Therefore, the burden tended to be on the tutor to supply information. The percentage of short-answer questions was somewhat in higher in the structured tutoring than in the normal tutoring, 49% versus 43%.

Another way of cutting the pie is to segregate deep questions (i.e., why, why not, how, what-if) from shallow questions. Deep questions expose the listener's reasoning patterns and problem solving skills that are inherent in deep learning. They include the following categories of questions: antecedent, consequence, enablement, goal-orientation, instrumental/procedural, and expectational. The tutors asked deep questions 20% of the time whereas the students did 25% of the time. It is informative to note that such questions are very infrequent in classroom settings. Although previous researchers have never performed as detailed analysis of questions as we have done in this grant, it has been reported that only 4% of the teacher questions are sophisticated questions, i.e., those that require inferences, the application of an idea to a new domain of knowledge, the synthesis of a new idea from multiple information sources, or the critical evaluation of a claim (Dillon, 1988). The classroom is not an ideal environment to expose students' extended reasoning and problem solving, whereas these complex cognitive skills are more prevalent in tutoring sessions. Perhaps this explains why learning is superior in tutoring sessions.

Question generation mechanisms. The questions were classified into the four question generation mechanisms which were defined in Section I: correction of knowledge deficit, monitoring common ground, social coordination of action, and control of conversation. The distribution of questions among these four categories was not very different between structured and normal tutoring, but there were differences between students and turtors. When considering student questions, 24% of the questions were in the correction of knowledge deficit category, 67% were in the common ground category, 3% were in the social coordination category, and 6% were in the conversation control category. The corresponding percentages for tutors were 0%, 91%, 3%, and 6%.

It is informative that 24% percent of the student questions were in the correction of knowledge deficit category. Questions in this cluster reflect the extent to which students take an active role in self-regulating their knowledge; such questions indicate that students attempt to identify and repair their knowledge deficits and misconceptions. We also found that the majority of the students' common ground questions were attempts to verify the correctness of their own knowledge or beliefs, e.g., "Doesn't a factorial design have two independent variables?". Once again, this constitutes an active regulation of one's knowledge. It has been argued that students should take a more active role in constructing, regulating, and monitoring their own knowledge (Carroll, Mack, Lewis, Grischkowski, & Robertson, 1985; Flavell, 1978; Piaget, 1985; Papert, 1980). At the same time, it has been argued that students do not naturally acquire the skills of identifying knowledge deficits so they need to be taught and guided by the teacher (Baker & Brown, 1980; Bransford et al., 1985; Brown, 1988; Pressley et al., 1989). A tutoring environment clearly shows more promise than a classroom environment in facilitating self-regulated learning.

Most of the tutors' questions were in the common ground category. The tutors normally grilled the students with questions in order to find out what they knew about various subtopics and skills. Many of these questions were scripted, such that the tutor asked the same question of all students. At other times, the tutor detected a problem or gap in the student's knowledge base so the tutor asked questions to diagnose the problem.

Correlations between student questions and examination scores. Correlational analyses were performed in order to assess whether there is any relationship between the number of student questions and student achievement. Achievement was measured by the examination scores throughout the course; there were 150 four-alternative multiple questions altogether. We segregated the first three tutoring sessions (half 1) from the last three sessions (half 2). During the first half of the normal tutoring condition, there was a significant negative correlation between number of student questions and achievement, *r* = -.56; during the second half the correlation was negative but not significant, *r* = -.12. The corresponding correlations in the structured tutoring condition were -.25 and .15, both of which were nonsignificant.

The fact that the questions had a zero or negative correlation with achievement is consistent with some previous studies (Fishbein et al., 1990; Flammer, 1981) and appears to be incompatible with the hypothesis that good students actively monitor their own comprehension failures (Brown, Bransford, Ferraro, & Campione, 1983; Chi et al., 1989; Zimmerman, 1989). However, the fact that the picture changes as students have more exposure to tutoring suggests that the better students learn more effective questioning skills.

We performed some follow-up correlational analyses in order to assess whether achievement is correlated with good questions. None of the raw correlations were statistically significant. One analysis correlated achievement with deep questions (as defined earlier). During half 1 of the tutoring sessions, the correlations were -.19 and -.29 for normal versus structured tutoring, respectively; the corresponding correlations were higher in the second half, -.08 and .05, respectively. A second analysis correlated achievement with the number of questions that corrected knowledge deficits (as defined earlier). During half 1 of the tutoring sessions, the correlations were -.18 and -.25 in normal versus structured tutoring; the corresponding correlations in half 2 were higher, .00 and .25.

In all of the above analyses between achievement and question asking, the correlations were less negative (or more positive) in the second half than the first half. These shifts in correlations were statistically significant when multiple regression analyses were performed. This suggests that students can be taught how to ask good questions even if the process of asking good questions does not come naturally. Moreover, there is evidence in the literature that training students to ask good questions takes several hours of training and leads to substantial improvments in the comprehension and acquisition of technical material (Davey & McBride, 1986; Gavelek & Raphael, 1985; King, 1989; Palinscar & Brown, 1984; Singer & Donlan, 1982; Yopp, 1988).

All of the subsequent analyses focused exclusively on the normal tutoring condition. Our goal was to understand typical characteristics of question asking and answering in naturalistic tutoring. There were artificial constraints in the structured tutoring condition, so these sessions were not regarded as representative of normal tutoring.

Quality of answers to deep questions. Most of the students' answers to deep tutor questions were poor in quality, even when the students believed they understood the material. We extracted all tutor questions that were deep questions (as defined earlier). The answers to these questions were assigned to one of the following five categories: correct, partial, vague/incoherent, error-ridden, and no answer. The percentages of answers in these categories were 39%, 25%, 9%, 14%, and 14%, respectively. It might be noted that the quality of the tutors' answers to the deep questions asked by students were distributed among the five categories as follows: 46%, 30%, 13%, 4%, and 7%. Therefore, the quality of the tutors' answers were only moderately better than that of the students.

Because the students' answers were so low in quality, the tutor typically helped the student answer the question in the form of a collaborative exchange. This collaborative process took several conversational turns, 6.1, 8.2, and 5.7 turns per answer for why, how, and consequence/what-if questions, respectively. Although the tutor originally asked a question, the tutor ended up supplying much of the answer to it during these collaborative exchanges. Therefore, the process of answering a question was a collaborative process in which both parties supplied the answer. Such collaborative processes have been emphasized in contemporary models of discourse and social interaction (Carlsen, 1991; Clark & Schaefer, 1989; Fox, 1988; Resnick, Salmon, & Zeitz, 1991; Shrager & Callahan, 1991; Tannen, 1984).

Consider the typical collaborative exchange when the tutor asks the student a deep questions. The student begins by attempting to answer the question, but usually fails to correctly answer it; complete answers occur during the first turn only 24% of the time. During the next turn, the tutor usually pumps the student for more information with expressions such as "okay", "uh-huh", and "keep going". Then the student supplies additional information, but provides a complete answer only 19% of the time. At that point the two parties either converge on an answer collaboratively or the tutor supplies an answer. We found that the student and tutor collaboratively created a complete answer 63% of the time.

Structure of collaborative exchange. Figure 1 summarizes a typical structure of an exchange when a tutor asks a why question. There is a vestige of the classroom frame in which the teacher grills the student and evaluates the answer (Mehan, 1979): (1) tutor asks question, (2) student answers question, and (3) tutor gives feedback on answer. However, this standard classroom questioning frame is augmented in tutoring by the tutor improving on the quality of the answer through a variety of tactics (see Figure 1) and by assessing whether the student understands the answer.

We were very surprised about the pattern of feedback that tutors gave to students' answers. We performed an analysis on each of the students' contributions while answering deep questions. A student's contribution was defined as a turn in which the student either answered a question or elaborated on an answer -- with new information. Each contribution was categorized on answer quality: complete answer, partial answer, vague/incoherent answer, or error-ridden answer. For each contribution, we scored whether the tutor supplied poitive feedback (e.g. "yes", "that's right"), negative feedback (e.g., "no", "you're wrong about that", "not quite"), or a neutral acknowledgement of the answer ("okay", "uh-huh").

We found that tutors rarely gave negative feedback. The likelihood of tutors giving negative feedback to a student's contribution was .00, .03, .00, and .05 for complete, partial, vague, and error-ridden answers, respectively. This result is compatible with McArthur et al. (1990), so it is probably a general phenomenon in this culture that tutors are reluctant to give negative feedback on students' errors and poor answers. Instead of giving negative feedback to errors, tutors usually interrupted the student at a fine-grained level and "spliced" in correct information, in a similar manner as Anderson's LISP tutor (Anderson, Conrad, & Corbett, 1989). There are several reasons why negative feedback is not given by tutors. In some cases it might traumatize a sensitive student to the point of not participating. In any event, we speculate that tutoring could improve if the tutor was less polite and offered more discriminating feedback, including negative feedback. The tutor would need to warn the student up front that they should expect negative feedback periodically.

We were also surprised to learn that tutors were not particularly discriminating in giving positive feedback. When considering all deep tutor questions, the likelihood that a student's contribution received positive feedback was .30, .35, .33, and .24 for complete, partial, vague, and error-ridden contributions. The only case in which there was a discriminating gradient between feedback and answer quality was in the case of why questions, with likelihood scores of .48, .30, .30, and .25. The neutral acknowledgements also did not vary as a function of answer quality. The likelihood of giving neutral acknowledgements was .27, .32, .31, and .35 among the four levels of answer quality. In summary, tutors were simply not discriminating in giving appropriate feedback to students' contributions during the exchange. Tutoring presumably would improve to the extent that the tutors give more accurate feedback.

Tutors frequently asked a sequence of questions without waiting for the students to answer. The tutor essentially revised the question in order to improve it or to converge on an expression that the student could handle. We extracted all multiple questions that had at least one deep question. In many of these sequences (44%) the successive questions were progressively more specific (e.g., "So what is variance? How is it related to standard deviation?"). The other 56% of these sequences was classified as follows:

Questions are progressively easier (18%)
Questions are progressively better articulated (13%)
Question is rearticulated in slightly different words (6%)

Question re-establishes focus of original question (6%)
Questions are progressively more difficult on listener (3%)
Other -- unclassifiable (10%)

Nearly all of the question sequences (approximately 92%) had revised questions that made it easier for the student to interpret or answer the question. This outcome is consistent with a "sensitivity postulate." According to this postulate, the speaker should do everything possible to make the communicative exchange easy on the listener and to remove obstacles from effective communication (Allen, 1983; Francik & Clark, 1985). The fact that tutors tended to ask students short-answer questions more than long-answer questions, as reported earlier, is also consistent with this postulate. It is tempting to speculate that a good tutor might attempt to violate the sensitivity postulate in order to encourage the student to more actively contribute to the exchange.

The tutor inquired whether the student understood an answer to a deep question approximately 25% of the time. In most of these cases, the tutor asked a very open-ended question, e.g., "Do you understand?." Sometimes the tutor got more specific, e.g., "Do you understand why that graph would have a significant interaction?." Very rarely did the tutor ask a series of questions that diagnosed the student's degree of understanding the answer. However, such a diagnosis is important because the students' answers to the open-ended question ("Do you understand?") fail to reflect their true understanding (Chi et al., 1989). Indeed, it is often the case that the good students claim they don't understand whereas the poor students claim they do understand. Tutors should probably be trained to mistrust the students' claims about their level of understanding subtopics.

Answering strategies of tutors to why and how questions. We extracted the why and how questions that were posed by the tutors and analyzed the strategies for answering them. Table 3 presents the subclasses of these two types of questions and lists the answering strategies for those subclasses that were sufficiently frequent. Most of the answering strategies were compatible with the QUEST model of question answering (Graesser & Franklin, 1990, see Section I). We were relieved to learn that there was some generality in the model we had developed in the laboratory, under artificial experimental conditions and pragmatic constraints. At the same time, however, QUEST's arc search procedures would need to be modified and tuned to accomodate the knowledge base of mathematics, statistics, and research methods; the original question answering procedures were formulated and tested in the context of narrative and expository text.

A close inspection of the answering strategies revealed that goal structures, planning networks, and "teleological semantics" constituted major cognitive structures underlying these static quantitative problems (see also Greeno, 1982; Ohlsson & Rees, 1991; Brown & van Lehn, 1980). When individuals answer a why-question, they convey a goal structure that motivates an action or decision rather than providing a logical proof that justifies a conclusion. However, at this point we are uncertain about the extent to which conceptual knowledge and constructs about research methods are integrated with these planning networks.

Goals underlying questions and answers. Figure 2 summarizes the major goals that motivate questions and constrain answers during tutoring. These goals were identified in our analyses of tutoring. The ideal goal of tutoring is to expand the common ground so that the student acquires the relevant knowledge base of the tutor. Throughout the tutoring process the tutor and student attempt to expand the frontier, i.e., the tutor's knowledge at the fringe of the common ground. The tutor also corrects deep conceptual errors, misconceptions, and minor bugs in the knowledge base of the student. The left column lists the primary goals of the tutor whereas the right column identifies the goals of the student.

Most of the goals listed in Figure 2 are self-explanatory, so they need not be elaborated. The top questioning goal of the tutor was particularly pervasive. Tutors frequently used a "syllabus" (i.e., script, list of questions) that guided questioning in a top down fashion and that exposed anticipated problematic concepts (see also McArthur et al., 1990; Putnam, 1987). The students were grilled and evaluated on this list of questions; the tutor had preformulated ideas on what good answers were. Indeed, these scripts were much more pervasive than the tutor's diagnosing and repairing each student's idiosyncratic bugs and misconceptions. Perhaps tutors need to be trained to be more receptive to the student's particular problems and to revise the syllabus plan to accomodate such problems. Nevertheless, a modest proportion of tutor questions were inspired by the "diagnosis-remediation" model in which the tutor diagnoses student errors, identifies the causes of errors, and corrects faulty understanding (Brown & Burton, 1978; Van Lehn, 1991).

## Analysis of tutoring sessions involving 7th graders on mathematics

We collected, transcribed, and analyzed a sample of 22 tutoring sessions at a local middle school in Memphis. There were 13 7th-grade students in the sample who were having trouble with particular topics in their basic mathematics class. There were 10 tutors from a high school who normally provided these tutoring services. In fact, these tutoring sessions were almost all of the tutoring sessions that occurred in the middle school during a one month period. Two example tutoring topics were (a) exponents and (b) constructing equations from algebra word problems. Originally, 29 tutoring sessions were videotaped; 7 were dropped from the sample because of the poor quality of the auditory channel. The tutoring sessions lasted 45-60 minutes, which was quite comparable to that of the research methods tutoring sessions.

The primary purpose of collecting this sample of tutoring session was to assess how representative were the results of the tutoring study on research methods. Indeed, the results from the two tutoring studies were quite similar. In the subsequent comparisons, we included the normal tutoring condition but not the structured tutoring condition in the study on research methods. The number of student questions per hour in the mathematics sample was 32.2, which is quantitatively close to the rate in the research methods sample (21.1). The number of tutor questions per hour in the mathematics sample was 112.1, again rather close to the research methods sample (95.2). The students accounted for 22% of the questions in the mathematics sample and 18% of the questions in the research methods sample.

With respect to degree of question specification, the mathematics sample was very similar to that of the research methods sample. The percentages of questions that were high, medium, versus low in specification were 2%, 60%, and 39% for students, and 1%, 47%, and 52% for tutors. These percentages compare favorably to those in the research methods sample: for students (3%, 67%, and 31%) and for tutors (3%, 50%, and 47%).

The questions were classified according to the analysis of question categories in Tables 1 and 2. Percentages were scored for each of the 18 question categories, segregating student questions and tutor questions. We then correlated these percentages between the two samples, i.e., mathematics versus research methods. The correlation for student questions was very high (*r* = .61); for tutor questions the correlation was extremely high (*r* = .95). Once again, the two samples of tutoring sessions were quite compatible.

Finally, we analyzed the question generation mechanisms that underlied the questions. In the mathematics sample, the percentages of questions involving corrections of knowledge deficits, common ground, social coordination of action, and control of conversation were 24%, 72%, 3%, and 1%, respectively; the corresponding percentages in the research methods sample were 29%, 67%, 2%, and 3%. The percentages for tutors in the mathematics sample were 0%, 93%, 4%, and 3%, whereas the percentages for tutors in the research methods sample were 0%, 91%, 3%, and 7%.

In summary the two samples of tutoring sessions produced remarkably similar data. The questioning exhibited by high school tutors and 7th grade students on the topic of mathematics was very similar to the questioning exhibited by college students on research methods. Therefore, we are satisfied that the results of the tutoring study on research methods are quite general.

## III. Learning on the "Point and Query" Software

We designed a human-computer interface that facilitates the speed and quality of questioning (Graesser, Langston, & Lang, 1991, in press; Graesser, Langston, & Baggett, in press; Lang, Graesser, & Langston, 1991). The student learns entirely by asking questions and reading answers on the "Point & Query" (P&Q) system. When a question is asked, the student first points to a word or picture element on the computer screen and then to a question that is relevant to the element (from a menu of relevant questions). The menu of relevant questions is formulated on the basis of the background knowledge structures and theoretical question answering procedures of QUEST (see Section I). The P&Q software is embedded in a hypertext system so answers are preformulated and quickly retrieved. The P&Q software is quite similar to the ASK TOM system that has recently been developed by Schank et al. (1991).

The P&Q software forces the student to take an active role in learning because the only way the student can learn is to ask questions and interpret answers. The interface is consistent with a philosophy of education that advocates a learning environment in which students self-regulate their own learning. As discussed earlier, however, it is widely acknowledged that students need some guidance in this process if deep knowledge about a topic is to be acquired. Indeed, one conclusion from the tutoring research in Section II is that students need to be taught effective question asking skills, i.e., when to ask questions and what questions to ask. It does not come naturally. Students learn effective questioning skills from the P&Q software because there is a menu of good questions that are relevant to the word or picture element the student points to. Our hope is that the P&Q software will rekindle curiousity in the student and will provide the necessary scaffolding to effective question asking skills. As mentioned earlier, it is well documented that the learning and comprehension of technical material substantially improves after students learn and apply effective question asking skills (King, 1989; Palinscar & Brown, 1984).

There are an extraordinary number of advantages to the P&Q software, which are briefly summarized below.

(1) It is very easy to ask a question. The student can ask a question within two seconds with two clicks of a mouse: select a screen element and then select a question. All other interfaces suffer from the critical shortcoming that it takes several seconds or minutes to ask a question, including structured query languages, natural language interfaces, the Texas Instruments Menu Driven Natural Language interface (Tennant, 1987), and RABBITT's "retrieval by reformulation" (Williams, 1984).

(2) It is very easy to learn the P&Q system. It takes approximately five minutes to learn how to use the P&Q system for students who already know how to use a mouse. The training time of other question asking interfaces is measured in hours.

(3) The student has direct feedback on what questions the P&Q system can handle. This is because the list of relevant questions is displayed in a menu on the screen. In most other interfaces, it is is unclear to the user what questions the computer can handle so the boundaries of the system are ill-defined.

(4) The student learns what questions are good questions. The menu of relevant questions is contingent on the type of background knowledge structure, e.g., goal/plan hierarchy, causal network, taxonomic hierarchy, spatial information. For example, why, how, and what-if questions are

particularly relevant to causal networks but not to spatial information. The student will eventually learn such relationships between the content of the material and the relevant questions. Our formulation of relevant questions is based on a cognitive model of question asking and answering.

(5) The c -puter quickly answers the questions according to a psychological model of question answering. We have used the QUEST model of human question answering as a guide for formulating answers to questions (Graesser & Franklin, 1990; Graesser, Gordon, & Brainerd, 1991; see Section 1). Therefore, the answers a.e formulated on the basis of a psychological theory that has been validated empirically, which distinguishes it from all other query systems in computer science.

(6) Knowledge is organized around questions and answers to questions. Schank (1986) has discussed some of the advantages of organizing and indexing knowledge around questions. There is some evidence that a "question+answer" encoding format is particularly persistent in memory, with slow forgetting rates (Pressley et al., 1988).

(7) The P&Q system facilitates curiousity and active learning, with some guidance on how to navigate through the knowledge base. As will be discussed in this section, Graesser, Langston, and Baggett (in press) have collected data on a prototype P&Q system on woodwind instruments. We found that students asked approximately 135 questions per hour, which is 7 times the rate of question asking during normal tutoring (see Section II) and 800 times the rate of question asking in a classroom setting (per student).

P&Q Software on Woodwind Instruments

We developed some P&Q software for woodwind instruments (Graesser, Langston, & Lang, 1991, in press). This topic was selected because there are multiple levels of knowledge and each level has a tractable set of questions, at least according to the QUEST model of question answering. We specify below the levels of knowledge and the questions that were handled by the P&Q software on woodwind instruments.

Definitions. There are a large number of terms that need to be defined, which should stimulate definitional questions (i.e., "What does X mean?").

Taxonomic knowledge. As depicted in Figure 3, there is a hierarchical taxonomic structure that contrasts air reed, single reed, and double reed instruments at the highest, most abstract level. The terminal concepts are at the lowest level of the hierarchy, such as alto, tenor, and baritone saxophone. The questions associated with taxonomic knowledge were: "What does X mean?", "What are the properties of X?", and "What are the types of X?".

Spatial composition. This is the spatial layout of objects, parts, and features of parts. This knowledge specifies that a particular component contains subcomponents (e.g., a mouthpiece contains a reed and a ligature), that a component is connected to another component (e.g., that the ligature is clamped to the reed), and that there are spatial relations between parts of an instrument (e.g., the neck is between the mouthpiece and the tube). The question associated with spatial composition was "What does X look like?".

Procedural knowledge. This knowledge embodies the actions, plans, and goals of agents. For example, procedural knowledge specifies how a person assembles, holds, and plays an instrument. The question associated with procedural knowledge was "How does a person use/play X?".

Sensory informatio.i. This includes information about visual, auditory, kinesthetic, and other modalities. The two questions associated with sensory information were "What does X look like?" and "What does X sound like?".

Causal knowledge. This embodies causal networks of events and states in technological, biological, and physical systems. Causal knowledge specifies how air flows through the player's mouth and embouchure, continues through the instrument, gets modified by the size of the chamber and the holes, and produces a sound with a particular pitch, intensity, quality, and duration (see air flow diagram in Figure 3). The questions associated with causal knowledge were "How does X affect sound?", "How can a person create X?", "What causes X?", and "What are the consequences of X?".

Obviously, there are systematic relationships between these levels of knowledge. For example, the size of the instrument (spatial and visual information) constrains how it is played (procedural knowledge) and the resulting pitch (sensory information). Whereas a soprano saxophone is small in size, is high in pitch, and is held by the player with two hands, a bass saxophone is large, low in pitch, and rests on a stand. It would be impossible to have a large woodwind instrument produce a high pitch sound. A flute has a pure sound (sensory information) which is caused by the accoustics of the air reed mouthpiece (causal knowledge).

Individuals sample and integrate the information from these viewpoints when a they learn about woodwind instruments. The representation and organization of the acquired concepts are to some extent a product of exploratory processes during learning. For example, if adults never explore the causal knowledge that explains the operation of instruments, then we would not expect them to correlate the size of the instrument and its pitch, or to correlate the features of the mouthpiece with the quality of the sound (e.g., air reads are correlated with pure sounds). These exploratory processes were investigated in an experiment described later in this section.

The P&Q software was embedded in a hypertext system and was implemented on a MacIntosh-II microcomputer. There were approximately 500 "cards" (i.e., computer screen displays) in the hypertext system. There were two pivotal seed cards that the learner could directly revisit at any point in the session: a woodwind taxonomy card and an air flow diagram card (see Figure 3). All of the other cards were answe.s to questions.

There were four major windows on each screen display, as illustrated in Figure 4. These windows included a content window, a question window, a function window, and a context window. The content window was the bottom 75% of the screen that displayed either the answer to a question, the woodwind taxonomy, or the air flow diagram. Using a mouse, the subject pointed to an element in the content window in order to declare which word or picture element was to be queried. Any word in capital letters could be queried. The computer did not respond when the user pointed to a word in small letters. In Figure Y, the subject was curious about the word LAY so he pointed to this word element.

The question window was a menu of questions presented at the top center of the screen display. The questions were presented in two columns and included the 10 types of questions specified earlier in this Section. The computer varied the subset of questions displayed on the menu according to the queried content element. For example, two questions were relevant to the word LAY: "What does X mean?" and "How does X affect sound?". In the example in Figure 4, the subject pointed to the question "What does X mean?" by manipulating the mouse (top half of Figure 4). The answer to the question "What does lay mean?" was then presented in the content window of the subsequent display (bottom of Figure 4).

The function window (top left of a card) displayed four symbols that referred to special-purpose functions. These included: (1) backing up one display, which could be applied recursively, (2) stopping the session, (3) jumping to the woodwind taxonomy, and (4) jumping to the air flow diagram.

The context window (top right of a card) identified the P&Q software and presented a description of the current card. This was the least informative window from the learner's point of view.

The success of the P&Q software depends on both the answers to the questions and the set of questions on the question menu. The question options and the answers to the questions were formulated on the basis of the QUEST model of question answering (see Section I). QUEST answers a broad diversity of questions in the context of taxonomic, spatial, causal, and procedural knowledge. Good questions vary among these knowledge structures. The question options in the question menu depended on the type of knowledge structure (or structures) associated with the content element the learner points to. Suppose that the learner pointed to content element X and there was informative causal and procedural knowledge associated with X; then the question categories associated with causal and procedural knowledge would be presented in the question menu. If an answer to a question was trivial or uninformative, then the question option was not displayed. In summary, the questions in the question window were based on: (a) the QUEST model of question answering, (b) informative types of knowledge structures associated with the queried content element, (c) the good questions that are associated with the type of knowledge structure, and (d) the extent to which there would be an informative answer to a question.

The answers to most of the questions were formulated according to QUEST's strategies of answering questions. There is a unique strategy for each question category that operates on knowledge structures. Consider definitional questions in the context of taxonomic hierarchies, i.e., "What does X mean?". An answer is produced by a "genus-differentiae" strategy which includes (a) the superclass of X and (b) properties of X that distinguish X from its contrast concepts, i.e., the concepts that have the same superclass as X. For example, the definition for an alto saxophone is "a saxophone of medium size that is in the key of E-flat." In this case, the superclass is "saxophone" whereas the conjoint features of {medium size, key of E-flat} uniquely distinguish the alto saxophone from other types of saxophones.

The question answering strategies of other question categories are quite different from that of definitional questions. The question answering strategy for antecedent questions tap causal networks, and produce antecedent events and enabling states that explain a queried event. When asked "What causes vibrato to occur in a saxophone?", an appropriate answer would be "as air flows through the mouthpiece, the player alternatives between having a tight and loose embouchure." The answers to instrumental/procedural questions include (a) the plan that an agent executes while performing an intentional action and (b) an object, part, or resource that is needed to execute a plan. The question answering stategies for some types of questions were not handled by the original QUEST model but were quite obvious. For example, the answer to "What does X look like?" was a picture of X. The answer to "What does X sound like?" was a 10-second digitized recording of the actual instrument playing a scale.

Asking questions on the P&Q interface: An exploratory study of exploration

We conducted a study which explored how college students ask questions when they learn about woodwind instruments with the P&Q interface. The subjects learned about woodwind instruments entirely by asking questions and reading answers to questions displayed by the computer. The computer recorded the questions that the students asked and the order of the questions.

There is no empirical research that has investigated knowledge exploration for taxonomic-definitional, spatial, sensory, procedural, and causal knowledge in a rich domain such as woodwind instruments. Therefore, any predictions we might offer would be motivated by theoretical considerations and would be quite preliminary. The important contrast addressed in the study was the sampling of deep causal knowledge versus the comparatively superficial knowledge, i.e., the taxonomic, definitional, sensory, and procedural knowledge. We manipulated the goals of the learner and we compared learners with high versus low prior knowledge about music. The learners' goals in a Design Instrument condition encouraged subjects to sample deep causal knowledge whereas this causal knowledge was not needed in an Assemble Band condition. The details of these conditions are discussed later. If the goals of the learner have a substantial impact on exploratory processes, then subjects should explore more causal knowledge in the Design Instrument condition than in the Assemble Band condition; the opposite should be the case for the superficial knowledge. Regarding the level of the learners' prior knowledge, subjects should explore more causal knowledge if they already have a high amount of knowledge about music. Thus, superficial knowledge would presumably be explored before deep causal knowledge.

Predictions can also be made with respect to the time-course of exploring knowledge within a learning session. There is some foundation for expecting a fixed order of knowledge that gets explored, such that the superficial knowledge precedes deep causal knowledge (Dillon, 1984). One would intuitively expect a person to become familiar with the meaning of the terms and the physical features of a system before embarking on causal mechanisms that explain the operation of the system. If there is a canonical ordering of knowledge exploration, then one might expect taxonomic-definitional information to be sampled before causal information, regardless of the goals of the learner. On the other hand, there may be a more flexible ordering of knowledge exploration that directly corresponds to the learner's goals.

Methods. The subjects were 32 undergraduate students at Memphis State University who participated to fulfill a psychology course requirement. We screened the subjects so that half of the subjects had low knowledge about music and half had comparatively high knowledge. Those with high knowledge rated themselves as having moderate to high knowledge on a 6-point scale and also played an instrument.

We manipulated the goals of the learner so that half of the subjects were expected to acquire deep causal knowledge of woodwind instruments and the other half could manage with superficial knowledge. The Design Instrument condition required deep causal knowledge whereas the Assemble Band condition did not. In the Design Instrument condition, the subjects were told that their goal was to design a new instrument that had a deep pure tone. A solution to this problem required them to know that large instruments produce deep tones (i.e., notes with low pitch or frequency) and that pure tones are produced by woodwinds with air reeds rather than mechanical reeds. An ideal instrument would perhaps be a large flute, although it might be difficult for an average diaphragm to sustain an air flow in such an instrument. In any event, a solution to this problem required the subject to have knowledge about the causal relationships between physical features of instruments and the features of the sounds produced by instruments.

The subjects in the Assemble Band condition were instructed that they would be assembling a 6-piece band with woodwinds that would play at a New Years Eve party for 40-year old Yuppies. The subjects did not need a deep causal knowledge of woodwind instruments and sound characteristics in order to solve this problem. It would be satisfactory to have a superficial knowledge about what the instruments looked like, what they sounded like, and what their names were.

Half of the 16 subjects with high music knowledge were assigned to the Design Instrument condition and half to the Assemble Band condition. Similarly, 8 of the subjects with low music knowledge were in the Design Instrument condition and 8 were in the Assemble Band condition. Therefore, there was an orthogonal variation of prior music knowledge and learner goals.

When the subjects arrived, the experimenter described the goals of their task and then gave a 5-minute demonstration of how to use the P&Q interface. They were instructed how to ask a question by pointing to a content element in the content window and then to a question in the question window. They were also instructed and shown how to use the four functions: backup one screen, stop session, jump to taxonomic structure, and jump to air flow diagram.

The subjects had 30 minutes to explore the database on woodwind instruments. They were told they could ask as many questions as they wanted, in whatever order the questions came to mind. The computer recorded the cards that the subject explored and the questions that were asked, in the order that the subject asked them. After the 30-minute learning session, the subjects completed the tasks that were assigned to them. That is, they wrote down a design of a new instrument or a band that was to be assembled for a party. They completed this task at their own pace. For the purposes of this report, we were interested only in the questions that subjects asked during the 30-minute learning session.

Results. In an initial analysis, we simply computed the mean number of questions asked by the subjects during the 30-minute learning session. The subjects asked a mean of 75.6 questions per session in the Design Instrument condition and 59.9 questions in the Assemble Band condition. Therefore, the rate of asking questions in this P&Q interface was 135 questions per hour. This rate is about 7 times the rate of question asking during normal tutoring (see Section II) and 800 times the rate of student questions in a classroom setting (see Section I). The P&Q software clearly has some promise in facilitating curiousity and question asking behavior during learning.

We segregated the session into time blocks and the questions into knowledge categories. Each 30-minute session was segregated into three 10-minute time blocks, yielding time blocks 1, 2, versus 3. We clustered the 10 question categories specified earlier into four categories that tapped four different types of knowledge: taxonomic-definitional, sensory, procedural, and causal. We computed the baserate percentage of questions in these four knowledge categories when considering all possible cards, content elements, and unique questions in the P&Q system. These proportions were 35%, 9%, 1%, and 55% for taxonomic-definitional, sensory, procedural, and causal, respectively.

An Analysis of variance was performed on question asking frequencies using a mixed design with four independent variables: condition (Design Instrument versus Assemble Band), prior experience (high versus low music knowledge), time block (1, 2, versus 3), and knowledge type (taxonomic-definitional, sensory, procedural, and causal). Condition and prior experience were between-subjects variables whereas time block and knowledge type involved repeated measures variables.

We were surprised to learn that prior experience had absolutely no impact on the exploration of knowledge. Prior experience did not have a significant main effect on question frequency and was not part of any significant statistical interaction. This outcome could be explained in a number of ways. Perhaps there was not a sensitive variation of music expertise. Perhaps the knowledge on woodwind instruments was too specialized to be covered by a college student with relatively high music knowledge. Perhaps the constraints of learner goals are extremely robust and mask any effects of prior knowledge. For whatever reason, prior music knowledge had no impact on knowledge exploration.

The frequency of questions did not significantly vary as a function of time blocks, with means of 22.8, 23.4, and 21.2 questions in time blocks 1, 2, and 3, respectively. Therefore, the absolute volume of questions was approximately constant across the three 10-minute segments. More questions were asked in the Design Instrument Condition than in the Assemble Band condition, $F(1, 28) = 5.00$, $p < .05$. The number of questions per time block significantly differed among the four knowledge types, with means of 8.8, 5.3, 1.1, and 7.3 for taxonomic-definitional, sensory, procedural, and causal knowledge, respectively, $F(3,84) = 27.62$, $p < .05$. However, this outcome is not particularly surprising because the baserates were quite different among these four knowledge types. The empirical percentages were 39%, 24%, 5%, and 32% whereas the baserates were 35%, 9%, 5%, and 55% for the four respective types of knowledge. Compared to the baserates, the subjects undersampled causal knowledge and oversampled the other three types of knowledge (particularly the questions in the sensory category).

There was a significant three-way interaction between condition, time block, and type of knowledge, $F(6,168) = 2.89$, $p < .05$. Figure 5 plots the cell means that expose this three-way interaction. The figure and statistical analyses uncovered the following trends in the data.

(1) Taxonomic-definitional knowledge. The frequency of these taxonomic-definitional questions started out the same in both the Design Instrument condition and the Assemble Band condition. Exploration of this knowledge decreased over time in the Design Instrument Condition but remained constant in the Assemble Band condition. It appears that taxonomic knowledge and definitions of terms needed to be sampled during the initial phases of learning, regardless of the goals of the learner. After this basic knowledge was acquired, the learner could explore knowledge that directly addressed his goals.

(2) Causal knowledge. The frequency of questions was extremely high and increased over time in the Design Instrument condition. In contrast, the frequency was extremely low and constant in the Assemble Band condition. It appears that causal knowledge was rarely tapped unless the learner's goals forced the learner to tap this knowledge. Subjects needed to tap this deep causal knowledge in order to design an instrument but not to assemble a 6-piece band. There was a trade-off in sampling causal knowledge versus taxonomic-definitional knowledge.

(3) Sensory knowledge. The frequency of sensory knowledge questions was low and constant in the Design Instrument condition. The frequency was high in the Assemble Band condition but decreased robustly over time. The learners in the Assemble Band condition wanted to find out what the instruments looked like and sounded like early in their exploration. This superficial visual and auditory information was important to the subjects who were trying to assemble a band whereas deep causal knowledge was unimportant.

(4) Procedural knowledge. There was a floor effect in sampling this type of knowledge so it was difficult to decipher trends. Subjects in the Assemble Band condition asked approximately twice as many questions in this category as did subjects in the Design Instrument condition.

Conclusions and discussion. We have documented how individuals explore knowledge when they learn about woodwind instruments. Knowledge exploration was measured by observing the questions that college students asked about woodwind instruments when they used the Point and Query interface. We found that exploration patterns were unaffected by the college students' prior knowledge of music. In contrast, there were dramatic changes in exploration as a function of the their goals and the time course of the learning session.

The learners tended to sample taxonomic knowledge and definitions of terms during the first 10 minutes of the session. Approximately 45% of the questions tapped this type of knowledge in both the Design Instrument and Assemble Band conditions whereas the baserate for these types of questions was only 35%. The learners' goals were quite different in these two conditions: the Design Instrument condition called for deep causal knowledge whereas the subjects in the Assemble Band condition could rely on superficial knowledge. In spite of these differences in goals, the learners needed to know about definitions of words and about the taxonomic composition of the semantic field. Thus, this taxonomic-definitional knowledge apparently must be established before learners can branch out into regions of knowledge that more directly address their goals.

Once the taxonomic-definitional knowledge is established, the learner explores knowledge that addresses the learning goals. This conclusion is obviously supported when time blocks 2 and 3 are inspected in Figure 5. When deep causal knowledge was needed, as in the Design Instrument condition, then the learner explored causal knowledge at the expense of the taxonomic-definitional, sensory, and procedural knowledge. Causal knowledge indeed was not explored unless the learner was forced to seek this knowledge in the pursuit of a goal. The percentage of questions in blocks 2 and 3 that tapped causal knowledge was 52% in the Design Instrument condition but only 16% in the Assemble Band condition; the corresponding baserate was 55% for these questions. On the other hand, if the learners could rely on superficial knowledge in solving their problem, as in the Assemble Band condition, then exploration continued after block 1 by seeking more taxonomic-definitional knowledge (48% of questions, with a baserate of 35%), sensory knowledge (29% of questions, with a baserate of 9%), and a modest amount of procedural knowledge (7% of questions, with a baserate of 1%); once again, causal knowledge was rarely explored in this condition.

Causal knowledge was sampled only when the learner had the goal to solve causal problems. This has nontrivial implications for theories of knowledge organization. Suppose it is correct that people normally rely on superficial knowledge in the everyday world, as opposed to constructing deep mental models of causal systems (Bransford, et al., 1985; Brown, et al., 1983; Graesser & Clark, 1985; Kieras & Bovair, 1984). This superficial knowledge includes the definitions of concepts, the perceptual surfaces of objects, and the procedures of manipulating objects. To the extent that this is true, deep causal knowledge is not an important part of the representation of our knowledge. This would perhaps present a challenge to any explanation-based theory of concept representation that would require deep causal knowledge. We suspect that an explanation-based theory is psychologically plausible only if the explanations appeal to the goals of agents, planning failures, methods of repairing planning failures, and methods of circumventing obstacles (Hammond, 1990; Mitchell, Keller, & Kedar-Cabelli, 1986; Mooney, 1990; Owens, 1991; Schank, 1986). Thus, explanations that center around agents can be readily handled by the cognitive system whereas rigorous causal explanations in biology, physical science, and technology are difficult for the cognitive system to cope with.

This study has demonstrated the value of the P&Q interface in investigating the process of exploring knowledge. The P&Q interface and other similar new interfaces (Schank et al., 1991; Sebrechts & Swartz, 1991) have made it extremely easy for the user to ask questions. The user simply points to a question on a question menu. It was very awkward and time-consuming to pose questions on previous computer interfaces (Lang, Dumais, Kilman, & Graesser, in press) and this presented a serious barrier in investigations of information exploration. We have explored only one semantic field in the present study, namely that of woodwind instruments. We are currently in the process of investigating other domains of knowledge, such as mathematics and statistics.

It is possible that the P&Q interface could have a substantial impact on education and concept representation to the extent that it rekindles curiousity and good question asking skills. As discussed earlier in this report, the learning of complex material can robustly improve if students learn how to ask the right questions (King, 1989; Palinscar & Brown, 1984; Singer & Donlan, 1982). However, it takes many hours to train students how to use ideal question asking skills. One side effect of the P&Q interface is that students learn how to ask good questions in the context of particular types of knowledge structures. Suppose the P&Q interface was extensively integrated with computer software in the educational curriculum, such that students spent hundreds of hours learning how to ask the right questions. We would end up with a more curious, inquisitive, creative, and intelligent generation of learners (Dillon, 1988; Schank, 1986; Sternberg, 1987). The cognitive representation of concepts, both mundane and technical, might radically change.

## IV. Information that Triggers Questions

We conducted a series of experiments that investigated whether individuals ask questions when they encounter anomalous information (Graesser & McMahen, 1992; Graesser, McMahen, & Johnson, 1991). There are several models in cognitive science which predict that questions are triggered by anomalous information (Kass, 1992; Klahr & Dunbar, 1988; Ram, 1990; Reisbeck, 1988; Schank, 1986), but empirical tests of this prediction are conspicuously absent.

In these experiments, college students were instructed to generate questions while they solved quantitative problems or while they comprehended stories. There were different versions of each problem or story, as shown in Table 4 for an algebra word problem. The *original* version was a complete problem from a textbook (or alternatively, a complete story from a book of fables). The other four versions contained anomalous transformations that either deleted or added information. A critical piece of information was removed in the *deletion* version; it would be impossible to solve the problem or comprehend the story without it. The other three versions added a phrase, proposition or sentence to the original version in order to introduce a *contradiction*, a *salient irrelevancy*, or a *subtle irrelevancy*. It was technically impossible to solve the problems in the contradiction versions, whereas it was possible to solve them when the irrelevant information was added.

It is important to clarify what the "anomaly hypothesis" predicts about question asking behavior in the context of this study. The first prediction is that college students should generate more questions in the four anomalous versions than in the original versions. The hypothesis does not predict particular differences among the four transformed versions. Therefore, the mean frequency of questions in the four transformed versions should be greater than the frequency in the original versions. A second prediction is that a subset of the questions should address the anomalous transformations, i.e., the individual should spot the unusual feature of a transformed problem. Once again, the anomaly hypothesis does not have a principled way of discriminating among anomalies, so it assumes that all transformations are equally detectable. It is conceivable that the transformed versions might stimulate questions that are not directly relevant to the transformations (to a greater extent than the original versions). This might occur when individuals vaguely detect that something is unusual about the problem but cannot quite put their finger on the anomalous feature.

There is some foundation for predicting differences among the various types of anomalous transformations. An "obstacle hypothesis" contrasts those anomalous transformations that present an obstacle in planning, problem solving, or understanding from those transformations that do not present an obstacle. The deletion and contradiction versions present a serious obstacle to problem solving and comprehension whereas the salient and subtle irrelevancy transformations do not present such an obstacle. According to the obstacle hypothesis, the incidence of transformation-relevant questions should be higher in the deletion and contradiction versions than in the two irrelevancy versions.

Even though an individual detects an anomaly, the individual may not generate a question about it. Instead, the individual might repair or discount the anomaly in various ways (Markman, 1979; Otero & Companario, 1990). In the case of a deletion version, the comprehender might fill in the deletion with default values that are based on world knowledge. For example, the comprehender might assume that the Bears and Bulls played the same number of games and infer that the number of losses was 32 for the Bulls. In the case of a contradiction version, the comprehender might construct an unusual scenario that resolves the contradiction or that discounts one of the contradictory premises. For example, the comprehender might attribute the contradictory statement in Table 4 to an unintentional typographical error or to information that was elliptically deleted (e.g., the Bulls won the most games *between the Bulls and the Bears*). Such repair strategies might be invoked automatically and unconsciously when questions are not generated. In order for the comprehender to generate a transformation-relevant question, the comprehender must (a) detect the anomaly, (b) acknowledge that the transformation is indeed an anomaly, (c) articulate and refer to the anomaly in words, and (d) address the anomaly in the form of a question. If any of these four conditions fail, then the anomaly will not trigger a question.

Experiment 1: Question asking while solving mathematics and statistics problems.

Upper division undergraduate students generated questions while they solved quantitative problems in either original, deletion, or subtle irrelevancy versions. According to the anomaly hypothesis, the two anomalous versions should elicit more questions than the original problems. Also, a subset of the generated questions in the transformed versions should address the anomalous transformations. According to the obstacle hypothesis, the transformation-relevant questions should be more prevalent in the deletion versions than in the subtle irrelevancy versions.

Methods. The subjects were 30 upper division undergraduate students at Memphis State University who were enrolled in a course on research methods. The subjects participated in the experiment in order to fulfill a laboratory requirement. Two of the 30 students were dropped because they had not completed an undergraduate course in statistics prior to the research methods course even though statistics was a prerequisite.

Twelve quantitative problems were selected from textbooks. Six of the problems came from an introductory college statistics text and six were algebra word problems from a junior high algebra text. Table 4 includes the original version of one of the six algebra problems. The statistics problems were more difficult than the algebra problems, but did not require memorization of any complex formulas (e.g., the formula for a standard deviation, the formula for a $t$-test). The statistics problems were "word problems" that required a conceptual understanding of basic statistical concepts, such as the probability of an event, sample size, mean, standard deviation, null hypothesis, $p < .05$, and statistical significance. The problems emphasized reasoning and drawing inferences about these basic statistical concepts rather than being mechanical "number crunching" exercises. Whenever numerical quantities needed to be combined, the student could perform the quantitative computations by following basic rules in statistics and algebra (e.g., probabilities are multiplied when there are multiple independent events) or by constructing a table with a small number of alternative combinations (e.g., tracing all possible combinations of heads and tails when there are 4 tosses of a fair coin). An example statistics problems is presented below.

> Statistics problem 1. The probability that a person who enters a certain bookstore in a shopping mall will buy a book is 0.4. If 4 customers enter the store, what is the probability that: (a) Exactly one person will buy a book? (b) One or two people will buy a book.

There were three versions of each of the 12 problems. The original version was an exact copy of the problem in the textbook. The deletion version was exactly the same as the original, except that a phrase, clause, or sentence was removed. Moreover, it technically would be impossible to solve the problem without the deleted information; the subject either would encounter an obstacle that rendered the problem unsolvable or would need to construct the missing information inferentially. For example, the deletion version in Table 1 deleted the number of games the Bulls lost, so the winning percentage could not be computed for the Bulls (unless the subject inferred that both teams played an equal number of games). The deletion version in the statistics problem substituted "very high" for the value .04, so it would be impossible to compute the probabilities in "a" and "b". The subtle irrelevancy version was the same as the original problem except that an irrelevant clause or sentence was added. It should be noted that the irrelevant information blended with the general theme or semantic content of the problem. The added information was irrelevant by virtue of mathematical criteria rather than nonmathematical semantic content. Moreover, the subject would be able to solve the problem with the addition of the subtle irrelevancy. Table 4 shows the subtle irrelevancy for the example algebra problem; this version added the sentence "Nearly 20% of the games almost resulted in ties." In the example statistics problem the following sentence was added: "The probability of walking into a dress shop and buying a dress is only .01."

Each subject received a 12-page booklet with one of the 12 problems on each of the pages. The three versions of the problems were counterbalanced across subjects. For each subject, four of the problems were in the original version (i.e., 2 algebra and 2 statistics), four were in the deletion version, and four were in the subtle irrelevancy version. From the standpoint of the problems, each version of a problem was administered to an equal number of subjects. The order in which the problems were presented in the booklet was randomly determined for each subject separately.

The subjects were instructed that they would be generating questions while they were solving word problems in algebra and statistics. The subjects were encouraged to generate as many questions as they could think of. The time course of generating the questions and solving the problems was monitored by the experimenter. More specifically, the subject worked on each problem in three phases, which were timed by the experimenter. The subjects read the problem during phase 1 for 30 seconds. During phase 2, the subjects wrote down questions that came to mind about the problem for 90 seconds. The subject drew a line on the paper at the end of the questions that were generated in phase 2. During phase 3, the subjects solved the problem and generated additional questions for 150 seconds. The subjects were told to write down their work and their additional questions from phase 3 below the line that they had drawn after phase 2. Therefore, the phase 2 and phase 3 questions could be readily distinquished. The subjects were given 30 seconds of rest before starting on the subsequent problem.

Ideally, the phase 2 questions would correspond to a stage in which subjects attempted to understand and represent the problem whereas phase 3 questions would correspond to the process of implementing a solution. Fishbein et al. (1990) reported that questions substantially differ in the problem solving stage than in the problem representation stage.

Results and Discussion. We scored the number of questions that each subject generated for each problem. A description was counted as a question if it had a question mark (e.g., "How many games did the Bulls play?") or if it was an inquiry or comment conveying uncertainty in a noninterrogative form (e.g., "I need to know the number of games the Bulls played", "I don't know how many games the Bulls played."). Two trained judges could identify questions with a high degree of reliability (Chronbach's alpha consistently exceeded .90)

The mean number of questions per problem significantly differed among the three versions, with means of 1.81, 2.49, and 2.09 in the original, deletion, and subtle irrelevancy versions, respectively, $F(2, 54) = 5.61$, $p < .05$. As predicted by the anomaly hypothesis, the two transformed versions combined were significantly higher than the original versions, $t(27) = 3.81$, $p < .05$. Post hoc comparisons revealed the following differences: deletion > subtle irrelevancy = original. Therefore, the deletion transformations had a more robust impact on question generation than did the subtle irrelevancy transformations.

We scored those questions that addressed the transformations in the deletion and subtle irrelevancy versions. Two transformation-relevant questions that addressed the example deletion version were "How many games did the Bulls play?" and "Did the two teams play the same number of games?". A transformation-relevant question that addressed the subtle irrelevancy version was "What does the fact that some of the games almost resulted in ties have to do with the problem?" Questions were not counted if they did not address the transformations, e.g., "Did the two teams ever play each other?" and "What games were they playing?". Two judges could identify the transformation-relevant questions with a high degree of interjudge reliability (Chronbach's alpha of .90).

The mean number of transformation-relevant questions per problem was significantly higher in the deletion versions than in the subtle irrelevancy versions, .80 versus .25, respectively, $t(27) = 6.86$, $p < .05$. This outcome is consistent with the obstacle hypothesis. As predicted by the anomaly hypothesis, both the .80 and .25 values significantly differed from 0, the baseline level of the original versions. Therefore, both types of anomalous transformations caused an increase in the number of transformation-relevant questions.

Two additional analyses on the transformation-relevant questions addressed the materials and the time course of question generation. First, we compared the easy 7th-grade mathematics problems with the difficult college statistics problems. We found that the easy mathematics problems invoked more questions than did the difficult statistics problems, .80 versus .24, respectively, $\underline{t}(27) = 7.23$, $\underline{p} < .05$. However, within each of these types of problems, there were significantly more questions in the deletion versions than the subtle irrelevancy versions: 1.18 versus .43 within mathematics problems and .43 versus .05 within the statistics problems. Regarding the second analysis, we compared the incidence of generating questions during the initial 90-second question asking phase versus the 1[illegible]0-second problem solving phase. We found that most of the questions were generated in the initial pnase rather than the problem solving phase. We found that 74% of the transformation-relevant questions in the deletion condition occurred during the initial question asking phase; the corresponding percentage was 75% in the subtle irrelevancy condition.

In summary, the anomaly hypothesis was confirmed by the fact that the problems with anomalous transformations caused an increase in subject-generated questions. Moreover, a subset of the questions addressed the deletions and subtle irrelovancies. The deletions elicited more transformation-relevant questions than did the subtle irrelevancies, an outcome that supports the obstacle hypothesis. The finding that the deletion versions produced the most questions occurred in the case of the easy mathematics problems and also the more difficult statistics problems. Finally, most of the questions were asked during the initial stages of solving the problem, i.e., the first two minutes.

## Experiment 2: Question asking while solving mathematics and analytical problems

Lower division college students generated questions and solved quantitative problems that were presented in one of five versions: original, deletion, contradiction, salient irrelevancy, and subtle irrelevancy (see Table 4). Half of the 10 problems were algebra word problems whereas the other half were analytical "brain teasers".

Methods. The subjects were 40 undergraduate students at Memphis State University who were enrolled in an introductory psychology course.

Ten original problems were selected. Five of the problems were algebra word problems sampled from a junior high algebra text, as illustrated in Table 4. The other 5 questions were analytical brain teasers selected from an SAT sample examination. An example analytical problem is presented below.

> A half tone is the smallest possible interval between notes. Note T is a half tone higher than note V. Note V is a whole tone higher than note W. Note W is a half tone lower than note X. Note X is a whole tone lower tha note T. Note Y is a whole tone lower than note W. What is the order of the notes from lowest to highest?

There were 5 versions of each problem: original, deletion, contradiction, salient irrelevancy, and subtle irrelevancy. The deletion and subtle irrelevancy versions were constructed in the same manner as in Experiment 1. The contradiction versions involved explicit contradictions by adding one statement to the original version. For example, the contradictory statement "The Bulls won more games than the Bears" is incompatible with the earlier claims that the Bears won 40 games and the Bulls won 32 games. The contradiction version of the example analytical problem added the sentence "Note T is a half tone lower than note V"; this added sentence directly contradicts the earlier sentence "Note T is a half tone higher than note V." Each salient irrelevant transformation had one statement added to the original version. The statement was not only irrelevant to solving the quantitative problem, but also was semantically irrelevant to the problem theme. For example, the fact that the players were always getting into arguments has little semantic relevance to the win-loss records of the Bears and the Bulls. The salient irrelevancy in the example analytical problem added the sentence "April showers may bring May flowers," which is clearly unrelated to the ordering of notes on a musical scale.

Each subject received the 5 algebra word problems and the 5 analytical word problems in a 10-page booklet. Within each of these two types of problems, a particular subject received one problem in each of the five versions. The assignment of versions to problems was counterbalanced across subjects so that each problem had 8 subjects assigned to each of the five versions. The order in which the 10 problems was presented was determined randomly for each subject separately. Experiment 2 had the same procedure as in Experiment 1 except that the subjects received 10 problems instead of 12 problems.

Results and discussion. Table 5 presents the mean number of questions generated per problem. As predicted by the anomaly hypothesis, the mean of the four transformed versions combined (2.32) was significantly higher than the mean of the original versions (2.05), $t(39) = 2.76$, $p < .05$. More detailed comparisons indicated that the deletion and salient irrelevancy versions significantly differed from the original versions, whereas the contradiction and subtle irrelevancy versions did not significantly differ from the original.

Table 5 also presents the number of transformation-relevant questions per problem. Each of these scores significantly differed from 0 (the baserate for the original version), which is consistent with the anomaly hypothesis. The obstacle hypothesis predicts that the deletion and cont diction versions should produce more tranformation-relevant questions than the two irrelevancy versions; this trend was supported by the data in a planned comparison between the two pairs of conditions, .68 versus .31, $t(1, 39) = 5.78$, $p < .05$. Post hoc comparisons revealed that the deletion versions had higher scores than the other three versions. The contradiction versions had higher scores than the subtle irrelevancy versions, $t(39) = 2.45$, $p < .05$, but not the salient irrelevancy versions.

Table 5 also presents the likelihood that a particular transformation was detected by the subjects, as manifested by at least one transformation-relevant question. These likelihood scores were higher for the deletion and contradiction versions (mean = .41) than the two irrelevancy versions (mean = .27), $t(39) = 2.85$, $p < .05$. This outcome is consistent with the obstacle hypothesis. However, there was absolutely no difference between the contradiction and salient irrelevancy versions so there is only weak support for this hypothesis when fine-grained comparisons are made between individual conditions. Once again, the most robust difference was between the deletion versions and the other three types of transformations.

As in Experiment 1, we performed an analysis on the time-course of the transformation-relevant questions. We found that the vast majority of transformation-relevant questions occurred during the initial 90-second question asking phase (77%) rather than the 150-second problem solving phase (23%). This 77% figure compares favorably to the 75% figure in Experiment 1. Therefore, the questions were inspired quite early in the problem solving process. It is important to acknowledge, however, that this outcome might be limited to problems that are comparatively easy to solve. Perhaps more questions would be generated during later stages of problem solving to the extent that problems are more difficult.

## Experiment 3: Problem solving while comprehending stories

Experiments 1 and 2 contained quantitative problems that subjects were expected to solve. The subjects' problem solving task was well-defined in the sense that they were required to obtain solutions to objective problems involving statistics, algebra, and logic. At least for the original versions, the problem representations were complete, precise, and self-contained. In contrast, the task and materials in Experiment 3 were comparatively ill-defined. The subjects' task was to comprehend a set of stories and to generate questions during the process. The comprehension task is ill-defined because there are many levels of comprehension that could be achieved and sometimes multiple ways to interpret each level (Kintsch & van Dijk, 1978; Kintsch, 1988). Compared to word problems, stories are incomplete, imprecise, and open-ended. As a consequence of this flexibility in interpretation, it might be more difficult for the subjects to identify anomalous information. Experiment 3 permited us to assess the generality of the findings in the previous two experiments.

Methods. The subjects were 40 lower division students enrolled in an introductory psychology course at Memphis State University. The subjects were instructed to read each story carefully and to write down questions that came to mind as they comprehended each story. The subject were provided approximately 5 minutes to comprehend and write down questions for each story.

There were 10 original stories (5 parables and 5 fables) which were selected from anthologies of classical or famous writers. The stories were not extremely popular, so there was a low likelihood that the subjects had already read them. The stories had a short length of approximately 100 words.

The four types of anomalous transformations were composed according to particular criteria. A critical piece of information was removed in the deletion versions. The point or moral of the story was difficult or impossible to grasp without this information. In the contradiction versions, there as an added contradictory statement that was directly incompatible with a major idea in the original story. The contradiction would make it difficult to understand the point or moral of the story. In the salient irrelevancy versions, the added sentence conveyed an episode that would probably not occur in the story (and this would be obvious to most adult readers). In the subtle irrelevancy versions, the added episode might have occurred in the story, but was irrelevant to the main plot and point of the story.

A 10-page booklet was prepared for each subject, with one story per page. Half of the stories were adult fables and half were parables, although the distinction is not important for the purposes of this report. Each subject received one of the 5 versions within the set of fables and one of the five versions within the set of parables. The assignment of versions to stories was counterbalanced across subjects. The order of stories was randomized for each subject separately

Results and discussion. The results of this study essentially replicated those in Experiment 2 even thought the materials were radically different. Table 5 presents the results of both Experiments 2 and 3. The data are strikingly similar.

The anomaly hypothesis was once again supported by two findings. First, the four types of anomalous transformations as a group (mean = 3.37 questions) had significantly more questions than did the original stories (2.99 questions), $t(39) = 2.09$, $p < .05$. Second, the mean number of transformation-relevant questions was significantly higher than 0 in all four anomalous versions.

As predicted by the obstacle hypothesis, the mean number of transformation-relevant questions in the deletion and contradiction versions combined (mean = .50) was significantly higher than that of the two irrelevancy versions (mean = .35), $t(39) = 1.70$, $p < .05$. More detailed comparisons indicated that the deletion versions had significantly more transformation-relevant questions than each of the other three transformed versions. In contrast, the contradiction versions were not significantly higher than the two irrelevancy versions.

In conclusion, the impact of the anomalous transformations on question asking appear to be quite general. The reported results generalize to both well-defined and ill-defined tasks and materials.

Experiment 4: Self-induced question asking while solving quantitative problems

The subjects in the previous three experiments were instructed to generate questions while they solved problems or comprehended stories. The purpose of Experiment 4 was to examine the incidence of questions under "self-induced" rather than "task-induced" conditions. Specifically, the subjects solved the same problems that subjects solved in Experiment 2, but they were not instructed to generate questions. While solving the problems, they had the opportunity to ask questions of an experimenter in an adjacent room, but only if they needed or wanted to ask a question. The experimenter recorded any questions that the subjects asked.

Experiment 4 permitted us to assess the extent to which social constraints present barriers to asking questions. In Experiment 2, we reported the likelihood that subjects could cognitively generate and articulate questions that addressed the transformations. These likelihood scores varied from .23 to .50 (mean = .34), depending on the transformed version. Only a subset of these questions presumably would be asked once social constraints are taken into consideration under self-induced questioning conditions.

Van der Meij (1988) identified several social barriers to question asking in the classroom. These include: (a) interruption of the flow of conversation, (b) difficulty in setting up the context for the question, (c) the social context not being viewed as a help context, (d) the teacher not having the competence to answer the question, and (e) the question not being relevant to the current topic. We attempted to set up a context in the present experiment that removed most, if not all, of these barriers. The fact that the experimenter mentioned to the students that they were free to walk into an adjacent room and ask the experimenter questions would presumably remove barriers a, c, and e. The fact that the experimenter designed the problems and administered the booklet would presumably remove barriers b and d. We implemented additional methods of removing social barriers. The subjects worked on the problems at a table in groups of 3-6 individuals. One of the individuals was a confederate who got up from the table and asked the experimenter a question on three occasions during the 1-hour session. Therefore, there was a social precedence for asking the questions. In summary, we designed the context in Experiment 4 in a manner that removed many social barriers to asking questions.

Methods. The subjects were 25 undergraduate students at Memphis State University who participated in order to fulfill a psychology course requirement. The materials included the same conditions, booklets, and counterbalancing that was used in Experiment 2.

The experimenter passed out the booklets and instructed the students to solve the problems at their own pace. They were told to write down the answers to the problems in the booklets, as well as any notes they needed to assist them in solving the problems. The experimenter mentioned that they might have questions about the problems during the experiment; if so, they were free to ask the experimenter questions in an adjacent room. The subjects were seated around a table in groups of 3-6. One of the individuals was a confederate who asked the experimenter questions on three occasions during the experiment.

The experimenter tape recorded all questions that the subjects asked in the adjacent room. The experimenter made sure to mention the subject number and the problem number during the course of the conversation. These conversations were transcribed and analyzed.

Results and discussion. There essentially was a floor effect in the incidence of questions under self-induced questioning conditions. The overall mean numbers of questions per problem were .04, .06, .12, .04, and .08 in the original, deletion, contradiction, salient irrelevancy, and subtle irrelevancy conditions, respectively. The collective mean of the four transformed versions (.08) was not significantly different from the mean of the original versions, an outcome which could be attributed to the obvious floor effect. Most of the questions (73%) in the four transformed versions were transformation-relevant questions. This indicates that there typically was an anomaly in the problem before a subject expended the effort in asking the experimenter a question. It should be noted that the corresponding percentage was 22% in Experiment 2, where there was task-induced question asking.

The mean numbers of transformation-relevant questions were .06, .12, .00, and .04 in the deletion, contradiction, salient irrelevancy, and subtle irrelevancy conditions, respectively. In spite of the obvious floor effect, the 9 transformation-relevant questions in the deletion and contradiction versions was significantly greater than the 2 transformation-relevant questions in the two irrelevancy versions, chi-square (1) = 4.45, $p < .05$. Therefore, the obstacle hypothesis was supported in spite of the obvious floor effect. As in the previous experiments, we computed the likelihood that a particular transformed problem was detected by a subject, as manifested by one or more transformation-relevant question for a particular problem. These scores were .04, .08, .00, and .02 in the respective conditions.

### Experiment 5: Anomaly detection in quantitative problems

Experiment 5 examined the relationship between detecting an anomaly in a problem and articulating a question about an anomaly. The subjects in this experiment worked on each problem and then gave an "anomaly detection" rating as to whether there was something unusual or irregular about the problem. The instructions on this scale were purposely left vague in order to avoid emphasizing one or more types of anomalous transformations. The rating scale discretely segregated YES from NO responses, with varying degrees of confidence. The problems in Experiment 5 were the same as those in Experiments 2 and 4.

An *anomaly detection likelihood* was computed which corrects for response bias, e.g., the inclination to say YES to all problems. The baserate likelihood of saying YES was manifested in the original versions of the problems. Formula 1 presents the measure of anomaly detection likelihood, which corrects for response bias.

$$\text{Anomaly detection likelihood} = [p(\text{YES}|\text{transformed version}) - p(\text{YES}|\text{original version})] / [1.0 - p(\text{YES}|\text{original version})] \qquad (1)$$

It is possible to detect that something is unusual or irregular about a problem, but not be able to articulate a question that addresses the problem. To the extent that this is the case, then the anomaly detection likelihood should be greater than the likelihood of posing a transformation-relevant question. Conversely, a person might not detect an anomaly until the person is forced to ask questions about the problem. Stated differently, the act of expression might be instrumental to the identification of an anomaly. To the extent that this alternative is the case, then the anomaly detection likelihood should be less than the likelihood of posing a transformation-relevant question.

*Methods*. The subjects were 25 undergraduate students at Memphis State University who participated to fulfill a requirement in a psychology course. The problems were the same as those used in Experiments 2 and 4.

The booklet contained 10 pages with problems, interleaved with 10 pages with rating scales. That is, a set of rating scales was completed after each problem was solved. There actually were four rating scales on each test page but only the "anomaly detection" scale is relevant to this study. This scale had the question "Does this problem have something unusual or irregular about it?" and the following six points: (1) definitely no, (2) moderately certain no, (3) undecided but guess no, (4) undecided but guess yes, (5) moderately certain yes, and (6) definitely yes.

The subjects were instructed that they would attempt to solve a series of problems for about 5 minutes per problem. They would also rate each problem on a number of scales after they attempted to solve each problem. The subjects wrote down their notes and their answers on the sheets with the problems. After attempting to solve each problem for 4 minutes, they rated the problem on four 6-point scales.

Results and discussion. Mean anomaly detection ratings were computed for each of the four versions. There were significant differences among the ratings in the five conditions, with means of 2.66, 3.86, 3.44, 2.62, and 2.52 in the original, deletion, contradiction, salient irrelevancy, and subtle irrelevancy conditions, respectively, $F(4, 96) = 7.06$, $p < .05$. As predicted by the anomaly hypothesis, the mean of the four transformed versions was significantly higher than the rating of the original, 3.11 versus 2.66, respectively, $t(24) = 1.93$, $p < .05$. As predicted by the obstacle hypothesis, the mean of the deletion and contradiction versions combined was significantly higher than the mean of the other three conditions combined, 3.65 versus 2.60, $t(24) = 4.40$, $p < .05$. Post hoc comparisons were consistent with the following pattern: deletion > contradiction > original = salient irrelevancy = subtle irrelevancy. We analyzed the "proportion of YES" responses and found precisely the same pattern of significant differences; the mean proportions were .28, .62, .46, .28, and .20 in the original, deletion, contradiction, salient irrelevancy, and subtle irrelevancy conditions, respectively.

We computed anomaly detection likelihoods, as specified in Formula 1. These scores were .47, .25, .00, and -.11 in the deletion, contradiction, salient irrelevancy, and subtle irrelevancy conditions, respectively. An ANOVA and planned comparisons showed the following pattern among the means: deletion > contradiction > salient irrelevancy = subtle irrelevancy versions. These anomaly detection likelihood scores can be meaningfully compared to the likelihood scores for asking transformation-relevant questions in Experiment 2 (i.e., line 3 in Table 5). The question asking likelihood scores were .50, .31, .31, and .23 in the deletion, contradiction, salient irrelevancy, and subtle irrelevancy versions, respectively. When considering the deletion and contradiction versions, the anomaly detection likelihoods did not significantly differ from the question asking likelihoods. In contrast, the anomaly detection likelihoods were essentially zero and significantly lower than the question asking likelihoods in the two irrelevancy versions. Therefore, the irrelevancies were not detected unless the subjects were forced to ask questions about the versions with irrelevancies.

## General discussion and conclusions

The five experiments in Section IV have consistently supported the anomaly hypothesis, which predicts that individuals will ask more questions when there are anomalous transformations of original problems or stories. These anomalous transformations included deletions of critical information, contradictions, salient irrelevancies, and subtle irrelevancies. The finding that the transformed versions triggered more questions than the original versions was a consistent finding across different types of material: difficult statistics problems, easy algebra problems, analytical brain teasers, and stories. Therefore, the trends exist both for complete, precise, self-contained problem representations and for incomplete, imprecise, open-ended stories. Indeed, these results appear to be quite general, and are compatible with cognitive models of question asking which have emphasized the importance of anomalies in stimulating questions (Collins, 1985; Graesser, Person, & Huber, 1992, in press; Kass, 1992; Ram, 1990; Schank, 1986).

The fact that the anomalous transformations triggered questions supports the claim that students to some extent are able to monitor their comprehension and to self-regulate their deficiencies in knowledge. Researchers have frequently reported that these metacognitive abilities are prevalent in good comprehenders (Brown et al., 1983; Chi et al.,1989; Flavell, 1978; Zimmerman, 1989) and that students frequently need to be trained how to use these strategies (Bransford et al., 1985; Collins, 1985, 1988; Palinscar & Brown, 1984; Pressley et al., 1989). The present study is encouraging to the extent that we have documented that the vast majority of college students *can* ask questions that address knowledge deficits if they are forced to and if the material is not extremely difficult. The results of the tutoring data in Section II also revealed that there were some attempts to correct knowledge deficits, so once again, there is some hope that college students are capable of inquiring about deletions, gaps, anomalies, and other problems in their knowledge base. The present findings also bolster the models of cognition that assert that question generation is a critical component in comprehension, learning, problem solving, and intelligence (Collins et al., 1980; Schank, 1986; Sternberg, 1987).

The obstacle hypothesis predicted that there would be more transformation-relevant questions asked in the deletion and contradiction versions than the two irrelevancy versions. Deletions and contradictions present an obstacle to the goal of solving a problem or comprehending a text, whereas there is no clear-cut obstacle in the irrelevancy versions. We found weak support for this predicted trend. The deletion and contradiction versions together did produce significantly more transformation-relevant questions than did the two irrelevancy versions. Deletion versions consistently produced more questions than did the two irrelevancy versions, but the contradiction versions sometimes had question generation rates comparable to the irrelevancy versions. For some reason, contradiction versions were not as disruptive to comprehension as we expected. Perhaps individuals discount or rationalize away these contradictions, as some researchers have suspected in the "contradiction detection" paradigm (Baker, 1979; Epstein et al.,1984; Glenberg et al., 1982; Markman, 1979; Otero & Companario, 1990). For example, the contradictions might be attributed to a misprint or to missing assumptions that would resolve the apparent contradiction. It should be noted that subjects did have a comparatively high likelihood of detecting these contradictions (see results of Experiment 5) even though they did not always ask questions about the contradictions.

There frequently are social barriers to asking questions even when individuals can detect anomalies and can articulate questions (van der Meij, 1987, 1988). Experiment 4 dramatically demonstrated the powerful impact of social barriers. The likelihood that students asked a question about an anomaly in a self-induced setting was extremely low (4%) compared to the likelihood of asking a question in a task-induced setting that forced students to ask questions (34%, Experiment 2). It should be noted that Experiments 2 and 4 had the same materials and subject population. Moreover, Experiment 4 was designed to remove social barriers by having a confederate get up from the table and ask the experimenter questions on three occasions during the experiment. Even when we removed several social barriers, subjects rarely asked a question in a self-induced condition in which they were capable cognitively of articulating a good quesiton (4%/34%=.12). This suggests that social barriers, rather than cognitive barriers, provide the most feasible explanation of the well-documented phenomenon that students rarely ask questions in a classroom (Dillon, 1988; Kerry, 1987). According to available estimates, an average student asks .2 questions per hour in a classroom whereas a student asks 20 questions per hour in a tutoring session (Graesser, Person, & Huber, in press; Section II). According to the results in Section III, a student asks 135 questions per hour on a computer system that is built around student questions and computer answers, e.g., educational software with a "Point & Query" interface (Graesser, Langston, & Baggett, in press).

The results from these experiments are consistent with a model of question asking that has three components: (1) anomaly detection, (2) question articulation, and (3) social editing. The anomaly detection process identifies gaps, contradictions, irregularities, and other knowledge deficits, as discussed by Graesser, Person, and Huber (1992, in press). The question articulation process constructs a question from a knowledge deficit that gets detected. The social editing process assesses the costs and benefits of asking a question in the particular social setting, as discussed by van der Meij (1987, 1988). Each of these components presents a potential barrier to asking a question.

It is important to acknowledge that the three components of the model may be executed in an interactive fashion rather than a sequential fashion. It would be tempting to propose a sequential stochastic model in which anomaly detection precedes question articulation, which in turn precedes social editing. However, the sequential model does not explain some of the findings of this study. For example, consider the anomaly detection likelihoods in Experiment 5 (an index of the anomaly detection process) and the likelihoods of generating transformation-relevant questions in Experiment 2 (an index of the question articulation process). In the case of deletion and contradiction transformations, these two likelihoods were about the same; the mean anomaly detection likelihood was .36 whereas the question generation likelihood was .40. In contrast, for the two irrelevancy versions, the anomaly detection likelihood was essentially zero whereas the question generation likelihood was .27. Therefore, the subjects might not have detected these irrelevancies unless they were forced to ask questions. Stated differently, question asking is sometimes a necessary prerequisite to detecting an anomaly, as opposed to vice versa.

In a similar fashion, the social editing process may interact with both anomaly detection and question articulation. Consider the questions that passed the social editing process in Experiment 4, i.e., students asking questions in the self-induced condition. Nearly all of these questions were about deletions and contradictions, the two types of transformations that presented obstacles and that were most likely to be detected. Most of the self-induced questions were transformation-relevant questions rather than frivolous questions; in contrast most of the questions under forced questioning were frivolous questions rather than transformation-relevant questions. Quite clearly, when a student passes through the social barriers and asks a question, it better be a good question.

## References

Allen, J. (1983). Recognizing intentions from natural language utterances. In M. Brady, & R. C. Berwick (Eds.), Computational models of discourse. Cambridge, MA: MIT press.

Allen, J. (1987). Natural language understanding. Menlo Park, CA: Benjamin-Cummings.

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. Cognitive Science, 13, 467-505.

Appelt, D. E. (1984). Planning English referring expressions, Artificial Intelligence, 26, 1-33.

Bach, K., & Harnish, R. M. (1979). Linguistic communication and speech acts. Cambridge, MA: MIT Press.

Baker, L. (1979). Comprehension monitoring: Identifying and coping with text confusions. Journal of Reading Behavior, 11, 363-374.

Baker, L., & Brown, A. L. (1980). Metacognitive skills in reading. In D. Pearson (Ed.), Handbook of reading research. New York: Plenum.

Bamber, D. (1990). Knowledge acquisition for the development of expert systems: An analysis. Technical Document 1322 of the Naval Ocean Systems Center, San Diego, CA.

Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 4-16.

Blum-Kulka, S., & Weizmann, E. (1988). The inevitability of misunderstandings: Discourse ambiguities. Text, 8, 219-241.

Bransford, J. D., Arbitman-Smith, R., Stein, B. S., & Vye, N. J. (1985). Analysis--improving thinking and learning skills: An analysis of three approaches. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), Thinking and learning skills,(Vol. 1). Hillsdale, NJ: Erlbaum.

Brown, A. L. (1988). Motivation to learn and understand: On taking charge of one's own learning. Cognition and Instruction, 5, 311-321.

Brown, A. L. Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markman (Eds.), Handbook of child psychology (4th ed.).cognitive development (Vol. 3). New York: Wiley.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. Cognitive Science, 4, 379-426.

Bruce, B. C. (1982). Natural communication between person and computer. In W. G. Lehnert & M. H. Ringle (Eds.), Strategies for natural language processing. Hillsdale, NJ: Erlbaum.

Burbules, N. C., & Linn, M. C. (1988). Response to contradiction: Scientific reasoning during adolescence. Journal of Educational Psychology, 80, 67-75.

Carlsen, W. S. (1991). Questioning in classrooms: A sociolinguistic perspective. Review of Educational Research, 61, 157-178.

Carroll, J. M., Mack, R. L., Lewis, C. H., Grischkowsky, N. L., & Robertson, S. R. (1985). Exploring a word-processor. Human-Computer Interaction, 1, 283-307.

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

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), Discourse production and comprehension (pp. 1-40). Norwood, NJ: Ablex.

Clark, H. H. (1979). Responding to indirect speech acts. Cognitive Psychology, 11, 430-477.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. Cognitive Science, 13, 259-294.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. American Educational Research Journal, 19, 237-248.

Cohen, P. R., Perrault, C. R., & Allen, J F. (1982). Beyond question answering. In W. G. Lehnert & M. H. Ringle (Eds.), Strategies of natural language comprehension. Hillsdale, NJ: Erlbaum.

Collins , A. (1988). Different goals of inquiry teaching. Questioning Exchange, 2, 39-45.

Collins, A. (1985). Teaching and reasoning skills. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), Thinking and learning skills (Vol. 2). Hillsdale, NJ: Erlbaum.

Collins, A., Brown, J. S., & Larkin, K. M. (1980). Inference in text understanding. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), Theoretical issues in reading comprehension. Hillsdale, NJ: Erlbaum.

Collins, A., Warnock, E. H., Aiello, N., & Miller, M. L. (1975). Reasoning from incomplete knowledge. In D. G. Bobrow & A. Collins (Eds.), Representation and understanding. New York: Academic Press.

Coombs, M. H., & Alty, J. L. (1980). Face-to-face guidance of university computer users--II: Characterizing advisory interactions. International Journal of Man-Machine Studies, 12, 407-429.

D'Andrade, R. G., & Wish, M. (1985). Speech act theory in quantitative research on interpersonal behavior. Discourse Processes, 8, 229-259.

Dahlgren, K. (1988). Naive semantics for natural language understanding. Boston: Kluwer Academic Press.

Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. Journal of Educational Psychology, 78, 256-262.

Dillon, J. T. (1984). The classification of research questions. Review of Educational Research, 54, 327-361.

Dillon, J. T. (1987). Question-answer practices in a dozen fields. Questioning Exchange, 1, 87-100.

Dillon, J. T. (1988). Questioning and teaching: A manual of practice. New York: Teachers College Press.

Edwards, D., & Mercer, N. M. (1989). Reconstructing context: The conventionalization of classroom knowledge. Discourse Processes, 8, 229-259.

Epstein, W., Glenberg, A. M , & Bradley, M. M. (1984). Coactivation and comprehension: Contribution of text variables to the illusion of knowing. Memory and Cognition, 12, 355-360.

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

Fishbein, H. D., Eckart, T., Lauver, E., Van Leeuwen, R., Langmeyer, D. (1990). Learners' questions and comprehension in a tutoring setting. *Journal of Educational Psychology*, *82*, 163-170.

Flammer, A. (1981). Towards a theory of question asking. *Psychological Research*, *43*, 407-420.

Flavell, J. H. (1978). Metacognitive development. In J. M. Scandura & C. J. Brainerd (Eds.), *Structural process theories of complex human behavior*. Leyden: Sijthoff.

Fox, B. (1988). Cognitive and interactional aspects of correction in tutoring. Technical report #88-2, University of Colorado, Boulder, Colorado.

Francik, E. P., & Clark, H. H. (1985). How to make requests that overcome obstacles to compliance. *Journal of Memory and Language*, *24*, 560-588.

Gall, M. D. (1970). The use of questions in teaching. *Review of Educational Research*, *40*, 707-721.

Gavelek, J. R., & Raphael, T. E. (1985). Metacognition, instruction, and the role of questioning activites. In D. L. Forrest-Pressley, G. E. Mackinnin, & T. G. Waller (Eds.), *Metacognition, cognition, and human performance* (Vol. 2, pp. 103-136). Orlando, FL: Academic Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Gibbs, R. W., & Mueller, R. A. G. (1988). Conversational sequences and references for indirect speech acts. *Discourse Processes*, *11*, 101-116.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory and Cognition*, *10*, 597-602.

Goffman E. (1974). *Frame analysis*. Cambridge, MA: Harvard University Press.

Good, T. L., Slavings, R. L., Harel, K. H., & Emerson, H. (1987). Student passivity: A study of question-asking in K-12 classrooms. *Sociology of Education*, *60*, 181-199.

Gordon, P., & Lakoff, G. (1971). Conversational postulates. *Papers from the seventh regional meeting, Chicago Linguistics Society*, *7*, 63-84.

Graesser, A. C., & Clark, L. F. (1985). *Structures and procedures of implicit knowledge*. Norwood, NJ: Ablex.

Graesser, A. C., & Franklin, S. P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes*, *13*, 279-303.

Graesser, A. C., & Hemphill, D. (1991). Question answering in the context of scientific mechanisms. *Journal of Memory and Language*, *30*, 186-209.

Graesser, A. C., Gordon, S. E., & Brainerd, L. E. (1992) QUEST: A model of question answering. *Computers and Mathematics with Applications*.

Graesser, A. C., Lang, K. L., & Roberts, R. M. (1991). Question answering in the context of stories. *Journal of Experimental Psychology: General*, *120*, 254-277.

Graesser, A. C., Langston, M. C., & Baggett, W. B. (in press). Exploring information about concepts by asking questions. In G. V. Nakamura & R. M. Taraban (Eds.), *Acquisition, representation, and processing of categories and concepts: The contributions of exemplars and theories*. San Diego: Academic Press.

Graesser, A. C., Langston, M. C., & Lang, K. L. (in press). Designing educational software around questioning. *Journal of Artificial Intelligence in Education*.

Graesser, A. C., & McMahen, C. L. (1992). Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. Unpublished manuscript, Memphis State University, Memphis, TN.

Graesser, A. C., McMahen, C. L., & Johnson, B. K. (1991). Tests of some mechanisms that trigger questions. Proceedings of the Cognitive Science Society, (pp.13-18). Hillsdale, NJ: Erlbaum.

Graesser, A. C., Person, N. K., & Huber, J. D. (1992). Mechanisms that generate questions. In T. E. Lauer, E. Peacock, & A. C. Graesser (Eds.), Questions and information systems. Hillsdale, NJ: Erlbaum.

Graesser, A. C., Person, N. K., & Huber, J. D. (in press). Question asking during tutoring and in the design of educational software. To appear in a book edited by Mitch Rabinowitz.

Graesser, A. C., Roberts, R. M., & Hackett-Renner, C. (1990). Question answering in the context of telephone surveys, business interactions, and interviews. Discourse Processes, 13, 327-348.

Green, G. M. (1989). Pragmatics and natural language understanding. Hillsdale, NJ: Erlbaum.

Greeno, J. (1982). Forms of understanding in mathematical problem solving. In S. G. Paris, G. M. Olson, & H. W. Stevenson (Eds.), Learning and motivation in the classroom. Hillsdale, NJ: Erlbaum.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and Semantics, Vol 3: Speech acts (pp. 41-58). New York: Academic Press.

Grishman, R. (1986). Computational linguistics: An introduction. Cambridge: Cambridge University Press.

Halliday, M. A. K. (1967). Transivity and theme, part 2. Journal of Linguistics, 3, 199-244.

Hammond, K. J. (1990). Case-based planning: A framework for planning from experience. Cognitive Science, 14, 385-443.

Hudson, R. A. (1975). The meaning of questions. Language, 51, 1-31.

Kaplan, S. J. (1983). Cooperative response from a portable natural language system. In M. Brady & R. C. Berwick (Eds.), Computational models of discourse. Cambridge, MA: MIT Press.

Kass, A. (1992). Question-asking, artificial intelligence, and human creativity. In T. Lauer, E. Peacock, & A. C. Graesser (Eds.), Questions and information systems. Hillsdale, NJ: Erlbaum.

Kass, R., & Finin, T. (1988). Modeling the user in natural language systems. Computational Linguistics, 14, 5-22.

Kempson, R. M. (1979). Semantic theory. Cambridge UK: Cambridge University Press.

Kerry, T. (1987). Classroom questions in England. Questioning Exchange, 1, 32-33.

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. Cognitive Science, 8, 255-274.

King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. Contemporary Educational Psychology, 14, 366-381.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A constructive-integration model. Psychological Review, 95, 163-182.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Klahr, D., & Dunbar, K. (1988). Dual search space during scientific reasoning. Cognitive Science, 12, 1-48.

Labov, W., & Fanshel, D. (1977). Therapeutic discourse: Psychotherapy as conversation. New York: Academic Press.

Lang, K. L., Graesser, A. C., Dumais, S. T., Kilman, D. (1992). Question asking in human-computer interfaces. In T. Lauer, E. Peacock, and A. C. Graesser (Eds.), Questions and information systems. Hillsdale, NJ: Erlbaum.

Lang, K. L., Graesser, A. C., & Langston, M. (1991). Question asking in electronic documentation. Proceedings of the Electronic Document Delivery Conference, (pp.201-214). East Brunswick, NJ: Belcore Documentation.

Lauer, T., Peacock, E., & Graesser, A. C. (1992). Questions and information systems. Hillsdale, NJ: Erlbaum.

Lehnert, W. G. (1978). The process of question answering. Hillsdale, NJ: Erlbaum.

Lindfors, J. W. (1980). Children's language and learning. Englewood Cliffs, NJ: Prentice Hall.

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

Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. Child Development, 50, 643-655.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. Cognition and Instruction, 7, 197-244.

McKeown, K. R. (1985). Discourse strategies for generating natural-language text. Artificial Intelligence, 27, 1-41.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. Machine Learning, 1, 47-80.

Miyake, N., & Norman, D. A. (1979). To ask a question, one must know enough to know what is not known. Journal of Verbal Learning and Verbal Behavior, 18, 357-364.

Mooney, R. J. (1990). A general explanation-based learning mechanism and its application to narrative understanding. San Mateo, CA: Morgan Kaufman.

Neber, H. (1987). Self directed questioning. Questioning Exchange, 1, 189-191.

Needham, W. P. (1990). Semantic structure, information structure, and intonation in discourse production. Journal of Memory andLanguage, 29, 455-468.

Norman, D. (1973). Memory, knowledge, and the answering of questions. In R. Solso (Ed.), Contemporary issues in cognitive psychology. Washington, DC: Winston.

Ohlsson, S., & Rees, E. (1991). The function of conceptual understanding in the learning of arithmetic procedures. Cognition and Instruction, 8, 103-179.

Otero, J. C., & Campanario, J. M. (1990). Comprehension evaluation and regulation on learning from science texts. Journal of Research in Science Teaching, 27, 447-460.

Owens, C. (1991). A functional taxonomy of abstract plan failures. The Thirteenth Annual Conference of the Cognitive Science Society (167-172). Hillsdale, NJ: Erlbaum.

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension monitoring activites. Cognition and Instruction, 1, 117-175.

Papert, S. (1980). Mindstorms: children, computers and powerful ideas. New York: Basic Books.

Piaget, J. (1952). The origins of intelligence. New York: International University Press.

Pressley, M., & Levin, J. R. (1983). Cognitive strategy training: Educational applications. New York: Springer-Verlag.

Pressley, M., Goodchild, F., Fleet, J., Zajchowski, R., & Evans, E. (1989). The challenges of classroom strategy instruction. The Elementary School Journal, 89, 301-342.

Pressley, M., Symons, S., McDaniel, M. A., Snyder, B. L., & Turnure, J. E. (1988). Elaborative interrogation facilities in the acquisition of confusing facts. Journal of Educational Psychology, 80, 301-342.

Putnam, R. T. (1987). Structuring and adjusting context for students: A study of live and simulated tutoring of addition. American Educational Research Journal, 24, 13-48.

Ram, A. (1990). Decision models: A theory of volitional explanation. In The Twelfth Annual Proceedings of the Cognitive Science Society (198-205). Hillsdale, NJ: Erlbaum.

Reder, L. M., & Cleeremans, A. (1990). The role of partial matches in comprehension: The Moses illusion revisited. In A. C. Graesser & H. Bower (Eds.), The Psychology of Learning and Motivation: Inferences and Text Comprehension (pp. 233-258). San Diego: Academic Press.

Reisbeck, C. K. (1988). Are questions just function calls? Questioning Exchange, 2, 17-24.

Resnick, L., Salmon, M. H., & Zeitz, C. M. (1991). The structure of reasoning in conversation. In The Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society (pp 388-393), Hillsdale, NJ:Erlbaum.

Schank, R. C. (1986). Explanation patterns: understanding mechanically and creatively. Hillsdale, NJ: Erlbaum.

Schank, R. C., & Abelson, R. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Erlbaum.

Schank, R., Ferguson, W., Birnbaum, L., & Greising, M. (1991). ASK TOM: An experimental interface for video case libraries. In The Proceedings of the 13th Annual Conference for the Cognitive Science Society (pp. 570-575). Hillsdale, NJ: Erlbaum.

Searle, J. R. (1969). Speech acts. London: Cambridge University Press.

Sebrechts, M. M., & Swartz, M. L. (1991). Question-asking as a tool for novice computer skill acquisition. Proceedings of the International Conference on Computer-Human Interaction, 293-297.

Shrager, J., & Callahan, M. (1991). Active language in the collaborative development of cooking skill. In Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society (pp. 394-399), Hillsdale, NJ: Erlbaum.

Singer, M., & Donlan, D. (1982). Active comprehension: Problem solving schema with question generation for comprehension of complex short stories. Reading Research Quarterly, 17, 166-186.

Souther, A., Acker, L., Lester, J., & Porter, B. (1989). Using view types to generate explanations in intelligent tutoring systems. In the Proceedings of the 11th Annual Conference of the Cognitive Science Society (pp. 123-130), Hilldale, NJ: Erlbaum.

Sternberg, R. J. (1987). Questioning and intelligence. Questioning Exchange, 1, 11-13.

Stevens, A., Collins, A., & Goldin, S. E. (1982). Misconceptions in students' understanding. In D. Sleeman & J. S. Brown (Eds.), Intelligent tutoring systems. New York: Academic Press.

Stokal, R. R. (1974). Classification. Science, 185, 115-123.

Tannen, D. (1984). Coherence in spoken and written discourse. Norwood NJ: Ablex.

Turner, E. H., & Cullingford, R. E. (1989). Using conversation MOPs in natural language interfaces. Discourse Processes, 12, 63-90.

van der Meij, H. (1987). Assumptions of information-seeking questions. Questioning Exchange, 1, 111-117.

van der Meij, H. (1988). Constraints on question asking in classrooms. Journal of Educational Psychology, 80, 401-405.

van Lehn, K. (1991). Mind bugs: The origins of procedural misconceptions. Cambridge, MA: MIT.

Webber, B. (1988). Question answering. In S. C. Shapiro (Ed.), Encyclopedia of Artificial Intelligence. New York: John Wiley and Sons.

Williams, M. D. (1984). What makes RABBIT run? International Journal of Man-Machine Studies, 21, 333-352.

Woods, W. A. (1977). Lunar rocks in natural English: Explorations in natural language question answering. In A. Zampoli (Ed.), Linguistic structures processing. New York: Elsevier North-Holland.

Yopp, R. (1988). Questioning and active comprehension. Questioning Exchange, 2, 232-238.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. Journal of Educational Psychology, 81, 329-339.

TABLE 1

Question Categories in Graesser, Person, and Huber (1992) Scheme

| Question Category | Abstract Specification | Example |
|---|---|---|
| SHORT ANSWER | | |
| Verification | Is a fact true? Did an event occur? | Is the answer five? |
| Disjunctive | Is X or Y the case? Is X, Y, or Z the case? | Is gender or female the variable? |
| Concept completion | Who? What? What is the referent of a noun argument slot? | Who ran this experiment? |
| Feature specification | What qualitative attributes does entity X have? | What are the properties of a bar graph? |
| Quantification | What is the value of a quantitative variable? How many? | How many degrees of freedom are on this variable? |
| LONG ANSWER | | |
| Definition | What does X mean? | What is a t-test? |
| Example | What is an example label or instance of the category? | What is an example of a factorial design? |
| Comparison | How is X similar to Y? How is X different from Y? | What is the difference between a t-test and an f-test? |
| Interpretation | What concept or claim can be inferred from a static or active pattern of data? | What is happening in this graph? |
| Causal antecedent | What state or event causally led to an event or state? | How did this experiment fail? |
| Causal consequence | What are the consequences of an event or state? | What happens when this level decreases? |
| Goal orientation | What are the motives or goals behind an agent's action? | Why did you put decision latency on the y-axis? |
| Instrumental/procedural | What instrument or plan allows an agent to accomplish a goal? | How do you present the stimulus on each trial? |
| Enablement | What object or resource allows an agent to perform an action? | What device allows you to measure stress? |
| Expectational | Why did some expected event not occur? | Why isn't there an interaction? |
| Judgmental | What value does the answerer place on an idea or advice? | What do you think of this operational definition? |
| Assertion | The speaker makes a statement indicating he lacks knowledge or does not understand an idea. | I don't understand main effects. |
| REQUEST/DIRECTIVE | The speaker wants the listener to perform an action. | Would you add those numbers together? |

Table 2

Type of question

| | Student | | Tutor | |
|---|---|---|---|---|
| | Normal | Structured | Normal | Structured |
| **SHORT-ANSWER QUESTIONS** | | | | |
| Verification (Yes/No) | .28 | .16 | .43 | .23 |
| Disjunctive (A or B?) | .02 | .03 | .03 | .02 |
| Concept Completion (who? what?) | .14 | .15 | .09 | .17 |
| Feature Specification | .05 | .05 | .02 | .04 |
| Quantification (how many?) | .03 | .06 | .03 | .06 |
| **LONG-ANSWER QUESTIONS** | | | | |
| Definition (what does x mean?) | .03 | .04 | .05 | .03 |
| Comparison (how does x compare to y?) | .02 | .01 | .02 | .01 |
| Example (what is an example of x?) | .01 | .02 | .01 | .01 |
| Interpretational (what's happened?) | .09 | .09 | .09 | .12 |
| Judgmental | .03 | .05 | .02 | .04 |
| Antecedent (why?) | .02 | .01 | .02 | .01 |
| Consequence (what if?) | .03 | .02 | .04 | .02 |
| Goal Orientation (why?) | .01 | .02 | .02 | .03 |
| Enablement (why/how?) | .00 | .01 | .01 | .01 |
| Instrumental/Procedural (how?) | .13 | .19 | .06 | .13 |
| Expectational (why not?) | .03 | .04 | .01 | .02 |
| **OTHER** | | | | |
| Assertion | .08 | .04 | .00 | .00 |
| Request/directive | .03 | .02 | .05 | .05 |

Table 3

Question answering strategies for why and how questions under conditions of normal tutoring

| Type of question | Answer strategy |
|---|---|
| WHY-ACTION (49%) | |
| why <agent do action A> | Superordinate goal of A (.68)<br>State/event that initiates plan of A (.42)<br>Example that justifies A (.29) |
| why <agent do action A rather than B> | Action B is not feasible or desirable (.70)<br>Action A has positive consequences (.30)<br>Hypothetical example justifying either A or B (.30) |
| WHY-EVENT (2%) | |
| WHY-NOT-X (13%) | |
| why-not *<X exist>* | |
| why-not *<agent do action A>* | |
| why-not *<event occurs>* | |
| why-not *<(X is a Y) or (X has property P)>* | Present example that illustrates it (1.00)<br>X doesn't have features of P or Y (.80) |
| WHY-STATE (36%) | |
| why *<concept/instance X is a concept Y>* | X has properties of Y (.80)<br>Specify contrast concept of Y (.30) |
| why *<concept/instance X has property P>* | X has features of property P (1.00)<br>If X exists, then features of P exists (.43)<br>If X not exist, then features of P don't exist (.14)<br>Specify contrast property of P (.29)<br>Example (.29) |
| why *<concept X is important>* | If X exists, positive consequences occur (.75)<br>If X not exist, negative consequences occur (.50)<br>X is more than alternative Y (.50)<br>Define concept X as background (.50)<br>Example (.25)<br>X is not merely Z (.25) |
| why *<X is better than Y>* | |
| why *<value on variable>* | |
| HOW-ACTION (75%) | |
| how *<agent do action A>* | Subordinate action or plan of A (.62)<br>Draw on or point to board (.27)<br>Logical derivation (.10)<br>Give example (.06) |
| HOW-EVENT (10%) | |
| how *<X affect Y>* | Specify processes during event (.60)<br>Causal chain between X and Y (.40)<br>Give example (.40)<br>Similar to Z (.40) |
| how *<event X occurs>* | Specify processes during event (1.00) |
| HOW-STATE (13%) | |
| how <person know X> | Logical trace (.50)<br>Specify properties of X (.50)<br>Causal antecedents of X (.17)<br>Point to board (.17)<br>Give evidence that X is true (.17) |

Table 4

Five Different Versions of a Mathematical Word Problem.

ORIGINAL The Bears won 40 games and lost 24, while the Bulls won 32 games and lost 18. Which team had the higher percentage of wins?

DELETION The Bears won 40 games and lost 24, while the Bulls won 32. Which team had the higher percentage of wins?

CONTRADICTION The Bears won 40 games and lost 24, while the Bulls won 32 games and lost 18. The Bulls won more games than the Bears. Which team had the higher percentage of wins?

SALIENT IRRELEVANCY The Bears won 40 games and lost 24, while the Bulls won 32 games and lost 18. Several Bear team members were always getting into arguments with some members of the Bulls team. Which team had the higher percentage of wins?

SUBTLE IRRELEVANCY The Bears won 40 games and lost 24, while the Bulls won 32 games and lost 18. Nearly 20% of the games almost resulted in ties. Which team had the higher percentage of wins?

TABLE 5

Questions generated while solving quantitative problems: Experiment 2.

|  | VERSION OF PROBLEM |  |  |  |  |
|---|---|---|---|---|---|
|  | Original | Deletion | Contradiction | Salient Irrelevancy | Subtle Irrelevancy |
| Overall number of question | 2.05 | 2.43 | 2.18 | 2.45 | 2.21 |
| Number of transformation-relevant questions |  | .90 | .46 | .35 | .27 |
| Likelihood that a subject detected a particular transformation |  | .50 | .31 | .31 | .23 |

Questions generated while comprehending stories: Experiment 3.

|  | VERSION OF PROBLEM |  |  |  |  |
|---|---|---|---|---|---|
|  | Original | Deletion | Contradiction | Salient Irrelevancy | Subtle Irrelevancy |
| Overall number of questions | 2.99 | 3.46 | 3.15 | 3.62 | 3.28 |
| Number of transformation relevant questions |  | .72 | .28 | .48 | .23 |
| Likelihood that a subject detected a particular transformation |  | .47 | .24 | .35 | .20 |

Structure of tutor-student exchange
(T=tutor; S=student)

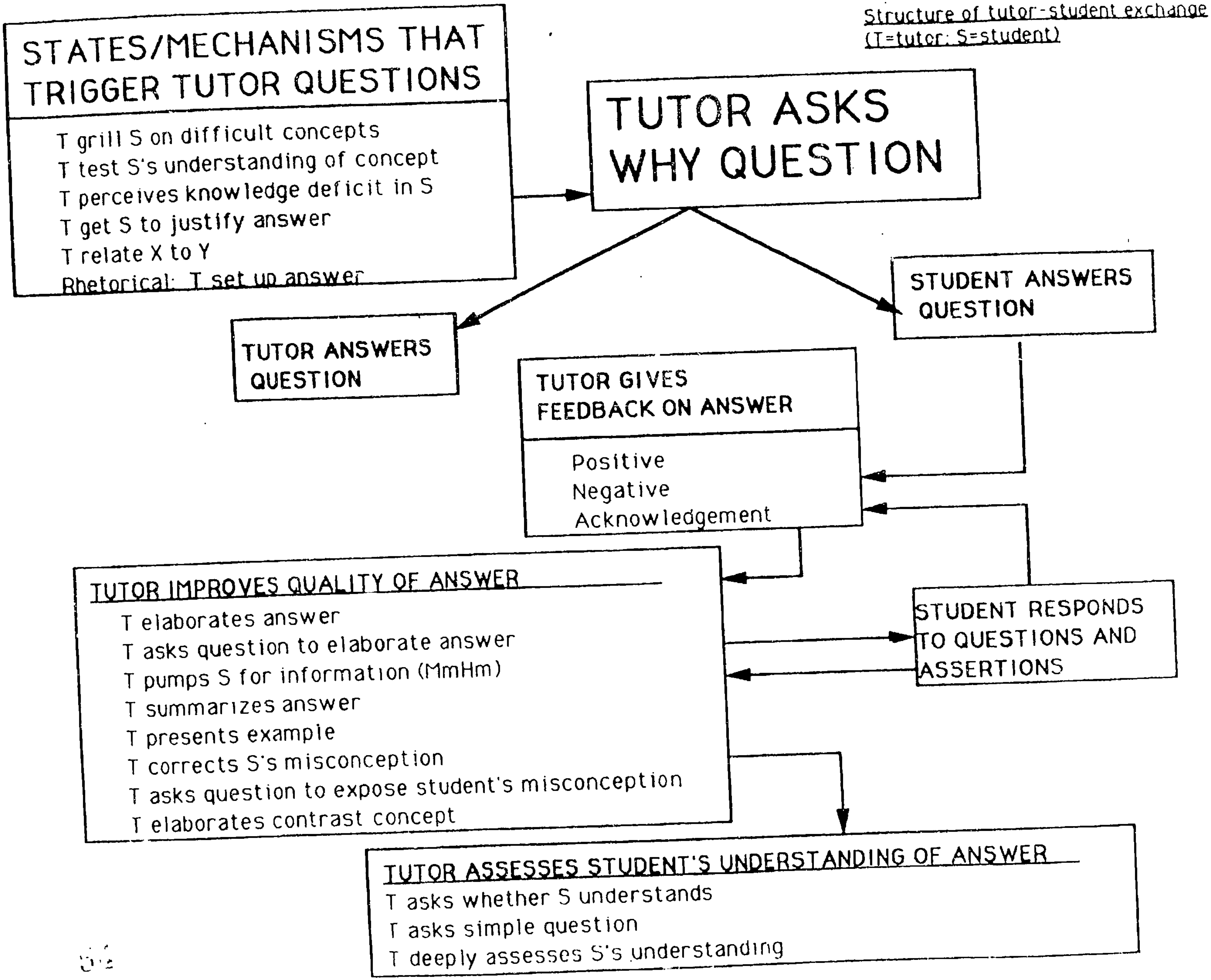
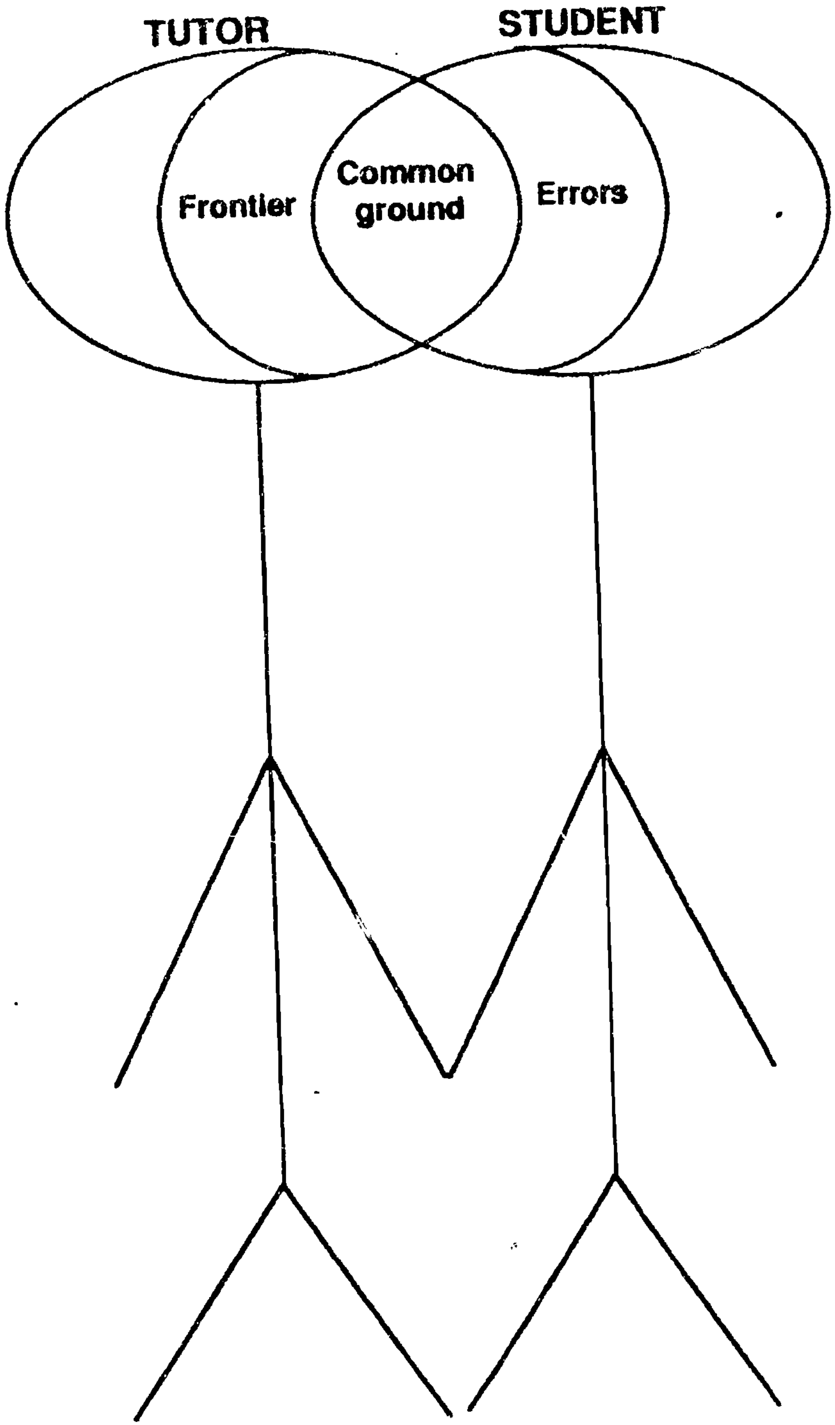

STATES/MECHANISMS THAT TRIGGER TUTOR QUESTIONS
T grill S on difficult concepts
T test S's understanding of concept
T perceives knowledge deficit in S
T get S to justify answer
T relate X to Y
Rhetorical: T set up answer
TUTOR ASKS WHY QUESTION
TUTOR ANSWERS QUESTION
STUDENT ANSWERS QUESTION
TUTOR GIVES FEEDBACK ON ANSWER
Positive
Negative
Acknowledgement
TUTOR IMPROVES QUALITY OF ANSWER
T elaborates answer
T asks question to elaborate answer
T pumps S for information (MmHm)
T summarizes answer
T presents example
T corrects S's misconception
T asks question to expose student's misconception
T elaborates contrast concept
STUDENT RESPONDS TO QUESTIONS AND ASSERTIONS
TUTOR ASSESSES STUDENT'S UNDERSTANDING OF ANSWER
T asks whether S understands
T asks simple question
T deeply assesses S's understanding

Figure 2



QUESTIONING GOALS

Quiz student on syllabus and problematic concepts

Identify common ground and student knowledge

Diagnose student errors (bugs, misconceptions)

Extend frontier

Get student to defend concepts

Clarify message

Gauge student's level of understanding

ANSWERING GOALS

Extend/elaborate frontier

Repair student errors

Justify claims

Clarify

Summarize

QUESTIONING GOALS

Correct knowledge deficits:
- contradictions
- anomalies
- obstacles to goals
- gaps in knowledge

Verify whether student's knowledge is correct

Clarify message

ANSWERING GOALS

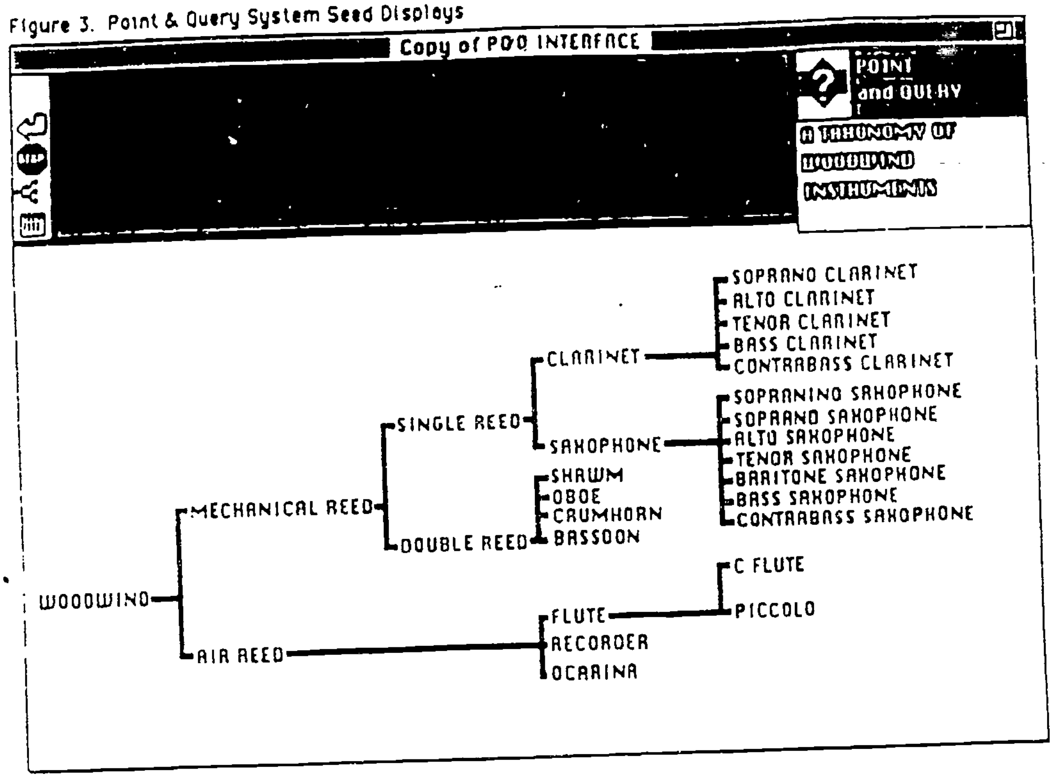
Generate correct answer

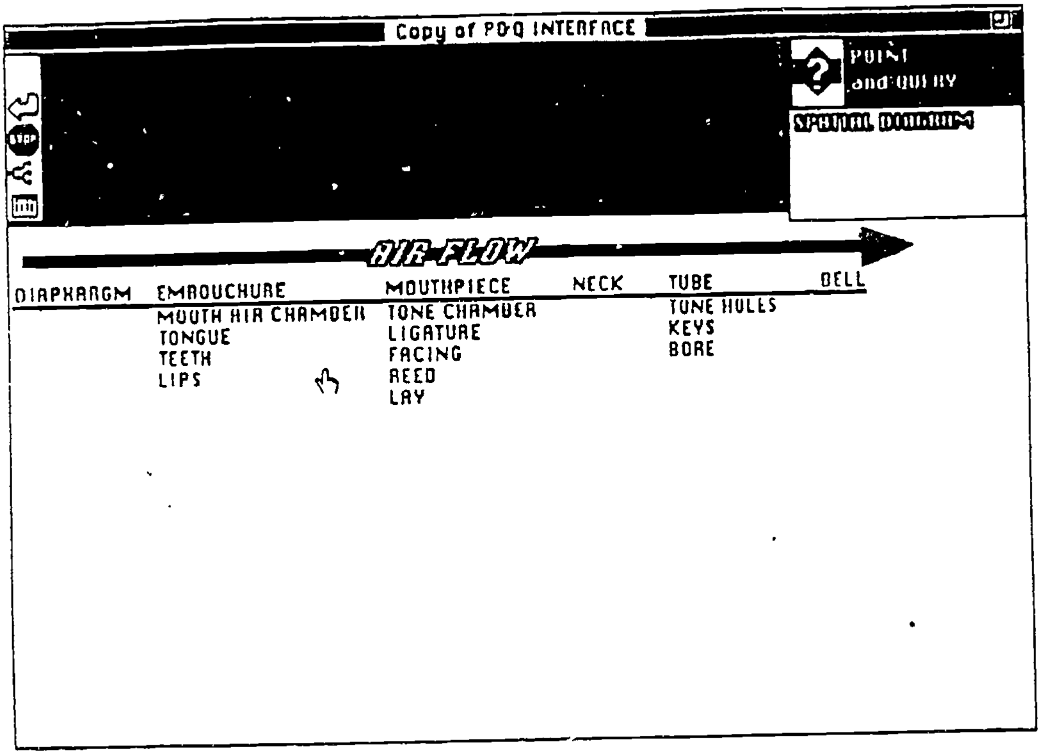
Elaborate Information

Display correct reasoning

Convince tutor that student's knowledge is adequate

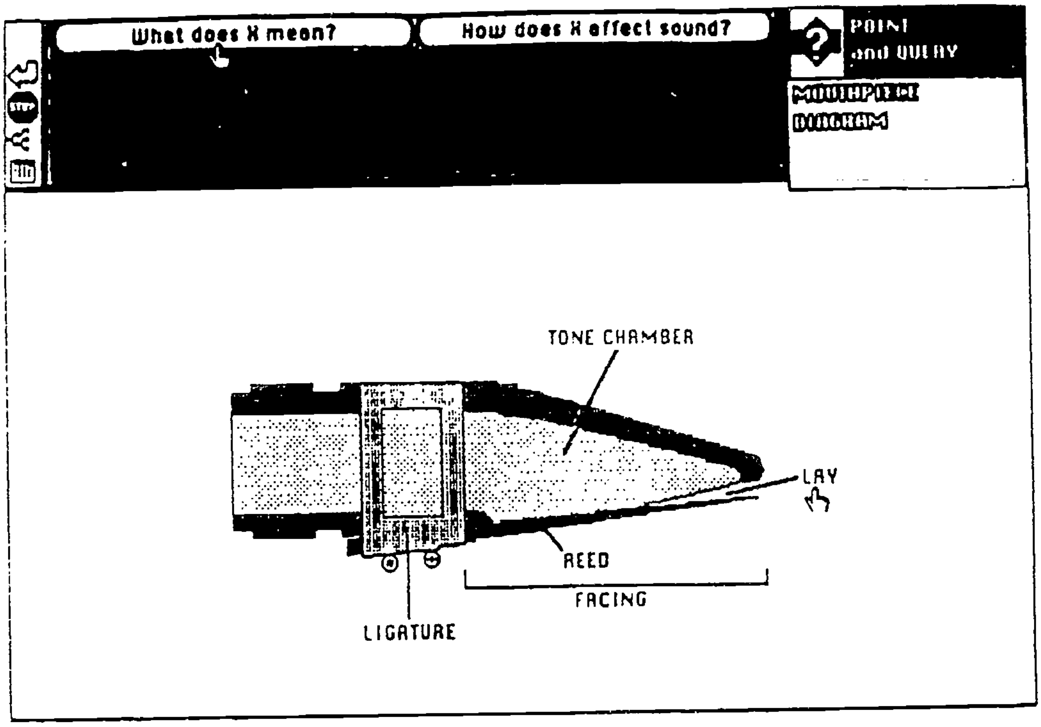
Figure 3. Point & Query System Seed Displays

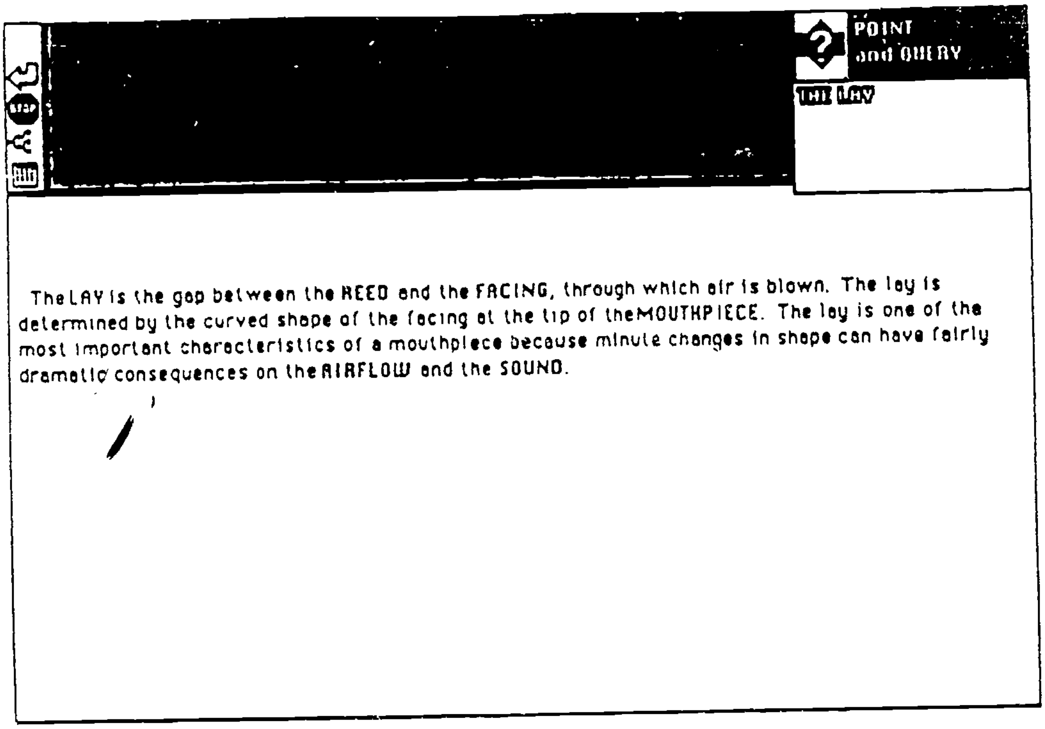




| DIAPHRAGM | EMBOUCHURE | MOUTHPIECE | NECK | TUBE | BELL |
|---|---|---|---|---|---|
| | MOUTH AIR CHAMBER | TONE CHAMBER | | TONE HOLES | |
| | TONGUE | LIGATURE | | KEYS | |
| | TEETH | FACING | | BORE | |
| | LIPS | REED | | | |
| | | LAY | | | |

Figure 4. An example question and answer on the Point & Query Interface





The LAY is the gap between the REED and the FACING, through which air is blown. The lay is determined by the curved shape of the facing at the tip of the MOUTHPIECE. The lay is one of the most important characteristics of a mouthpiece because minute changes in shape can have fairly dramatic consequences on the AIRFLOW and the SOUND.

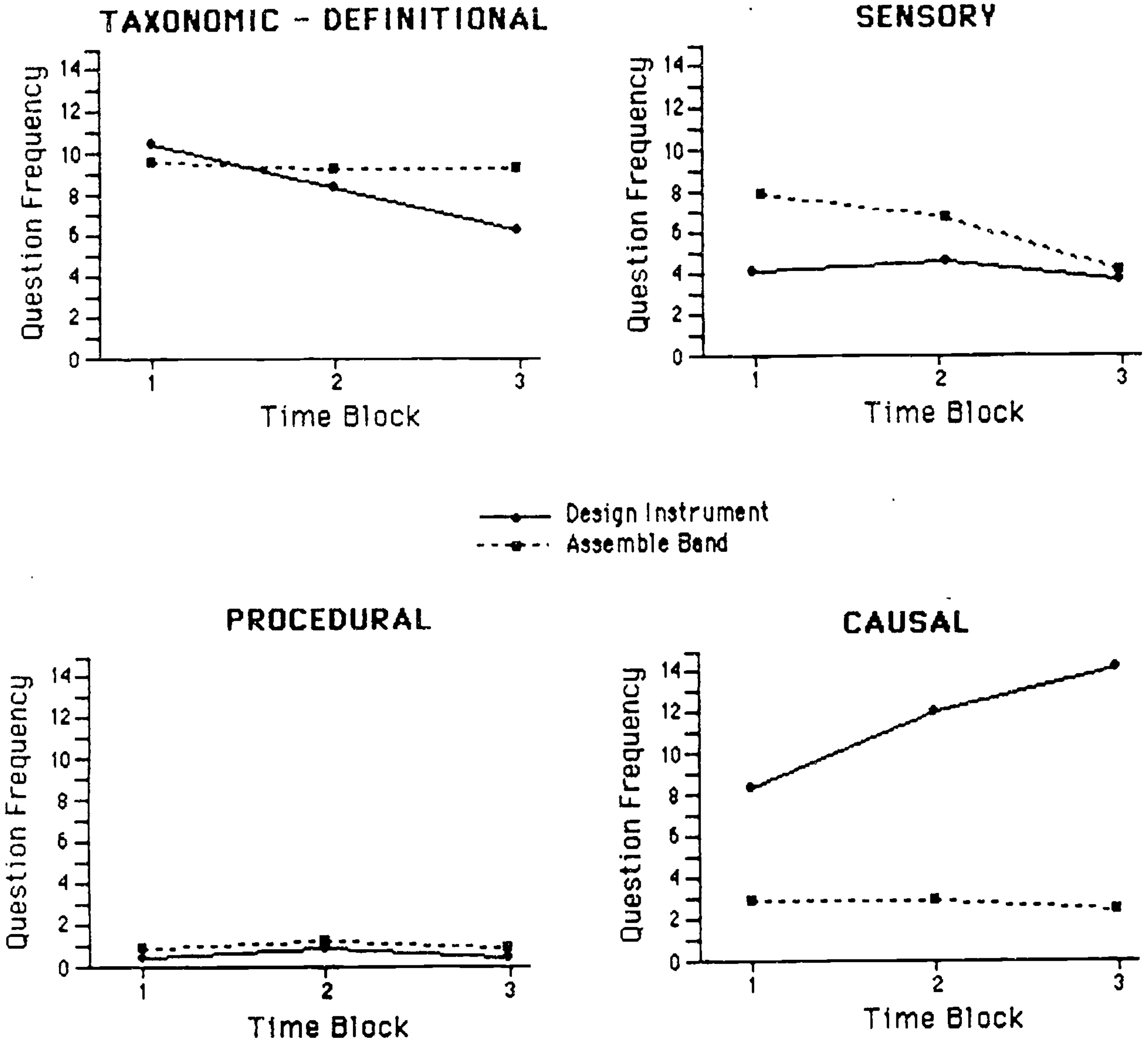


Figure 5. Number of questions asked on Point & Query System