INFERRING WHAT THE STUDENT KNOWS IN ONE-TO-ONE TUTORING: THE ROLE OF STUDENT QUESTIONS AND ANSWERS

NATALIE K. PERSON, ARTHUR C. GRAESSER, JOSEPH P. MAGLIANO, AND ROGER J. KREUZ

MEMPHIS STATE UNIVERSITY

ABSTRACT: We analyzed 44 one-to-one tutoring sessions in which undergraduates were tutored by graduate students on troublesome topics in research methods. The primary goal of this research was to determine the relationship between measures of student achievement and measures of student questions and answers. First, our results indicated that the quality of students' answers were the most reliable source for inferring student understanding. Second, the quality of the students' questions was only a marginal indicator of student understanding. And third, students' answers to tutors' comprehension-gauging questions (e.g., Do you understand?) proved to be very misleading in regard to student understanding.

The purpose of tutoring sessions is to help students learn new skills and master problematic topics. Ideally, the tutor should be able to adjust the level of instruction to the idiosyncratic needs and knowledge deficits of a particular student. This requires the tutor to have a valid way of assessing what the student understands. The developers of many intelligent tutoring systems, for example, have attempted to take the knowledge level of the student into account by designing systems that adjust instruction according to the misconceptions and reasoning strategies of a particular student (Anderson, Boyle, & Reiser 1985; Brown & Burton 1982; Clancey 1983; Dede 1986; McArthur, Stasz, & Zmuidzinas 1990; Van Lehn 1991).

Existing intelligent tutoring systems have modeled the knowledge of the stu-

dent by adopting a number of different approaches (Eberts & Brock 1988). Two of the less sophisticated approaches involve topic marking and establishing a student context model. The topic marking approach simply keeps track of all the information that has been presented to the student. The student context approach uses the student's dialogue and questions as an indicator of understanding. More sophisticated approaches include the "buggy" and "overlay" approaches. These approaches compare student knowledge to ideal expert knowledge about a particular topic. According to the "buggy" approach, student knowledge is represented in terms of misconceptions, incorrect rules, inappropriate metaphors or other knowledge that is clearly not a subset of expert knowledge. According to the "overlay" approach, the student's knowledge is considered a subset of the ideal expert knowledge. Once the student has demonstrated correct use of a new fact or procedure, this new information is presumed to be in the student's knowledge base.

Although some intelligent tutoring systems provide accurate inferences about student knowledge, it is plausible to assume that they are not as effective as expert human tutors (Clancey, Bennett, & Cohen 1982). The comparative effectiveness of human tutors versus computer tutors has not yet been demonstrated for a wide variety of tasks and topics, so it is too early to offer a firm conclusion. We will operate under the assumption that human tutors are more effective than computer tutors in assessing a student's knowledge and providing accurate feedback.

Assuming that humans provide the best informative feedback, how do human tutors infer student understanding? This question becomes nontrivial when some of the proclivities of human tutoring are considered. For example, Putnam (1987) documented that tutors follow pre-planned curriculum scripts rather than structuring the instruction to meet the needs of a particular student. In these curriculum scripts, the tutor and student work through a set of problems that is predetermined by the tutor. The goal of these curriculum scripts is to achieve correct answers to the predetermined problems. Hence, the misconceptions of a particular student are rarely diagnosed and systematically repaired.

In many cases in which tutors do inquire about student understanding, tutors rely on students' self-reports as indicators of their understanding. That is, if a student expresses that he or she understands, the tutor frequently accepts this as totally accurate and moves to the next topic. A typical exchange is illustrated in the following excerpt.

Excerpt 1

Tutor: . . . . . so that's what a main effect is. Do you understand?
Student: Yeah.

Accepting this feedback from the student is problematic because students are surprisingly unsuccessful in calibrating their own knowledge (Graesser & Person in press; Glenberg, Wilkinson, & Epstein 1982; Pressley, Ghatala, Woloshyn, & Pirie 1990). It has been well documented that students have difficulty identify-

ing contradictions and inconsistencies in orally presented messages, scientific texts, mathematical word problems and other text forms (Baker 1979, 1985; Burbules & Linn 1988; Epstein, Glenberg, & Bradley 1984; Graesser & McMahen 1993; Markman 1979; McDevitt & Carroll 1988; Otero & Campanario 1990). Students have problems distinguishing information that is relevant for a solution to a problem from information that is irrelevant to the solution (Dillon 1988). Given that students are often unaware of their knowledge deficits, tutors should be reluctant to accept a student's self-report of his or her knowledge base.

Another limitation of human tutoring arises from the normative rules of polite conversation (Brown & Levinson 1987; Graesser 1993; Grice 1975). In normal conversation, speakers typically structure the dialogue so that the impositions on the hearer are minimized (Graesser, Person, & Huber 1993). Tutors may avoid asking difficult questions that will embarrass or intimidate the student even though such questions would expose the knowledge deficits of the student. On the other side of the coin, students may be reluctant to inform the tutor that they do not understand because they do not want to appear impolite. That is, when a student acknowledges that he or she doesn't understand a topic previously covered by the tutor, this acknowledgment may serve as a form of negative feedback to the tutor about the tutor's abilities.

Aside from the politeness issue, there are other social reasons that might explain students' reluctance to seek help and ask questions. It has been well documented that student-generated questions in classrooms are both infrequent and unsophisticated (Dillon 1987, 1988; Gall 1970; Good, Slavings, Harel, & Emerson 1987; Kerry 1987; Lindfors 1980). Several researchers have documented the potential costs to posing questions in classroom settings (Newman & Goldin 1990; van der Meij 1987, 1988). Some of these costs include: (1) revealing ignorance and losing status when a bad question is posed, (2) not being viewed as an independent problem solver, (3) succumbing to the stereotypes often associated with certain abilities (e.g., girls are more reluctant to seek help with math than boys), and (4) imposing on a teacher who does not want to be interrupted. It is certainly plausible that these costs also exists in one-to-one tutoring sessions.

Consequently, one might wonder if it is possible for a tutor to make valid and reliable inferences about a student's knowledge. In tutoring sessions, the tutor can infer student understanding in one of two ways. One option is to analyze the extent to which the student is actively self-regulating his or her own learning. During self-regulated learning, the student is responsible for establishing the tutoring agenda and for posing questions that address particular knowledge deficits. Researchers in education and developmental psychology have frequently advocated educational settings that engage students in active problem solving and that train students how to acquire self-regulatory learning strategies (Bransford, Arbitman-Smith, Stein, & Vye 1985; Collins 1985, 1988; Palincsar & Brown 1984; Papert 1980; Paris & Newman 1990; Piaget 1952; Pressley, Goodchild, Fleet, Zajchowski, & Evans 1989; Pressley & Levin 1983; Zimmerman 1989). Unfortunately, active self-regulated learning rarely occurs in naturalistic tutoring sessions (Graesser 1993; Graesser & Person in press). The second option

requires that the tutor infer student understanding from the student's contribu-
tions at a micro-level (i.e., questions and answers) in the tutorial dialogue that is
structured by the tutor at a macro-level. These student contributions at a micro-
level were investigated in this study.

The primary goal of this research was to determine whether students' mastery
of the material (i.e., achievement) is reflected in their questions and answers
during one-to-one tutoring. Student achievement was measured by the stu-
dents' examination scores and the final grade received in a research methods
course. The final grade was included as a measure because it included lab re-
ports (37.5% of final grade) and homework assignments (25% of final grade) in
addition to the examination scores (37.5% of final grade). We analyzed the writ-
ten transcripts of 44 one-to-one tutoring sessions. In these tutoring sessions,
undergraduates were tutored by graduate students on troublesome topics in a
research methods course. Student understanding could in principle be inferred
from three types of student contributions: student questions, student answers,
and student answers to "comprehension-gauging" questions (e.g., as illustrated
in *Excerpt 1*). Altogether there were five measures of student questions and
answers: (1) the overall number of questions asked by the student, (2) the
proportion of student questions that addressed knowledge deficits, (3) the pro-
portion of student questions that involved deep reasoning, (4) the proportion of
positive versus negative student answers to comprehension-gauging questions
posed by the tutor (e.g., "Do you understand?"), and (5) the quality of the
students' answers to questions. The following sections describe the theoretical
schemes used for analyzing the students' questions and answers in the tutorial
dialogues.

THEORETICAL SCHEMES FOR ANALYZING QUESTIONS

Schemes for analyzing the qualitative characteristics of questions have been
proposed by researchers in education (Dillon 1987; Flammer 1981; van der Meij
1987), psychology (Clark & Schaefer 1989; Graesser, Person, & Huber 1992), and
artificial intelligence (Allen 1987; Kaplan 1983; Lehnert 1978; Schank 1986; Web-
ber 1988). This study adopted the scheme developed by Graesser et al. (1992).
The Graesser-Person-Huber (GPH) scheme was influenced by the theoretical
work in each of these fields and by preliminary analyses of the tutoring proto-
cols. We adopted this scheme for several reasons. First, the GPH scheme can
accommodate a speech act that is either an interrogative expression (an utterance
that is followed by a question mark), a genuine inquiry (see Van der Meij 1987),
or both. For example, the following expressions are all inquiries, but only the
first in an interrogative expression:

1. What is a main effect?

2. Tell me what a main effect is.

3. I need to know what a main effect is.

In an effort to be inclusive in our analysis of questions, each of these expressions would be counted as a question. Second, previous research has established that the GPH scheme is reliable and valid in categorizing questions that occur in the context of tutoring (Graesser & Person in press). Finally, two dimensions on the GPH scheme have revealed systematic relationships with achievement levels (Graesser & Person in press). These two dimensions specify (1) question content categories and (2) psychological mechanisms that trigger questions.

QUESTION CONTENT CATEGORIES

Table 1 presents the 18 question content categories in the GPH scheme. These categories are defined according to the content of the information requested rather than by the question stems (i.e., how, why, what, etc.). For example, *how* stems frequently occur in quantitative questions ("How many independent variables are in this factorial design?"), instrumental/procedural questions ("How do you compute a *t*-test?"), and antecedent questions ("How did this experiment fail?"). Defining categories on the basis of the content of the information requested rather than form is consistent with theories of question answering in the cognitive sciences (Graesser & Franklin 1990; Lehnert 1978; Schank & Abelson 1977; Souther, Acker, Lester, & Porter 1989).

Two theoretical perspectives provided most of the categories in the taxonomy. D'Andrade and Wish (1985) identified eight major speech act categories that appear to be exhaustive in categorizing virtually all speech acts that occur in conversation. The GPH taxonomy includes three of the eight speech act categories (i.e., questions, assertions, and requests/directives) because these are the only categories that represent genuine inquiries. The major category of "questions" is segregated into 16 subcategories. Most of these subcategories were extracted from Lehnert's (1978) taxonomy that categorized the questions posed by subjects after they read short narrative passages.

The question content categories vary according to the length of the expected answers. The 18 categories in Table 1 are segregated into *short-answer* questions and *long-answer* questions. Short-answer questions place fewer cognitive demands on the answerer because a satisfactory answer requires only a word or short phrase. For example, an acceptable answer to a verification question is "yes," "no," or "maybe." It should be noted that although one can imagine short-answer questions that require a considerable amount of thought to answer (e.g., Do Descartes and Kant resolve the mind-body problem in the same way?), Graesser and Person (in press) have reported that these deep short-answer questions rarely occur during tutoring. Long-answer questions tend to impose a burden on the answerer because lengthy coherent answers are expected. For example, the answer to the interpretational question "What is happening in this

TABLE 1
Question Categories in Graesser, Person, and Huber (1992) Scheme

Question Category	Abstract Specification	Example
Short Answer Question		
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 Question		
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?
*Casual Antecedent	What state or event causally led to an event or state?	How did this experiment fail?
*Casual 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 agents actions?	Why do you put decision latency on the y-axis?
*Instumental/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 these numbers together?

*Denotes deep-reasoning questions.

graph?" would involve several sentences. One might expect a good tutor to ask more long-answer questions in order to expose misconceptions in the student's knowledge. Alternatively, a student might try to avoid revealing his or her ignorance by asking long-answer questions and placing the greater cognitive burden on the tutor (Graesser, Person, & Huber 1993).

A subset of the long-answer questions in Table 1 are considered *deep reasoning* questions. Deep reasoning questions are questions that elicit patterns of reasoning in logical, causal, or goal-oriented systems. In logical reasoning, the answer

consists of premises and conclusions in a logical syllogism. In causal reasoning, the answer conveys events and states in causal chains. In goal-orientation reasoning, the answer traces the goals and planning structures of the agent. The following six categories in Table 1 are considered deep-reasoning questions: antecedent, consequence, goal orientation, instrumental/procedural, enablement, and expectational.

The six categories of deep-reasoning questions are highly correlated with the cognitively demanding levels of knowledge identified in Bloom's taxonomy of educational objectives in the cognitive domain (Bloom 1956). Graesser and Person (in press) reported that the questions classified as deep-reasoning questions were positively correlated with the cognitively demanding levels in Bloom's taxonomy, $r = .64$, $p < .05$. (For a complete description of the five values on Bloom's taxonomy, see Graesser and Person in press). In tutoring sessions, deep-reasoning questions allow the tutor and student to explore deeper levels of knowledge. One would expect good students and tutors to ask these deep-reasoning questions.

The taxonomy in Table 1 can be used as either a polythetic or a monothetic classification scheme (Stokal 1974). In a polythetic scheme, a question can be assigned to more than one category. However, in a monothetic scheme, any given speech act can be assigned to one and only one category. For example, the following questions are speech acts that could be assigned to more than one category.

Did sleep deprivation cause an increase in anxiety level? (verification + antecedent)

What are the consequences of the caffeine levels being higher for children than for adults? (comparison + consequence)

Priority rules were established in order to categorize these hybrid questions monothetically. These rules were based on the length of the answer that is typical for a particular content category. For example, answers to verification questions (e.g., "Is X true?") and concept completion questions are shorter than answers to goal orientation questions (e.g., "Why did X do Y?"). Questions that entailed shorter answers received lower precedence than those that elicited longer answers. Therefore, the two preceding example questions would be categorized as antecedent and as consequence, respectively.

QUESTION GENERATION MECHANISMS

The GPH scheme specifies four major mechanisms that trigger questions: (1) correction of knowledge deficit, (2) monitoring common ground, (3) social coordination of action, and (4) control of conversation and attention. Some of these mechanisms have been identified in computer models of question generation, problem solving, and reasoning (Kass 1992; Klahr & Dunbar 1988; Laird, Rosenbloom, & Newell 1987; Schank 1986), in theories of natural language comprehen-

sion and conversation (Clark & Schaefer 1989), and in theories of learning (Collins 1988; Palincsar & Brown 1984; Sleeman & Brown 1982). Other mechanisms were identified during the preliminary analyses of the tutoring data (Graesser, Person, & Huber 1992, 1993).

We should emphasize that the question generation mechanisms take into account the discourse context of the question. That is, each question was analyzed according to the underlying mechanisms that elicit or motivate the inquiry. Moreover, these question generation mechanisms should not be confused with the question content categories in Table 1. For example, a verification question could in principle be motivated by any of the four major question generation mechanisms.

Knowledge Deficit. The first major question generation mechanism consists of information-seeking questions that occur when the questioner detects a knowledge deficit in his or her knowledge base. These questions occur under the following conditions: (a) when the questioner encounters an obstacle in a plan or problem (e.g., What do I do next?), (b) when the questioner is missing a crucial piece of information (e.g., How do you compute the degrees of freedom?), (c) when the questioner encounters a contradiction (e.g., Doesn't the dependent variable go on the y-axis?), or (d) when the questioner must decide between two equally attractive alternatives (e.g., Should I have three or four levels on my independent variable?).

As discussed earlier, it has been argued that students should take an active role in constructing, regulating, and monitoring their own learning and comprehension activities (Bransford et al. 1985; Collins 1985, 1988; Palincsar & Brown 1984; Pressley, Goodchild et al. 1989; Pressley & Levin 1983; Zimmerman 1989). Knowledge deficit questions reflect the extent to which students take an active role in self-regulating their own knowledge. Hence, one might expect good students to ask many knowledge deficit questions. On the other hand, poor students have more knowledge deficits so they might ask more of these questions. As a third alternative, a curvilinear relationship might exist between knowledge deficit questions and student achievement; good students and poor students may ask more of these questions than intermediate level students. Finally, a fourth alternative would be that intermediate level students ask more knowledge deficit questions than good and poor students.

Common Ground. This question generation mechanism involves the monitoring of common ground between participants. In order to achieve successful communication, speech participants frequently have to establish, negotiate, or update their mutual knowledge (Clark & Schaefer 1989). In tutoring sessions, the common ground dynamically changes over the course of the tutorial dialogue. This change occurs at both the level in which new referents and ideas of the a particular conversation are introduced, as well as the more global level in which the domain knowledge that is shared by the tutor and student is developed. In order to develop this shared knowledge, students frequently ask common

ground questions in order to confirm that their own beliefs are correct (e.g., "Doesn't the dependent variable go on the vertical axis?"). Tutors, on the other hand, ask common ground questions in order to: (1) inquire whether the student knows anything about a topic (e.g., "Have you had statistics?"), (2) assess the student's knowledge about a particular topic (e.g., "What is the difference between a bar graph and a histogram?"), and (3) gauge whether the student is understanding at a global level (e.g., "Do you understand?"). This last type of common ground question is defined as a *comprehension-gauging* question.

Social Coordination. This question generation mechanism coordinates the social actions between speech participants. Social coordination questions are posed in order to manipulate agents or to coordinate their actions (Austin 1962; Clark 1979; Gibbs & Mueller 1988; Searle 1969). These questions include indirect requests (e.g., "Would you read this card?"), indirect advice (e.g., "Why don't you put this variable on the x-axis?"), permission (e.g., "Can I label the graph?"), and negotiations (e.g., "If I get the next one right, can we stop?").

Conversation Control. The fourth question generation mechanism includes questions that are asked in order to control the flow of conversation and the attention of speech participants. These questions include greetings (e.g. How's it going?), gripes (e.g., If I do one more, can we stop?), replies to summons (e.g., The tutor says "Hey, Joe!," and the student replies "What?"), and questions that change the flow of conversation (e.g., The student says "My boyfriend is a real jerk," and the tutor attempts to change the subject by asking "How do you do on the last exam?"). It should be noted that these questions are quite rare in tutoring (Graesser et al. 1993).

SCHEME FOR ANALYZING STUDENTS' ANSWERS TO TUTOR QUESTIONS

Tutors frequently ask questions that assess a student's knowledge about a particular topic. Student understanding was measured according to the quality of the students' answers to these tutor-posed questions. Each student answer was assigned to one of four qualitative categories: (1) complete answer, (2) partial answer, (3) vague answer, or (4) error-ridden answer. These categories of answers are illustrated in the following examples in the context of the tutor question "What is a *t*-test?".

Complete answer	It's a statistical procedure used to determine whether the means of two groups statistically differ.
Partial answer	It looks at the means of the groups.
Vague answer	Ya know, two samples and all.
Error-ridden answer	It tells you if there is an interaction.

One would expect that good students are more likely to provide complete and partially correct answers, whereas poor students are more likely to provide vague and error-ridden answers.

STUDENTS' ANSWERS TO COMPREHENSION-GAUGING QUESTIONS

As mentioned earlier, tutors often ask comprehension-gauging questions (see *Excerpt 1*) in order to monitor the students' knowledge at a global level. One would intuitively expect that good students answer positively more often than poor students. That is, good students should be more likely to answer "Yes, I understand." Given that poor students have more knowledge deficits than good students, one would expect the poor students to provide more negative answers to comprehension-gauging questions. That is, the poor students should be more likely to say "No, I do not understand." However, Chi, Bassok, Lewis, Reimann, and Glaser (1989) reported findings contrary to this assumption in an analysis that examined college students' self-monitoring statements while attempting to work physics problems. Their results indicated that good students are more capable of accurately monitoring their comprehension failures, whereas poor students are less accurate in detecting such failures (Chi et al. 1989). Specifically, the good students were more likely to say they did not understand than did the poor students.

Our analyses examined the proportion of positive, negative, versus "other" answers that students provided to the comprehension-gauging questions that were asked by the tutor. These proportions were correlated with measures of student achievement. It should be noted that failing to give an answer is also a possible response to a comprehension-gauging question. These "other" responses are ambiguous with respect to student understanding, however.

METHOD

STUDENTS AND TUTORS

Tutoring protocols were collected from 27 undergraduate students enrolled in a scientific research methodology class at Memphis State University. The sample included nine males between the ages of 18 and 25, twelve females between the ages 18 and 25, and six females over the age of 25. All students participated in the tutoring sessions in order to fulfill a course requirement (6% of the total points in the course). Therefore, the tutoring protocols involved a representative sample of college students rather than a sample restricted to students who were having difficulty in the course.

The tutors were three psychology graduate students who had each performed well in undergraduate and graduate level research methodology courses. Each of the tutors had tutored students on a few occasions prior to this study, but none in the area of research methods. Therefore, the tutors had a modest amount of tutoring experience, but they did not have extensive training in the tutoring process. It is important to point out that these characteristics are representative of most of the tutoring that takes place in school systems. That is, tutors are usually older students, paraprofessionals, or adult volunteers who have moderately high domain knowledge and minimal training on the tutoring process (Cohen, Kulik, & Kulik 1982; Fitz-Gibbon 1977). Each tutor was paid $500 for serving as a tutor in 18 sessions.

LEARNING MATERIALS

The course instructor selected six topics that are normally troublesome for students in the course. Each topic had related subtopics that were to be covered in the tutoring session. An index card that listed the main topic and the subtopics was prepared for each of the six topics. The topics and subtopics are 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, dependent variables, statistics, main effects, cells, interactions; and

Interactions: independent variables, main effects, types of interactions, statistical significance.

The students were exposed to the material on two occasions prior to their participation in the tutoring sessions. First, each topic was covered in a lecture by the instructor before that topic was covered in the tutoring session. Second, each student was required to read specific pages in a research methods text (*Methods in Behavioral Research*, Cozby 1989) prior to the tutoring session. The students, therefore, had multiple chances to learn the material.

The tutoring sessions spanned an eight-week period. Only one topic was covered per week. The topics covered during the first three weeks were variables, graphs, and statistics, respectively. A two-week break followed the first three weeks of tutoring. The remaining three topics were covered during the subsequent three weeks.

EQUIPMENT AND SETTING

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

PROCEDURE

Prior to the tutoring session, the students were told that they would receive tutoring on particular pages in the Cozby text. When a student entered the tutoring room, the student was instructed to sit in view of the camera and to read the index topic card aloud. The tutoring session then proceeded in the direction that the tutor and student saw fit. The three tutors were not given a specific format to follow. They were instructed, however, to avoid simply lecturing to the student. Each tutor was also instructed to make the tutoring session last at least 45 minutes and not to exceed 60 minutes.

Each of the 27 students participated in two tutoring sessions. A counterbalancing scheme was designed so that (a) a student never had the same tutor twice, (b) each tutor covered all six topics, (c) each tutor was assigned to 18 tutoring sessions, and (d) a student was tutored once during the first three weeks and once during the second three weeks. Therefore, each tutor tutored three students on each of the six topics, which yielded 54 tutoring sessions. Ten of the 54 sessions could not be transcribed due to audio problems.

TRANSCRIPTION AND CODING OF THE TUTORING SESSIONS

Transcribers received a one-hour training session on how to transcribe the protocols. They were instructed to transcribe the entire tutoring sessions verbatim, including all "ums," "ahs," word fragments, broken sentences, and pauses. The transcribers specified whether an utterance was made by the student or tutor. In addition, transcribers noted messages that appeared on the marker board, hand gestures, head nods, and simultaneous speech acts that occurred between the student and tutor. Each written transcription was verified for accuracy by a research assistant who spot sampled random segments of the videotape.

Four trained judges coded the questions and answers in the transcripts according to the schemes that were described earlier. These judgments were made in the context of the tutorial dialogue rather than speech acts in isolation. The judges were graduate and undergraduate research assistants who were funded by a grant from the Office of Naval Research.

QUESTION IDENTIFICATION

Two judges were trained to determine whether or not a particular speech act in a transcript was a question, as specified in the Introduction. The judges had a high degree of reliability in making these judgments. The proportion of common responses be the two judges was .95.

QUESTION CONTENT CATEGORY

Two judges were trained to classify the question content categories listed in Table 1. Each question was assigned to one of the 18 content categories in Table 1. For questions that could be assigned to more than one category, the judges followed the set of priority rules that were discussed earlier. Before rendering judgments on the questions used in this study, the judges were trained on a set of practice protocols from a different tutoring sample. The judges were retrained on the practice protocols until their proportion of common responses was .70 on two consecutive protocols. After training, each judge scored approximately one half of the questions from the 44 tutoring sessions. In order to ensure that the judges remained accurate in their judgments, we randomly spot sampled four protocols in which both judges provided judgments. The proportion of common responses on these four protocols was .81.

QUESTION GENERATION MECHANISMS

Two different judges were trained to assign questions to the four question generation mechanisms presented in the Introduction. These judges were also trained on a set of practice protocols until their proportion of common responses was .70. Each question was assigned to only one question generation mechanism category. In the case of hybrid questions, the knowledge deficit category had higher priority than the other three categories because we were particularly interested in the extent to which students take an active role in self-regulating their knowledge. Common ground questions had higher priority than the social coordination and conversational control categories. The question generation mechanisms were spot sampled in the same manner as the question content categories. The proportion of common responses on the four randomly selected protocols was .78.

STUDENTS' ANSWERS

The co-authors of this article made the qualitative judgments regarding the quality of students' answers. As specified in the Introduction, answer quality had four levels: (1) complete, (2) partial, (3) vague or no answer, and (4) error-ridden. Independent judges were not trained to make these judgments because these judgments require extensive domain knowledge and familiarity with discourse analysis. For example, a judge would have to be knowledgeable in the

area of factorial designs in order to detect a student's misconceptions about main effects. During these coding sessions, two judges worked together to assign a qualitative value to each student answer. To assess reliability, the collaborative judgments made by two judges were compared to judgments made by a single judge (also one of the co-authors) on six randomly selected protocols. The proportion of common responses was .89 for these six protocols.

Student answers are typically embedded in a collaborative exchange that occurs between the tutor and the student (Graesser 1993). In this study, a typical collaborative exchange consisted of a tutor question followed by a series of conversational turns. Each student turn in a collaborative exchange was assigned to a qualitative answer category. This particular analysis included the qualitative judgments assigned to student turns that contained information relevant to the tutoring topic, as opposed to extraneous topics.

RESULTS AND DISCUSSION

The results are divided into two sections. The first section reports descriptive statistics on the various measures of achievement, questions, and answers. The second section reports the correlations between the achievement measures of students in the Research Methods course and the five measures of student questions and answers: (1) the overall number of questions asked by the student, (2) the proportion of student questions that addressed knowledge deficits, (3) the proportion of student questions that involved deep-reasoning, (4) the quality of the students' answers to questions, and (5) the proportion of positive versus negative student answers to comprehension-gauging questions posed by the tutor (e.g., "Do you understand?").

DESCRIPTIVE STATISTICS ON QUESTIONS AND ANSWERS

Number of Questions. The mean number of student questions per tutoring session was 21.1 (SD = 13.0). This rate is higher than the number of questions typically asked by students in a classroom setting (.2 questions per student per hour, Graesser & Person in press). The finding that student questions are more prevalent in one-to-one tutoring than in classrooms should be moderately encouraging to educators who seek educational environments that promote active learning. The increase in student question asking may explain why learning is better in tutoring environments than in classrooms (Bloom 1984). Alternatively, it may be important to consider the quality of the student questions rather than simply the frequency (Fishbein, Eckart, Lauver, Leeuwen, & Langmeyer 1990; Flammer 1981, Graesser, Person & Huber 1993).

Question Content Categories. Each question was assigned to one of the 18 question content categories listed in Table 1. The proportion score for the content categories indicated that verification questions were clearly the most prevalent type asked by students (.28, $SD = .20$). The next most prevalent question category for students was concept completion questions (.14, $SD = .11$). These categories were followed by instrumental/procedural questions (.13, $SD = .09$), and interpretational questions (.09, $SD = .10$). The proportion score for the six deep-reasoning question categories combined (i.e., antecedent, consequence, goal orientation, enablement, instrumental/procedural, and expectational) was .22 ($SD = .14$).

Although the distribution of the question content categories indicated that students prefer short-answer questions (.52, $SD = .18$) to long-answer questions (.40, $SD = .17$), this difference is not robust. The comparatively high proportion of long-answer questions suggests that students were attempting to place the burden on the tutor to articulate the difficult material. The assertion and request/directive categories (.08, $SD = .07$) were omitted from this comparison because they are ambiguous with respect to answer length.

Question Generation Mechanisms. Knowledge-deficit questions comprised .33 of the student questions ($SD = .11$). This suggests that students are to some extent taking an active role in self-regulating their knowledge. Most of these knowledge-deficit questions (79%) were triggered by contradictions or anomalous events. In the following example, a student had identified a contradiction in her knowledge base. The student is viewing a graph in which the lines representing each variable do not intersect. The analysis indicates that the interaction is statistically significant.

> Student: But I thought the lines had to intersect for the interaction to be significant?

Another 17% of the knowledge deficit questions were triggered by gaps in the students' knowledge. The following exchange illustrates a gap in a student's knowledge base.

> Tutor: Is there a main effect for gender?
> Student: What is a main effect?

The majority of student questions in the tutoring sessions were common ground questions (.89, $SD = .31$). These common ground questions posed by the students were attempts to confirm the validity of their own knowledge (e.g., "Intersecting lines means there's an interaction, right?"). Only 3% of the students' common ground questions were assessments of the tutor's knowledge. As expected, the efforts of both the student and tutor were focused on the knowledge base of the student.

Quality of Student Answers. As mentioned earlier, each student answer was classified as either complete, partial, vague, no answer, or error-ridden. Complete answers were the most prevalent in the collaborative exchanges, followed by partial answers, then vague/no answers, and error-ridden answers, .40 (SD = .18), .33 (SD = .13), .14 (SD = .12), and .14 (SD = .10), respectively. The proportion of complete answers is probably due to the large number of short-answer questions, which are comparatively easy to answer. We found that 60% of the tutor questions were short-answer questions. The tutors minimized the likelihood that the students would provide partial, vague, or error-ridden answers by asking comparatively few questions that require long-winded answers.

Answers to Comprehension-gauging Questions. Tutors frequently asked comprehension-gauging questions to gauge whether a student was understanding a particular topic or subtopic. In these tutoring sessions, such questions comprised .35 (SD = .33) of all tutor questions. As mentioned earlier, we were particularly interested in the answers that students give to these comprehension-gauging questions. Most of the student answers (.57, SD = .15) were positive in nature, whereas only a small percentage were negative (.07, SD = .05); the other .36 of the responses involved instances where the student did not answer the tutor's comprehension-gauging question. One explanation of the comparatively high proportion of positive answers addresses the politeness strategies of natural conversation: The students claimed they understood because they did not want to appear impolite. Another possibility is that the students actually thought they understood the material.

Measures of Achievement. Achievement was measured by examination scores and the final grade received in the research methods course. The examination scores were based on three objective examinations that consisted of 150 four-alternative, forced-choice items. Most of the exam items (67%) were selected from the Instructor's Manual associated with Cozby's text; the other items were prepared by the course instructor and graduate teaching assistants. Ten percent of the exam items contained information that was covered in the tutoring sessions. All of the exam items were prepared prior to the onset of the lectures or the tutoring sessions. It should be noted that the tutors were not informed about the content of the examination items. The mean examination score was 100.6 (SD = 11.4). Students correctly answered 67% of the exam items (100.6/150 = .67) indicating neither a ceiling nor a floor effect. Regarding the final grade received in the course, four students received an "A," nine received a "B," ten received a "C," and four received a "C−" or a "D."

CORRELATIONS BETWEEN ACHIEVEMENT MEASURES AND MEASURES OF STUDENT QUESTIONS AND ANSWERS

This section reports the correlations between the student achievement measures and the five measures of student questions and answers. The correla-

tions between these measures and the achievement measures are provided in Table 2.

Overall Number of Student Questions. One would expect that good students would take a more active role in regulating their knowledge, and would therefore ask more questions. Previous researchers, however, have reported a zero or negative correlation between the incidence of student questions in tutoring and the achievement level of the students (Fishbein et al. 1990; Flammer 1981). Our results are compatible with these previous findings. The correlation between the final grade received in the course and the overall number of student questions was significantly negative. The correlation between examination scores and the overall questions was also negative, but not statistically significant. Therefore, good students do not necessarily ask more questions. The low correlation between the examination scores and the overall number of questions might be explained by a U-shaped function. That is, poor students and good students might ask more questions than students with an intermediate level of achievement. Our analysis did not uncover a significant curvilinear relationship, however.

Student Knowledge-deficit Questions. There are several hypotheses regarding students' use of knowledge deficit questions. One hypothesis predicts that poor students have more gaps in their knowledge, and should therefore ask more questions to rectify these gaps. A second hypothesis predicts that the good students should ask more knowledge-deficit questions to the extent they are

TABLE 2
Correlations between Student Achievement and Properties of Student Questions and Answers

	Achievement Measure	
Measures of Student Questions and Answers	Examination Scores	Final Grade
Total Number of Student Questions	−.22	−.34**
Proportion of Student Questions That Are Knowledge Deficit Questions	.15	.32
Proportion of Student Questions That Are Deep-Reasoning Questions	.44*	.58*
Proportion of Students' Answer Contributions That Are:		
Completely Correct	.32	.43*
Partially Correct	.09	−.09
Vague or No Answer	−.30	−.46*
Error-Ridden	−.32	−.10
Error-Ridden, Vague, or No Answer	−.52*	−.49*
Proportion of **Yes** Answers (by student) to Comprehensive-Gauging Questions (by tutor)	.07	.05
Proportion of **No** Answers (by student) to Comprehension-Gauging Questions (by tutor)	.42*	.20

*$p < .05$, two tailed.
**$p < .10$, two-tailed.

taking an active role in self-regulating their knowledge. A third hypothesis would accommodate both of these possibilities and thereby yield a U-shaped function. A fourth hypothesis predicts that the average students (B and C students) will ask more knowledge deficit questions because it is these students who have failed to master the material and are to some extent sensitive to what they do not know. If the fourth hypothesis is the case, the analysis should yield an inverted U-shaped function between student achievement and the proportion of knowledge deficit questions posed by the student. None of these hypotheses were confirmed in our analysis, however. There was no evidence that student achievement was correlated with the proportion of knowledge-deficit questions asked by the students. A curvilinear relationship also proved to be non-significant. Hence, the knowledge level of the student cannot be validly inferred from the incidence of knowledge-deficit questions asked by the student.

One plausible explanation for this finding is that students frequently fail to identify their own knowledge deficits. As mentioned earlier, it is well documented that students have difficulty identifying inconsistencies and contradictions in various text forms. Graesser and McMahen (1993) reported modest question-generation rates when adults were instructed to generate questions about mathematical word problems that contained contradictions. The questioning rates for identifying contradictions in the word problems were in fact comparable to the rates of the questions generated for word problems with irrelevant information. Several researchers have suggested that students may discount or rationalize away the contradictions they encounter in various text forms (Baker 1979; Epstein et al. 1984; Glenberg et al. 1982; Graesser et al. 1993; Markman 1979; Otero & Campanario 1990). Perhaps this rationalization takes place when a student encounters a contradiction or anomaly in his or her knowledge base during one-to-one tutoring.

Deep-reasoning Questions Asked by the Student. The six deep-reasoning questions are denoted by asterisks in Table 1. As discussed earlier, these deep-reasoning categories are defined as questions that tap the steps and rationale in logical reasoning, in problem solving procedures, in plans, and in causal chains (e.g., *why, how, what-if* questions). As mentioned earlier, each question had also been classified according to the cognitive demand it places on the questioner using Bloom's taxonomy of educational objectives in the cognitive domain (Bloom 1956).

Regarding student understanding, there was a positive correlation between the proportion of deep-reasoning questions asked by the student and both measures of achievement. This result would lead one to suspect that tutors could infer student understanding from the depth of their questions. However, this result should be tempered by the fact that a small proportion of the questions involved deep-reasoning. That is, only 22% of the student questions were deep-reasoning questions; only 8% of these questions were classified as cognitively demanding on Bloom's taxonomy (levels 2-6).

The Quality of the Students' Answers. The tutors in these tutoring sessions frequently asked questions that assessed a student's knowledge about a particular topic (40%). We measured the quality of the students' answers to these tutor-posed questions. As mentioned earlier, each student contribution (i.e., student assertions during the collaborative exchange) was assigned to one of four qualitative categories: (1) complete answer, (2) partial answer, (3) vague answer or no answer, or (4) error-ridden answer. The good students should provide more complete and partially correct answers, whereas the poor students should provide more vague and error-ridden answers. This expectation was confirmed by the correlations that appear in Table 2. The proportion of complete answers given by students was positively correlated with the students' final grades. That is, the good students were more likely to provide complete answers to tutor questions than poor students. Student achievement was negatively correlated with the proportion of student answers in the vague/no answer and error-ridden categories. Thus, the poor students were more likely to provide vague answers, error-ridden answers, or no answer at all to tutor-posed questions. These robust correlations between student achievement and student answers suggest that student understanding is best assessed through analyses of students' answers rather than analyses of student questions.

Students' Answers to Tutors' Comprehension Gauging Questions. Comprehension-gauging questions were defined by Graesser, Person, and Huber (1992) as questions that allow the tutor to gauge the student's degree of comprehension. Such questions also allow students to assess their own knowledge about the material. In tutoring sessions, comprehension-gauging questions typically occur after a particular topic has been discussed. For example, after a dialogue about operational definitions, the tutor might ask the student "Do you understand?" or "Okay?" We found a positive correlation between the students' examination scores and the proportion of answers to comprehension-gauging questions that were "no" (i.e., "I don't understand."). That is, the good students were more likely to say they did not understand. This counterintuitive finding is compatible with the findings reported by Chi et al. (1989).

Unlike Chi and her colleagues, we examined the positive responses that students gave to comprehension-gauging questions. The linear relationship between student achievement and the proportion of positive responses was not statistically significant. However, we found a marginally significant curvilinear relationship between student examination scores and the proportion of "yes" answers to comprehension-gauging questions, $F(2,30) = 2.16$, $p < .10$. "Yes" answers to comprehension-gauging questions were typically given by the average students who made B's and C's on the examinations, .62 and .61, respectively; the proportion scores were lower for the C$-$/D students (.52) and the A students (.46). This curvilinear relationship suggests that the good students are more sensitive to what they do know versus do not know (Brown, Bransford, Ferrara, & Campione 1983). This finding is also compatible with Miyake and

Norman's (1979) claim that a certain amount of material has to be mastered before a learner can accurately assess his or her own knowledge base. There are several reasons why the poor students gave fewer positive answers. Perhaps the information is unclear and becomes too overwhelming for the poor student, so the poor student is reluctant to offer positive answers. Given that the curvilinear relationship was only marginally significant, additional research needs to be conducted to verify this serendipitous finding.

SUMMARY AND CONCLUSIONS

The primary goal of this research was to determine the relationship between measures of student achievement in a Research Methods course and measures of student questions and answers that occur in one-to-one tutoring sessions. Our results supported several interesting conclusions which are summarized below.

Conclusion 1. Good students do not necessarily ask more questions. Our results showed a small negative correlation between the overall number of questions asked by students and the final grades they received in the course. Hence the degree to which students take an active role in self-regulating their own knowledge is not robustly reflected in the overall number of questions they ask.

Conclusion 2. Student knowledge-deficit questions are a poor information source for gauging student understanding. Our results indicated that the proportion of knowledge-deficit questions asked by students is unrelated to student achievement. One hypothesis predicted that the poor students would ask more of these questions because they have more knowledge deficits. One reason why the poor students do not ask more knowledge-deficit questions is possibly that they are unable to calibrate their own understanding. A second reason involves the social barriers to asking questions. Even if the student does detect a knowledge deficit, he or she may be reluctant to rectify it in the social context (Graesser & McMahen 1993; van der Meij 1987, 1988).

An alternative hypothesis predicted that good students would ask more knowledge-deficit questions to the extent that they are taking an active role in self-regulating their knowledge. There are several possible reasons why good students did not ask a higher proportion of knowledge-deficit questions. One reason is that they understood the material, and therefore did not need to ask questions. Another reason is that good students may use other strategies to rectify their knowledge deficits. Finally, the social barriers that exist for poor students may also exist for good students.

Conclusion 3. Student deep-reasoning questions are a moderately reliable source for gauging student understanding. Our findings indicated a positive correlation between the proportion of deep-reasoning questions asked by students and

measures of student achievement. The relatively high proportion of deep-reasoning questions asked by good students suggests that, to some degree, good students acquire knowledge at deeper and more sophisticated levels than poor students. However, this finding is only moderately encouraging. Our analysis of student questions revealed that a comparatively small percentage of the questions were sophisticated, deep-reasoning questions. The small percentage would make it difficult for the tutor to assess the sophistication of the students' knowledge base.

Conclusion 4. Students' answers to comprehension-gauging questions are misleading. Our analysis revealed a positive correlation between student achievement and students' negative answers to comprehension-gauging questions (e.g., "No, I do not understand."). This somewhat counterintuitive result is compatible with the findings of Chi et al. (1989). This finding suggests that good students realize that they do not understand more often than do the poor students. Hence, tutors should be suspect of students who rarely say they do not understand.

We also found a marginally significant curvilinear relationship between student achievement and students' positive answers to comprehension-gauging questions (e.g., "Yes, I understand."). That is, the poor students and the good students say "Yes" less often than the students at an intermediate level. This finding is compatible with Miyake and Norman's (1979) claim that learners have to master a certain amount of the material before they can accurately assess their own knowledge bases. Good students had mastered a sufficient amount of the material and were therefore sensitive to what they did versus did not know. To some extent, the good students' positive answers were an accurate gauge of their understanding. The poor students' reluctance to give positive answers to comprehension-gauging questions is perhaps an indicator that they mastered very little of the material.

Conclusion 5. Students' answers are the most reliable source for inferring student understanding. We expected that good students would be more likely to provide complete and partially complete answers, whereas the poor students would provide more vague and error-ridden answers. These intuitively plausible predictions were confirmed. We found a positive correlation between the student achievement measures and the proportion of complete answers given by the student. That is, good students are more likely to provide complete answers to tutor questions than poor students. We also found student achievement to be negatively correlated with the proportion of student answers that belonged to the error-ridden, vague, or no answer categories. These correlations were high, and tutor questions were frequent. Therefore, based on our correlational analyses, tutors primarily infer student understanding from the students' answers rather than from the students' questions.

In closing, we propose that the properties of conversation and dialogue are critical in explaining why one-on-one tutoring is more effective than typical

classroom instruction. This study has uncovered some intuitive, as well as counterintuitive, properties about the nature of questions and answers in conversation and the achievement levels of students. Future research will hopefully reveal additional patterns of dialogue and clarify their relationship with both achievement levels and learning outcomes.

REFERENCES

Allen, J.F. (1987). *Natural language understanding*. Menlo Park, CA: The Benjamin-Cummins Publishing Co.

Anderson, J.R., D.F. Boyle, & B.J. Reiser. (1985). "Intelligent tutoring systems." *Science, 228*, 456–462.

Austin, J.L. (1962). *How to do things with words*. Oxford: Oxford University Press.

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

———. (1985). "Differences in the standards used by college student to evaluate their comprehension of expository prose." *Reading Research Quarterly, 20*, 297–313.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives. Handbook I: Cognitive domain*. New York: McKay.

———. (1984). "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring." *Educational Researcher, 13*, 4–16.

Bransford, J.D., R. Arbitman-Smith, B.S. Stein, & N.J. Vye. (1985). "Analysis—Improving thinking and learning skills: An analysis of three approaches." Pp. 133–206 in *Thinking and learning skills* (Vol. 1), edited by S.F. Chipman, J.W. Segal, & R. Glaser. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, A.L., J.D. Bransford, R.A. Ferrara, & J.C. Campione. (1983). "Learning, remembering, and understanding." Pp. 77–166 in *Handbook of child psychology (4th ed.). cognitive development* (Vol. 3), edited by J.H. Flavell & E.M. Markman. New York: Wiley.

Brown, R.R. & J.S. Burton. (1982). "An investigation of computer coaching for informal learning activities." Pp. 79–98 in *Intelligent tutoring system*, edited by D. Sleeman & J.S. Brown. New York: Academic Press.

Brown, P. & S.C. Levinson. (1987). *Politeness: Some universals in language use*. Cambridge: Cambridge University Press.

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

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

Clancey, W.J. (1983). "Guidon." *Journal of Computer-Based Instruction, 10*(1&2), 8–15.

Clancey, W.J., J.S. Bennett, & P.R. Cohen. (1982). "Applications oriented AI research:

Education." In *The handbook of artificial intelligence* (Vol. 2), edited by A. Barr & E.A. Feigenbaum. Los Altos, CA: Kaufman.

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

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

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

Collins, A. (1985). "Teaching and reasoning skills." Pp. 579–586 in *Thinking and learning skills* (Vol. 2), edited by S.F. Chipman, J.W. Segal, & R. Glaser. Hillsdale, NJ: Erlbaum.

———. (1988). "Different goals of inquiry teaching." *Questioning Exchange, 2,* 39–45.

Cozby, P.C. (1989). *Methods in behavioral research.* Mountain View, CA: Mayfield Publishing Company.

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

Dede, C. (1986). "A review and synthesis of recent research in intelligent computer-assisted instruction." *International Journal of Man-Machine Studies, 24,* 329–353.

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

———. (1988). *Questioning and teaching: A manual of practice.* New York: Teachers College Press.

Eberts, R. & J.F. Brock. (1988). "Computer-based instruction." Pp. 599–627 in *Handbook of human-computer interaction,* edited by M. Helander. Amsterdam: Elsevier Science Publishers.

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

Fishbein, H.D., T. Eckart, E. Lauver, R.V. Leeuwen, & D. Langmeyer. (1990). "Learner's questions and comprehension in a tutoring setting." *Journal of Educational Psychology, 82,* 163–170.

Fitz-Gibbon, C.T. (1977). An analysis of the literature of cross-age tutoring. (ERIC Document Reproduction Service No. ED 148 807).

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

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

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

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

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

Graesser, A.C. (1993). "Dialogue patterns and feedback mechanisms during naturalistic tutoring." Pp. 126–130 in *Proceedings of the 15th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

———. (1993). *Questioning mechanisms during complex learning.* Memphis, TN: Memphis State University (ERIC Document Reproduction Service ED No. 350306).

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

Graesser, A.C. & C.L. McMahen. (1993). "Anomalous information triggers questions when adults solve quantitative problems and comprehend stories." *Journal of Educational Psychology, 85,* 136–151.

Graesser, A.C. & N.K. Person. (in press). "Question asking during tutoring." *American Educational Research Journal.*

Graesser, A.C., N.K. Person, & J. Huber. (1992). "Mechanisms that generate questions." Pp. 167–187 in *Questions and information systems,* edited by T. Lauer, E. Peacock, & A. Graesser. Hillsdale, NJ: Erlbaum.

———. (1993). "Question asking during tutoring and in the design of educational software." Pp. 149–172 in *Cognitive science foundations of instruction,* edited by M. Rabinowitz. Hillsdale, NJ: Lawrence Erlbaum Associates.

Grice, H. (1975). "Logic and conversation." In *Syntax and semantics* (Vol. 3), edited by P. Cole & J. Morgan. New York: Academic.

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

Kass, A. (1992). "Question asking, artificial intelligence, and human creativity." Pp. 303–360 in *Questions and information systems,* edited by T. Lauer, E. Peacock, & A.C. Graesser. Hillsdale, NJ: Erlbaum.

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

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

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

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

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

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

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

McDevitt, T.M. & M. Carroll. (1988). "Are you trying to trick me? Some social influences on children's responses to problematic messages." *Merrill-Palmer Quarterly, 34,* 131–145.

Miyake, N. & D.A. Norman. (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.

Newman, R.S. & L. Goldin. (1990). "Children's reluctance to seek help with school work." *Journal of Educational Psychology, 82,* 92–100.

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

Palincsar, A.S. & A.L. Brown. (1984). "Reciprocal teaching of comprehension-fostering and comprehension monitoring activities." *Cognition and Instruction, 1,* 117–175.

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

Paris, S.G. & R.S. Newman. (1990). "Developmental aspects of self-regulated learning." *Educational Psychologists, 25,* 87–102.

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

Pressley, M., E.S. Ghatala, V. Woloshyn, & J. Pirie. (1990). "Sometimes adults miss the main ideas and do not realize it: Confidence in responses to short-answer and multiple choice comprehension questions." *Reading Research Quarterly, 25,* 232–249.

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

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

Putnam, R.T. (1987). "Structuring and adjusting content for students: A live and simulated tutoring of addition." *American Educational Research Journal, 24,* 13–48.

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

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

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

Sleeman, D.H. & J.S. Brown. (Eds.) (1982). *Intelligent tutoring systems.* New York: Academic Press.

Souther, A., L. Acker, J. Lester, & B. Porter. (1989). "Using view types to generate explanations in intelligent tutoring systems." Pp. 123–130 in *Proceedings of the 11th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stokal, R.R. (1974). "Classification." *Science, 185,* 115–123.

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

———. (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 Press.

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

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