What is a Good Question?

Art Graesser[1], Yasuhiro Ozuru[2], and Jeremiah Sullins[1]

1 The University of Memphis

2 Educational Testing Service

Send correspondence to:

Art Graesser
Psychology Department
202 Psychology Building
University of Memphis
Memphis, TN, 38152-3230
901-678-2742
901-678-2579 (fax)
a-graesser@memphis.edu

What is a good question?

Isabel Beck's contributions to education have spanned a broad landscape, with monumental advances in vocabulary learning, deep comprehension, discourse coherence, communication, classroom interaction, and questions. Questions are the focus of the present chapter, in recognition of the book published by Isabel and her colleagues, *Questioning the Author* (Beck, McKeown, Hamilton, & Kucan, 1997; McKeown, Beck, Hamilton, & Kucan, 1999). This work is both a scholarly advance and a practical solution to improving reading comprehension. Students learn to read text at deeper levels of comprehension by having teachers model and prompt the readers to ask good questions, such as those listed below.

What is the author trying to tell you?

Why is the author telling you that?

Does the author say it clearly?

How could the author have said things more clearly?

What would you say instead?

These are indeed excellent questions that encourage students to reflect on the process of communication through text, on the quality of the text in meeting communication goals, on text coherence, and on the possibility of a text having alternative forms of expression. Instead of viewing text as a perfect artifact that has been cast in stone, text can be viewed as a fallible and flexible medium of communication that merits critical scrutiny. The process of inquiry helps the readers shift their mindset from viewing text as a static object to text being part of a fluid communication process. This shift in the reader's mental model of the reading process results in deeper comprehension.

Questions have an important status in the work of Beck, many of her contemporaries, and our research as well. We believe that questions vary in quality with respect to their supporting learning. The first author of this chapter has been obsessed with answering a very simple research question: *What makes a good question?* An answer to this research question has the potential to profoundly improve the processes of classroom learning, tutoring, reading, exploration, hypothesis testing, motivation and a host of other activities in the educational enterprise (Graesser & McMahen, 1993; Graesser, McNamara, & VanLehn, 2005; Graesser & Olde, 2003; Graesser & Person, 1994; Wisher & Graesser, 2007).

This chapter begins by examining some alternative frameworks for categorizing questions and scaling them on quality. The subsequent section reviews some research that supports the claim that improved questions can promote learning at deeper levels. Next we describe some learning technologies that encourage the asking and answering of high quality questions. The chapter ends with a plea for someone to develop a Question Authoring Workbench (QAW). The Workbench would train instructors, students, textbook writers, curriculum developers, test constructors and other educational communities to compose better questions and thereby elevate the standards of comprehension beyond the current shallow standards.

The Landscape of Questions

One initial step in improving question quality is to consider the broad landscape of question categories, knowledge representations, and cognitive processes. We have defined a *landscape of questions* as the distribution of questions that tap different categories of knowledge and different cognitive proficiencies. If there are Q question categories, K categories of knowledge, and P cognitive processes, then there are Q x K x P *cells* in the total space of

questions, as illustrated in Figure 1. The broad landscape needs to be considered when a question writer generates questions for particular tasks and goals (AERA, APA, & NCME, 1999). Most question writers focus on a very narrow terrain whereas it would be beneficial to consider a broader vision, with greater analytical detail and relevance to the learning objectives.

INSERT FIGURE 1 AND TABLE 1 ABOUT HERE

Types of Questions

Schemes for classifying questions have been proposed in the fields of psychology, education, and artificial intelligence (Dillon, 1988; Graesser & Person, 1994; Lehnert, 1978; Mosenthal, 1996). Two of the promising schemes were developed by Graesser and Person (1994) and Mosenthal (1996).

Graesser–Person taxonomy. The Graesser–Person taxonomy (Graesser & Person, 1994) classifies questions according to the nature of the information being sought in a good answer to the question. Table 1 lists and defines these categories. The 16 question categories can be scaled on depth, which is defined by the amount and complexity of content produced in a good answer to the question. In some of our analyses, we have differentiated simple *shallow* questions (categories 1-4), *intermediate* questions (5-8), versus complex *deep* questions (9-16). This scale of depth is validated to the extent that it correlates significantly ($r = .60 \pm .05$) with both Mosenthal's (1996) scale of question depth and the original Bloom's taxonomy of cognitive difficulty (1956).

Although the Graesser-Person scheme has some degree of validity, the claim could be challenged that it has a perfect scale for depth. For example, one can readily identify *disjunctive* questions that require considerable thought and reasoning, as in the case of the difficult physics question: *When the passenger is rear-ended, does the head initially (a) go forward, (b) go*

backwards, or (c) stay the same? The reasoning behind generating an answer to this question requires a causal analysis, corresponding to question categories 10 and 11, so this question may functionally be a hybrid question. The existence of hybrid questions is technically incompatible with a unidimensional scale of questions on depth. As another example, an *instrumental/procedural* question is classified as complex deep, but some of these questions require minimal thought and reasoning, such as: *How do you open a refrigerator?* The Graesser-Person scale of depth of question categories is somewhat crude and approximate because it is also important to consider the knowledge representations and cognitive processes that are recruited during the course of question answering.

Mosenthal's taxonomy. Mosenthal (1996) developed a coding system to scale questions on abstractness, which has a reasonable correspondence with depth. This classification scheme also is based on the information sought in the answer and no doubt interacts with knowledge and cognitive processes in systematic ways. The five levels of abstractness are presented below.

(1) **Most concrete**. Identification of concrete entity (e.g., specific person, thing, action) based on explicit information (e.g., in a text, picture, or other information source)

(2) **Highly concrete**. Identification of objectively observable attributes (action, types, attributes, amounts) based on explicit information.

(3) **Intermediate**. Identification of manner, procedures, and goals that cannot be directly observed in explicit information.

(4) **Highly abstract**. Identification of cause, effect, reason, and evidence derived from explicit information.

(5) **Most abstract**. Identification of theories, equivalences, differences, and themes beyond the explicit information.

This classification schemes conflates a number of dimensions that are conceivably separable, such as depth, complexity, abstractness, and explicitness. However, it no doubt has some degree of validity in scaling questions on quality with respect to promoting learning.

Types of Knowledge

Knowledge representations in artificial intelligence and cognitive science. Researchers in cognitive science and artificial intelligence in the 1980's and 1990's spent considerable effort dissecting the formal and psychological properties of different classes of knowledge (Lehmann, 1992; Lenat, 1995; Schank, 1999; Sowa, 1983). These theories (a) identified particular elements, relations, and packages of knowledge and (b) specified the formal properties and constraints of each class of knowledge. The theoretical goal was to systematically dissect the components, relations, and constraints of each type of knowledge rather than to rely on intuitions and folklore. The question categories (e.g., in Table 1) operate systematically on particular types of knowledge in various computational models of question answering such as QUALM (Lehnert, 1978) and QUEST (Graesser, Gordon, & Brainerd, 1992).

As an illustration of some different types of knowledge, consider the categories proposed in Wisher and Graesser (2007).

Agents and entities: Organized sets of people, organizations, countries, and entities.

Class inclusion: One concept is a subtype or subclass of another concept.

Spatial layout: Spatial relations among regions and entities in regions.

Compositional structures: Components have subparts and subcomponents.

Procedures & plans: A sequence of steps/actions in a procedure accomplishes a goal.

Causal chains & networks: An event is caused by a sequence of events and enabling states.

Others: Property descriptions, quantitative specifications, rules, mental states of agents.

Each of these types of knowledge has a unique set of properties, relations, and constraints. For example an IS-A relation connects concept nodes in class inclusion knowledge, e.g., *a robin is a bird, a bird is an animal,* whereas a CAUSE relation would connect event nodes in a causal network. Question categories of the sort in Table 1 are systematically aligned with the types of knowledge illustrated here. Definition questions have a close affinity to class inclusion structures whereas goal-orientation questions have a close affinity to procedures and plans. The QUEST model of question answering (Graesser et al., 1992) provided a systematic mapping between the types of knowledge and many of the question classes in Table 1.

Levels of representation in discourse processing. Researchers in the field of discourse processing postulate that cognitive representations of texts can be separated into levels of explicit information, referential mental models (sometimes called situation models), rhetorical structure, and pragmatic communication (Graesser, Millis, & Zwaan, 1997; Kintsch, 1998; Perfetti, Britt, & Georgi, 1995; Snow, 2002). The *explicit information* preserves the wording, syntax, and semantic content of the material that is directly presented. The *mental model* is the referential content of what the explicit material is about. In a technical text that explains a device, for example, the mental model would include: the components of the device, the spatial arrangement of components, the causal chain of events when the system successfully unfolds, the mechanisms that explain each causal step, the functions of components, and the plans of humans who manipulate the system for various purposes. The *rhetorical structure* is the more global composition and genre that organizes the discourse. For example, the structure of a story is very

different from an expository text with a claim+evidence rhetorical structure. The *pragmatic communication* specifies the main messages or points that the author is trying to convey to the reader. These four levels of discourse can be ordered on depth. More inferences and deeper levels of processing are needed as one moves from the explicit information to the mental models to the rhetorical and pragmatic communication levels. Bloom's (1956) categories of recognition and recall correspond with explicit information, whereas the category of comprehension closely aligns with the mental models that result from the integration of explicit information with pre-existing knowledge. The questions in Beck et al.'s (1997) *Questioning the Author* have important alignments to rhetorical structure and pragmatic communication.

Types of Cognitive Processes

Cognitive processes need to operate on the knowledge in order for the knowledge to have an impact on a person's question-answering behavior. It is therefore important to identify the types of cognitive processes during question answering and how different types of knowledge are recruited in these processes (Goldman, Duschl, Ellenbogen, Williams, & Tzou, 2003; Graesser, Lang, & Roberts, 1991; Guthrie, 1988; Kyllonen & Roberts, 2003; Reder, 1987; Rouet, 2006; Singer, 2003). It is beyond the scope of this section to cover the rich cognitive literature on process models of question asking and answering. Instead, we will briefly identify two older models that are widely recognized in the field of education.

Bloom's taxonomy. One of the early analyses of cognitive processes was Bloom's taxonomy (1956). The major categories in the original system are presented below:

(1) **Recognition**. The process of verbatim identification of specific content (e.g., terms, facts, rules, methods, principles, procedures, objects) that was explicitly mentioned in the learning material.

(2) **Recall**. The process of actively retrieving from memory and producing content that was explicitly mentioned in the learning material.

(3) **Comprehension**. Demonstrating understanding of the learning material at the mental model level by generating inferences, interpreting, paraphrasing, translating, explaining, or summarizing information.

(4) **Application**. The process of applying knowledge extracted from the learning material to a problem, situation, or case (fictitious or real-world) that was not explicitly expressed in the learning material.

(5) **Analysis**. The process of decomposing elements and linking relationships between elements.

(6) **Synthesis**. The process of assembling new patterns and structures, such as constructing a novel solution to a problem or composing a novel message to an audience.

(7) **Evaluation**. The process of judging the value or effectiveness of a process, procedure, or entity, according to some criteria and standards.

The cognitive processes tend to be progressively more difficult the higher the number. Recognition and recall are the easiest, comprehension is intermediate, and classes 4-7 are the most difficult. However, the relative ordering within categories 4-7 are not necessarily clearcut.

Bloom's taxonomy may not provide a perfect classification of cognitive processes, but the scheme has survived in many educational circles for several decades. Recognition and recall would be the primary processes associated with questions that access facts and events stored in long-term memory, such as *What is the value of gravity on the earth's surface?* or *When did Lance Armstrong first win the Tour de France?*. However, comprehension and synthesis would

be needed when a question inquires about a character's hidden intentions in a novel, and application is needed when a person inquires about possible causes of equipment breakdown.

Carroll's coding scheme for cognitive tasks. Carroll's (1976, 1987) coding scheme for items from cognitive tasks may be applied to the coding of questions on the process dimension. Carroll 's coding scheme included six general elements: (a) characteristics of the stimuli in the task, (b) the types of overt responses, (c) the temporal parameters of steps in the task, (d) the elementary information processes that are likely to be executed, (e) speed-related influences on task performance, and (f) the primary memory stores (e.g., short-term versus long-term memory stores) during task completion and the type of item content in these stores.

The elementary information processes are segregated into necessary *operations* versus more probabilistic *strategies.* The operations and strategies that are most relevant to typical questions that tap deep knowledge include: (1) Educing identities or similarities between two or more stimuli, (2) retrieving general information from memory, (3) retrieving or constructing hypotheses, (3) examining different portions of memory, (4) performing serial operations with data from memory, (5) recording intermediate results, (6) reinterpreting ambiguous items, (7) comprehending and analyzing language stimuli, and (8) judging stimuli with respect to specified characteristics. Carroll's (1976, 1987) coding scheme may be updated to incorporate recent advances in cognitive psychology and cognitive science (Floyd, 2005; Sternberg & Perez, 2005).

The landscape of questions discussed in this section has included example theoretical schemes for classifying questions, knowledge representations, and cognitive processes. The categories can be weakly ordered on depth so there is some principled foundation for evaluated the quality of questions in promoting learning. We present these schemes for purposes of

illustration rather than asserting they are ideal theoretical schemes. Future research will no doubt provide improvements in the classifications and theoretical analyses.

A Peek at the Quality of Multiple Choice Questions of Textbook Writers

It is well documented that school children and adult learners have substantial difficulties generating deep questions (Chi et al., 2004; Graesser & Person, 1994; Wisher & Graesser, 2007), as will be discussed later in this chapter. It is similarly difficult for experts to generate deep questions. One dramatic example of this is in a recent study we conducted on a corpus of multiple choice questions on psychology in college textbooks and a Graduate Record Examination practice book (written by a commercial publisher, not Educational Testing Service). We randomly selected 30 MC questions from the test banks associated with 3 textbooks and the GRE practice book, yielding 120 questions altogether. Cognitive psychologists and graduate students in cognitive psychology coded the questions on two different theoretical schemes on question depth. One was the question taxonomy of Graesser and Person (1994), which has a depth classification that significantly correlates with Bloom's (1956) taxonomy of cognitive objectives (r = .64). The other was Mosenthal's (1996) scale of question depth, which correlates well (r = .59) with the Graesser–Person taxonomy. The analyses revealed that only 23% of the questions were classified as deep questions according to the Graesser–Person taxonomy, and 21% were classified as deep questions according to the Mosenthal scale. Quite clearly, the textbook industry, teachers, and other experts need assistance in generating deep questions because it is hardly a natural inclination to generate them. This result speaks to the importance of having our QKP landscape of questions guide the generation of questions in the educational enterprise. The distribution of questions needs to be shifted to greater depth and relevance to the learning objectives of our school systems.

Questions in Comprehension and Learning

There is an idealistic vision that students are curious question generators who actively self-regulate their learning. They identify their own knowledge deficits, ask questions that focus on these deficits, and answer the questions by exploring reliable information sources. Unfortunately, this idealistic vision of intelligent inquiry is an illusion at this point in educational practice. Most learners have trouble identifying their own knowledge deficits (Baker, 1985; Hacker, Dunlosky, & Graessser, 1998) and ask very few questions (Dillon, 1988; Good, Slavings, Harel, Emerson, 1987; Graesser & Person, 1994). Graesser and Person's (1994) estimate from available studies revealed that the typical student asks less than .2 questions per hour in a classroom and that the poverty of classroom questions is a general phenomenon across cultures. The fact that it takes several hours for a typical student to ask one question in a classroom is perhaps not surprising because it would be impossible for a teacher to accommodate 25-30 curious students. The rate of question asking is higher in other learning environments (Graesser, McNamara, & VanLehn, 2005). For example, an average student asks 26 questions per hour in one-on-one human tutoring sessions (Graesser & Person, 1994) and 120 questions per hour in a learning environment that forces students to ask questions in order to access any and all information (Graesser, Langston, & Baggett, 1993).

Given the poverty of student questions, particularly questions at deeper levels, researchers in cognitive science and education have often advocated learning environments that encourage students to generate questions (Beck et al., 1997; Collins, 1988; Edelson, Gordin, & Pea, 1999; Palincsar & Brown, 1984; Pressley & Forrest-Pressley, 1985; Schank, 1999; van der Meij, 1994; Zimmerman, 1989). There are several reasons why question generation might play a central role in learning (Wisher & Graesser, 2007):

a) **Active learning**. The learner actively constructs knowledge in the service of questions rather than passively receiving information.

b) **Metacognition**. The learner becomes sensitive to his/her own knowledge deficits and comprehension failures while the learner attempts to comprehend the material.

c) **Self-regulated learning**. The learner takes charge of both identifying and correcting comprehension problems.

d) **Motivation and engagement**. The learner is more motivated and engaged in the material because the learning experience is tailored to the learner's own needs.

e) **Building common ground with author**. The learner achieves more shared knowledge with the author of the material.

f) **Transfer appropriate processing**. Learners are normally tested by answering questions, so the learner's generating questions should improve the overlap between comprehension representations and test representations.

g) **Coding of cognitive representation**. The cognitive representation is more precise, specific, and elaborate when the learner generates questions.

Empirical evidence supports the claim that improvements in the comprehension, learning, and memory of technical material can be achieved by training students to ask questions during comprehension (Ciardiello, 1998; Davey & McBride, 1986; Gavelrek & Raphael, 1985; King, 1989, 1992, 1994; Palincsar & Brown, 1984; Rosenshine, Meister, & Chapman, 1996; van der Meij, 1994; Wong, 1985). The process of question generation accounts for a significant amount of these improvements from question generated learning, over and above the information supplied by answers. Rosenshine et al. (1996) provided the most comprehensive analysis of the impact of question generation on learning in their meta-analysis of 26 empirical studies that

compared question generation learning to conditions with appropriate controls. The outcome measures in these studies included standardized tests, short-answer or multiple-choice questions prepared by experimenters, and summaries of the texts. The median effect size was .36 for the standardized tests, .87 for the experimenter-generated tests, and .85 for the summary tests.

Training students to ask deep questions would of course be desired in the interventions. One of the key predictors of deep questions during inquiry is the existence of goals, tasks, or challenges that place them in *cognitive disequilibrium*. Learners face cognitive disequilibrium when they encounter obstacles to goals, anomalies, contradictions, incompatibilities with prior knowledge, salient contrasts, obvious gaps in knowledge, and uncertainty in the face of decisions (Chinn & Brewer, 1993; Collins, 1988; Festinger, 1957; Flammer, 1981; Graesser & McMahen, 1993; Schank, 1999). Graesser and his colleagues have developed a cognitive model of question asking called PREG (Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Graesser & Olde, 2003; Otero & Graesser, 2001) that embraces cognitive disequilibrium in its foundation. The term PREG stems from part of the word *pregunta*, which means question in Spanish. The PREG model has a set of rules that predict the particular questions that readers should ask on the basis of the characteristics of the text, the type of disequilibrium, the reader's background knowledge, and metacognitive standards of comprehension (Otero & Graesser, 2001). It is beyond the scope of this chapter to describe the PREG model in detail, however.

It is important to acknowledge that questions are not always generated by knowledge deficits and cognitive disequilibrium. Graesser, Person, and Huber (1992) identified four very different types of question generation mechanisms that occur in naturalistic settings. Whereas the first category consists of bona fide *knowledge deficit* questions, the other three mechanisms addressed communication and social interaction processes. *Common ground* questions are asked

when the questioner wants to establish or confirm whether knowledge is shared between participants in the conversation (such as "Are we working on the third problem?," "Did you mean the independent variable?"). *Social coordination* questions are indirect requests for the addressee to perform an action or for the questioner to have permission to perform an action in a collaborative activity (e.g., "Could you graph these numbers?", "Can we take a break now?"). *Conversation-control* questions are asked to manipulate the flow of conversation or the attention of the speech participants (e.g., "Can I ask you a question?"). Sometimes a student's question is ambiguous as to whether it is a knowledge deficit question or an attempt to get attention from a teacher, tutor, or peer.

Many, if not most, questions posed by students and teachers are not sincere information-seeking (SIS) questions. Van der Meij (1987) identified 11 assumptions that need to be in place in order for a question to qualify as a SIS question.

1. The questioner does not know the information he asks for with the question.
2. The question specifies the information sought after.
3. The questioner believes that the presuppositions to the question are true.
4. The questioner believes that an answer exists.
5. The questioner wants to know the answer.
6. The questioner can assess whether a reply constitutes an answer.
7. The questioner poses the question only if the benefits exceed the costs.
8. The questioner believes that the respondent knows the answer.
9. The questioner believes that the respondent will not give the answer in absence of a question.
10. The questioner believes that the respondent will supply the answer.

11. A question solicits a reply.

A question is a non-SIS question if one or more of these assumptions are not met. For example, when a physics teacher grills students with a series of questions in a classroom (e.g., *What forces are acting on the vehicle in the collision?, What are the directions of the forces? What is the mass of the vehicle?*), they are not SIS questions because they violate assumptions 1, 5, 8, and 10. Teachers know the answers to most questions they ask during these grilling sessions, so they are not modeling bona fide inquiry. Similarly, assumptions are violated when there are rhetorical questions (*When does a person know when he or she is happy?*), gripes (*When is it going to stop raining?*), greetings (*How are you?*), and attempts to redirect the flow of conversation in a group (a hostess asks a silent guest *So when is your next vacation?*). In contrast, a question is a SIS question when a person's computer is malfunctioning and the person asks a technical assistant the following questions: *What's wrong with my computer? How can I get it fixed? How much will it cost?*

The social and pragmatic mechanisms that underlie questions are sometimes important in education on dimensions other than deep learning of an academic subject matter. They are important for acquiring skills of socialization and communication. This does little to clear the picture of what constitutes a good question. Nevertheless, if the goal is to learn deep knowledge of academic content, then good questions are at the deeper levels of Graesser-Person taxonomy and more abstract levels of Mosenthal's taxonomy, with example question stems/expressions such as *why, how, what-caused, what-are-the-consequences, what-if-, what-if-not,* and *so-what*.

Learning Technologies that Improve Questions and Learning

Most teachers, tutors, and student peers do not ask a high density of deep questions (Dillon, 1988; Graesser & Person, 1994) so students have limited exposure to high quality inquiry. There are few role models in school environments for the students to learn good question asking and answering skills vicariously. This presents a golden opportunity for turning to technology to help fill this gap.

In the early 1990s, Graesser, Langston, and Baggett (1993) developed and tested *Point & Query* (P&Q) software that pushed the limits of learner question asking and that exposed the learner to a broad profile of question categories. College students learned about musical instruments entirely by asking questions and interpreting answers to questions. This system was a combination of a hypertext/hypermedia system and a question asking and answering facility (see Graesser, Hu, Person, Jackson, & Toth, 2004, for a recent application of P&Q on the subject matter of research ethics). In order to ask a question, the student points to a hot spot on the display (e.g., *the double reed of an oboe*) and clicks a mouse. Then a list of questions about the selected object or area of an object (e.g., the double reed of an oboe) is presented. Example questions are: *What does a double reed look like? What does an oboe sound like?* and *How does a double reed affect sound quality?* The learner subsequently clicks on the desired question, and an answer immediately appears. Therefore, the learner can ask a question very easily by two quick clicks of a mouse.

Research on the P&Q software proved to be quite illuminating from the standpoint of the quantity and quality of questions. Regarding quantity, learners ended up asking a mean of 120 questions per hour, which is approximately 700 times the rate of questions in the classroom (Graesser & Person, 1994). This is a dramatic increase in the rate of question asking, but the

quality of question asking is also important to consider. The students in all conditions were exposed to both low- quality (shallow) and high-quality (deep) questions on the menu of question options on the P&Q software. The results revealed that the quality of student questions did not improve by simply exposing the students to menus of high-quality questions associated with hot spots in hypertext/hypermedia pages. When students explored the hypertext space on their own, they overwhelmingly tended to pose questions that tapped shallow knowledge much more often than deep knowledge. The only way to get the students to ask and explore deep questions was to give them task objectives that directly required deep learning, such as "Your goal is to design a new instrument that has a deep pure tone." A satisfactory solution to this task required an understanding of how the dynamics of air reeds causes changes in the quality of sound and how the size of an instrument determines the resulting pitch. The questions selected by the students were highly skewed to the shallow end of the continuum unless there was a task goal that required an understanding of the science of sound.

A different approach to using technology is to use animated pedagogical agents to model good inquiry and to have the student vicariously observe such skills. The student could observe a curious learner agent who asks good questions while narrating a journey through the learning materials. The student could observe two agents having a conversation, with one asking good deep questions about the learning materials and the other agent giving deep explanation-based answers. Instead of relying on humans to do this, computerized agents can provide the training both rigorously and tirelessly.

Animated pedagogical agents have become increasingly popular in recent advanced learning environments (Atkinson, 2002; Baylor & Kim, 2005; Graesser, Lu et al., 2004; McNamara, Levinstein, & Boonthum, 2004; Moreno & Mayer, 2004; Reeves & Nass, 1996). These agents

interact with students and help them learn by either modelling good pedagogy or by holding a conversation directly with the student. The agents may take on different roles: mentors, tutors, peers, players in multiparty games, or avatars in the virtual worlds. In some systems, such as AutoTutor (Graesser, Lu et al., 2004), the students hold a conversation with the computer in natural language. In other systems, the students vicariously observe agents that either present information in monologues, interact with each other in dialogues, or hold conversations with 3 or more agents.

One recent system with agents was designed with the explicit goal of modeling the asking of deep questions during learning. The system is called *iDRIVE*, which stands for *Instruction with Deep-level Reasoning questions In Vicarious Environments* (Craig et al., 2000; Driscoll et al., 2003; Craig, Sullins, Witherspoon, & Gholson, 2006). iDRIVE has dyads of animated agents train students to learn science content by modeling deep reasoning questions in question-answer dialogues. A student agent asks a series of deep questions (based on the Graesser-Person taxonomy) about the science content and the teacher agent immediately answers each question. Learning gains on the effectiveness of iDRIVE on question asking, recall of text, and multiple-choice questions have shown effect sizes that range from 0.56 to 1.77 compared to a condition in which students listen to the monologue on the same content without questions. The version of iDRIVE that asks deep questions produced better learning than (a) a version that asked shallow questions instead of deep questions and (b) a version that gave a monologue and substituted questions with silence to control for time on task (Craig et al., 2006).

AutoTutor is another computer system that is motivated by theories of question-based inquiry and deep learning. AutoTutor is an intelligent tutoring system with an animated pedagogical that helps students learn by holding a conversation with them in natural language (Graesser, Lu et al., 2004). One way of viewing AutoTutor is that it stimulates and continuously

maintains an optimal level of cognitive disequilibrium in learners' minds by presenting thought-provoking challenging questions, sustaining goal-driven inquiry through continuous dialogue, and providing deep answers that exhibit explanations of the material. AutoTutor presents a series of questions or problems that require deep reasoning, as in the case of the conceptual physics problem below.

> When a car without headrests on the seats is struck from behind, the passengers often suffer neck injuries. Why do passengers get neck injuries in this situation?"

Composing an answer is challenging for most students when the ideal answer is lengthy or requires deep explanatory reasoning. A typical student produces only one or two sentences when initially asked one of these conceptual physics problems whereas an ideal answer is a paragraph of information in length (roughly 10 sentences). AutoTutor assists the learner in the evolution of an improved answer that draws out more of the learner's knowledge, that fills in missing information, and that corrects the learner's misconceptions. The dialogue between AutoTutor and student is typically between 50 and 200 *turns* (i.e., the learner expresses something, then the tutor, then the learner, and so on) before a good answer to this single physics question emerges.

There are four interesting results from the AutoTutor research with respect to questions and learning. Two of the findings present optimistic news and the other two suggest there are limitations in this technology. First, assessments of AutoTutor on learning gains in 15 experiments have shown effect sizes of approximately 0.8 standard deviation units in the areas of computer literacy (Graesser et al., 2004) and Newtonian physics (VanLehn, Graesser et al., 2007) compared with suitable control conditions (e.g., pretests, textbook reading control). These evaluations place AutoTutor somewhere between an untrained human tutor (Cohen, Kulik, & Kulik, 1982) and an intelligent tutoring system with ideal tutoring strategies (Corbett, 2001).

Second, the interactive conversational aspects of AutoTutor show advantages over non-interactive content control conditions, but this advantage in interactivity only occurs when the subject matter content is more advanced than what the student already knows; otherwise, the interactive AutoTutor and non-interactive control conditions produce equivalent learning gains (Van Lehn et al., 2007). Third, AutoTutor is effective in modeling deep questions because the proportion of student questions that are deep increases as a consequence of the interactions with AutoTutor (Graesser, McNamara, & Van Lehn, 2005). Fourth, the modeling of deep questions is limited if AutoTutor provides poor answers to the questions. In fact, many students stop asking questions altogether if they are not satisfied with the quality of the answers that AutoTutor provides. An intelligent interactive learning environment may not be the best choice when the quality of the automated responses are poor or marginal. The more suitable alternative would be a choreographed dialogue between agents (such as iDRIVE) that exhibits excellent inquiry from which the student can vicariously learn.

A Question Authoring Workbench

The previous sections in this chapter have made a number of claims about the relationships between questions and learning. Available research indicates that: (1) most students ask few questions in most learning settings, (2) questions of students, teachers, and textbook writers tend to be shallow rather than deep, (3) training students to ask deeper questions facilitates comprehension and learning, (4) there are a number of psychological models that specify how to stimulate more questions and deeper questions, e.g., such as teachers or agents modeling deep questions or presenting challenges that place the students in cognitive disequilibrium. Given this research context, the time is ripe for a research team to develop a *Question Authoring Workbench*. The Workbench would train instructors, students, textbook

writers, curriculum developers, test constructors and other educational communities to compose better questions that elevate standards of learning in the educational enterprise. The Workbench would train question developers on theoretical principles of quality questions, present examples of high-quality questions, and guide the user of the Workbench in creating questions in different cells of a large landscape of questions.

A Sketch of the Question Authoring Workbench

The proposed Workbench would have several modules that vary in the degree of interactivity with the user and in the sophistication of its computational components that automatically analyze language, discourse, and world knowledge. For example, a *didactic instruction* module in Workbench would be a repository of guidelines for the creation of different categories of questions, at varying levels of depth and relevance to the learning objectives. This might be organized according to the landscape of questions in the first section in this chapter. A *scripted exemplar* module would augment the didactic instruction module by presenting and explaining questions that cover the broad landscape of question categories, types of knowledge, and cognitive skills. A *scripted interactive* module would go a step further by guiding the question writer in composing questions for a pre-selected corpus of texts, including explanatory feedback on the writer's contributions. A *question evaluation* module would critique questions created by the question writers on a pre-selected sample of texts. The four modules would hopefully create a learning environment that could be used by students in addition to researchers, teachers, textbook writers, and other professions in education.

Our Workbench vision is compatible with research efforts at Educational Testing Service that draw on the expertise of multidisciplinary research teams with cognitive science, artificial intelligence, linguistics, and education (Bejar, 1993, 2002; Deane & Sheehan, 2003; Graff,

Steffen, & Lawless, 2004). There are reasons for being optimistic that Workbench modules can to some extent analyze language and provide adaptive feedback because there have been major advances in computational linguistics (Jurafsky & Martin, 2000), statistical representations of world knowledge (Landauer, McNamara, Dennis, & Kintsch, 2007), and discourse processes (Graesser, Gernsbacher, & Goldman, 2003). The first author of this chapter has been involved with developing a number of systems that automatically analyze natural language in addition to the AutoTutor system described earlier. For example, *QUAID* (Question Understanding Aid) is a web tool that analyzes questions on the difficulty of the words, syntactic complexity, and working memory load (Graesser, Cai, Louwerse, & Daniel, 2006). *Coh-Metrix* (Graesser, McNamara, Louwerse, & Cai, 2004) is a tool on the web that analyzes texts on hundreds of measures of words, syntactic complexity, referential cohesion, coherence of mental models, and genre. A Question Authoring Workbench would be a feasible extension of these projects that integrate advances in computational linguistics, discourse processes, education, and cognitive science.

We imagine a QA Workbench that could accommodate questions in a *multiple choice* (MC) format in addition to questions without response options. Nearly all sectors of education and training have relied on the MC question format: K-12 teachers, university professors, the textbook industry, the College Board and Educational Testing Service (e.g., SAT, GRE), on-line universities, and training materials for business, industry, government, and the military. MC questions prevail in the formative assessments that occur during the learning process in addition to the summative assessments that occur at the end of training and the completion of education milestones. Assessment methods are currently moving toward more constructive response format such as open-ended questions (Leacock & Chodorow, 2003), essays (Burstein, 2003;

Foltz, Gilliam, & Kendall, 2000), mathematical derivations, and voice recordings (Zechner, Bejar, & Hemat, 2005), but these assessments are not likely to replace MC questions entirely. Moreover, the case can be made that (a) MC questions are more advantageous relative to open-ended questions under certain circumstances and (b) most of the criticisms of MC questions are largely based on the poor quality of multiple-choice items existing in the field, not necessarily because the multiple-choice question format is intrinsically unsuitable for assessment of deep knowledge or higher-level cognitive skills.

Quality of MC Questions

Most readers of this chapter will have completed thousands of multiple-choice tests during their academic history and will have constructed thousands of such questions themselves. However, the technical properties of good multiple choice questions are not widely known. Multiple choice questions typically have a question *stem* and a list of *response options*. The *key* is the most accurate response option whereas *distracters* are incorrect response options.

The quality of the MC questions can be evaluated with respect to the landscape of questions described in the first section of this chapter. However, an important additional consideration is the selection of the response options. To what extent is each response option a plausible answer and to what extent can the response options be differentiated? It is recommended that the distracters should vary by degree (Downing & Haladyna, 1997). For example, one distracter should be the *near miss*. This option is the most seductive distracter that reflects a common misconception that learners have. The discrimination between the key and the near miss should reflect an important learning objective or pedagogical point rather than testing an obscure, arbitrarily subtle, or unenlightening detail. The *thematic distracter* has content that is related to the topic at hand, but is not correct. A learner who quickly scans the learning materials

would have trouble discriminating the thematic distracter from the key and near miss. The *unrelated distracter* would seem reasonable to someone who never read the material, but might be plausible according to folklore and world knowledge. Our analyses of the corpus of MC questions about psychology (reported earlier) indicated that distracters rarely followed such systematic principles. As a result, the items were quite vulnerable to guessing or other unwanted processes that help learners identify the target without appropriately understanding the material. Instructors and question writers for textbooks rarely implement systematic methods of generating questions that are routinely implemented, at great expense, in the College Board and ETS.

There are circumstances when MC questions with appropriate answer options are more discriminating in assessment of learning than questions with open-ended formats without response options. As an example, in one of our research projects we compared AutoTutor with various control conditions, such as reading yoked chapters from a textbook or reading a text that has information equivalence to AutoTutor. Learning gains for AutoTutor versus comparison conditions were assessed with over a dozen measures: (a) MC questions that tap shallow knowledge, such as definitions, facts, and properties of concepts, (b) MC questions that tap deep knowledge, such as causal reasoning, justifications of claims, and functional underpinnings of procedures, (c) written essays, (d) cloze tasks that require students to fill in missing words in texts to articulate explanatory reasoning on the subject matter, and (e) assessment tasks requiring problem solving (Graesser, Lu, Jackson, et al., 2004; VanLehn, et al., 2007). We were happy to document the advantages in learning of AutoTutor over various controls (Graesser, Lu, Jackson, et al., 2004), but the important question arises as to which measures of evaluation were most sensitive in showing differences between AutoTutor and comparison conditions.

We have found that the MC questions that tap deep knowledge have consistently been the most sensitive measures of learning. However, this claim should be qualified by the constraints that we imposed on the construction of the MC questions. We adopted a principled framework for generating MC questions that tap the mental models that underlie science content. The framework was inspired by the field of qualitative physics (Forbus, 1984) and instantiated in the Force Concept Inventory for Newtonian physics (Hestenes, Wells, & Swackhamer, 1992). Suppose there is a set of N "nodes" in a scientific system; a node refers to a physical component, a part of a component, a system state, an event, a process, or an action. The set of N nodes are connected by a network of -, +, and 0 causal relations. If node C is disturbed or changed in some fashion (e.g., increased, broken, moved), how would it propagate its effects on the other nodes in the system (e.g., nodes X, Y and Z)? There may be three alternative answers that reflect the impact on an effected node, such as: (a) X increases, (b) X decreases, and (c) X stays the same. For example, the following question might be asked:

> When the passenger in a car is rear-ended, does the head initially (a) go forward, (b) go backwards, or (c) stay the same?

A deep comprehender is able to trace the causal antecedents and causal consequences of an event, whereas a poor comprehender is undiscriminating in tracking the impact of one event on other events in the scientific system.

The strong form of our claim, based on the above finding, is that well-constructed MC questions that tap deep knowledge can be a more sensitive measure than open-ended questions and essays. MC questions can be composed to be more discriminating and permit the researcher to target particular ideas and misconceptions. This may be particularly true when target knowledge involves groups of concepts that have formally defined conceptual relations. Each

distracter option can represent different (sometimes erroneous) conceptual relations with the target knowledge, the question stem, or both, as in the case of science or mathematics. In contrast, recall tests and open-ended questions run the risk of allowing students to get by with vague or verbose answers that sidestep subtle discriminations among interrelated concepts in scientific explanatory systems. MC questions can potentially assess mastery of a broad diversity of content and skills that vary in subtlety, granularity, and depth. It is an empirical question, of course, on whether the MC questions can achieve these ambitious goals. Ideally, MC questions could be developed that are firmly grounded in cognitive and learning sciences, as in the case of evidence-centered design (Mislevy, Steinberg, & Almond, 2003), tree-based regression (Sheehan, 2003), and rule space procedures (Buck, VanEssen, Tatsuoka, & Kostin, 1997).

Final Comments

In this chapter we have advocated the development of a Question Authoring Workbench that would assist students, teachers, textbook writers, and researchers in composing better questions. The current questions prepared by teachers and textbook writers are often both shallow and poorly crafted, which in part explains why so many students have shallow standards for what it means to comprehend. The College Board and ETS have much better questions, but it is extremely expensive to design questions that satisfy constraints that serve psychometric engineering goals, theoretical components in the cognitive and learning sciences, and the standards articulated in educational communities. There needs to be an interdisciplinary confluence between cognitive science, psychometrics, computational linguistics, and education in order to solve this problem.

The QA Workbench would expose the users to a landscape of questions that cross types of questions, knowledge, and cognitive processes. A simple QA Workbench would provide

didactic instruction on the landscape of questions and would present examples of both good and bad questions in the landscape. A traditional hypertext/hypermedia system would provide an adequate technology for providing these facilities. A more sophisticated QA Workbench would prompt the user to generate questions and would give feedback on the quality of generated questions. This would require more sophisticated technologies from computational linguistics, discourse processes, artificial intelligence, and cognitive science.

The value of the Workbench is bolstered by available research that has investigated relationships between learning and questions. Available research indicates that most students ask few questions in most learning settings; the questions of students, teachers, and textbook writers tend to be shallow rather than deep; training students to ask deeper questions facilitates comprehension and learning; and there is an abundance of cognitive, discourse, and pedagogical theories that specify how to stimulate deeper questions.

It should be apparent that there is not a simple answer to *What is a good question?* We hope that this chapter provides the reader with some new perspectives and a comprehensive snapshot of our fundamental and unending quest. And Isabel Beck's *Questioning the Author* (Beck et al., 1997; McKeown et al., 1999) is positioned smack dab in the center of the puzzle.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94*, 416-427.

Baker, L. (1985). How do we know when we don't understand? Standards for evaluating text comprehension. In D.L. Forrest-Pressley, G.E. Mackinnon, T.G. Waller (Eds) *Metacognition, cognition and human performance* (pp. 155-205). New York: Academic Press.

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15*, 95-115.

Beck, I.L., McKeown, M.G., Hamilton, R.L., & Kucan, L. (1997). *Questioning the Author: An approach for enhancing student engagement with text*. Delaware: International Reading Association.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Generating items from cognitive tests: Theory and practice* (pp. 199- 217). Mahwah, NJ: Lawrence Erlbaum

Bejar, I. I. (1993). A generative analysis of a three dimensional spatial task. *Applied Psychological Measurement, 14*, 237-245.

Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain*. New York: McKay.

Burstein, J. (2003). The *e-rater*® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Buck, G., Van Essen, T., Tatsuoka, K., & Kostin, I. (1997). Identifying the cognitive attributes underlying performance on the PSAT Writing Test. Unpublished proposal. Princeton, NJ: Educational Testing Service.

Ciardiello, A.V. (1998). Did you ask a good question today? Alternative cognitive and metacognitive strategies. *Journal of Adolescent & Adult Literacy, 42*, 210-219.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new structure of intellect. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27-56). Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1987). Psychometric approaches to cognitive abilities and processes. In S. H. Irvine & S. E. Newstead (Eds.), *Intelligence and cognition: Contemporary frames of reference* (pp. 217–251). Boston: Martinus Nijhoff.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533.

Chinn, C., & Brewer, W. (1993) The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*, 1-49.

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.

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

Corbett, A.T. (2001). Cognitive computer tutors: Solving the two-sigma problem. *User Modeling: Proceedings of the Eighth International Conference, UM 2001, 137-147.*

Craig, S. D., Gholson, B., Ventura, M., Graesser, A. C., & the Tutoring Research Group. (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11*, 242-253.

Craig, S. D., Sullins, J., Witherspoon, A. & Gholson, B. (2006). Deep-level reasoning questions effect: The role of dialog and deep-level reasoning questions during vicarious learning. *Cognition and Instruction, 24(4),* 563-589.

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

Deane, P. & Sheehan, K. (2003). Automatic item generation via frame semantics. Education Testing Service: http://www.ets.org/research/dload/ncme03-deane.pdf.

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

Downing, S. M. & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61-82.

Edelson, D.C., Gordin, D.N., & Pea, R.D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *The Journal of the Learning Sciences, 8,* 391-450.

Driscoll, D.M., Craig, S.D., Gholson, B., Ventura, M.,Hu, X., & Graesser, A.C. (2003). Vicarious learning: Effects of overhearing dialog and monolog-like discourse in a virtual tutoring session. *Journal of Educational Computing Research*, 29, 431-450.

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.

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

Foltz, W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*, 111-128.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence, 24*, 85–168.

Floyd, R. G. (2005). Information-processing approaches to interpretation of contemporary intellectual assessment instruments. In D. P. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp., 203–233). New York: Guilford Press.

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

Goldman, S. R., Duschl, R. A., Ellenbogen, K., Williams, S., & Tzou, C. T. (2003). Science inquiry in a digital age: Possibilities for making thinking visible. In H. van Oostendorp (Ed.), *Cognition in a digital world.* (pp. 253-284) Mahwah, NJ: Erlbaum.

Good, T.L., Slavings, R.L., Harel, K.H., & Emerson, M. (1987). Students' passivity: A study of question asking in K-12 classrooms. *Sociology of Education, 60,* 181-199.

Graesser, A.C., Cai, Z., Louwerse, M., Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that helps survey methodologists improve the comprehensibility of questions. *Public Opinion Quarterly, 70*, 3-22.

Graesser, A.C., Gernsbacher, M.A., & Goldman, S. (2003)(Eds.). *Handbook of discourse processes*. Mahwah, NJ: Erlbaum.

Graesser, A. C., Gordon, S. E., & Brainerd, L. E. (1992). QUEST: A model of question answering. *Computers and Mathematics with Applications, 23*, 733-745.

Graesser, A.C., Hu, X., Person, P.,Jackson, T., and Toth, J (2004). Modules and information retrieval facilities of the Human Use Regulatory Affairs Advisor (HURAA). *International Journal on eLearning, 3,* 29-39.

Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In G. V. Nakamura, R. M. Taraban, & D. Medin (Eds.), *The psychology of learning and motivation: Vol. 29. Categorization by humans and machines* (pp. 411-436). Orlando, FL: Academic Press.

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., Lu, S., Jackson, G.T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M.M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36,* 180-193.

Graesser, A.C., Lu, S., Olde, B.A., Cooper-Pye, E., & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory and Cognition, 33,* 1235-1247.

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

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36,* 193-202.

Graesser, A.C., McNamara, D.S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist, 40,* 225-234.

Graesser, A.C., Millis, K.K., & Zwaan, R.A. (1997). Discourse comprehension. *Annual Review of Psychology, 48,* 163-189.

Graesser, A.C., & Olde, B.A. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology, 95,* 524-536.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31,* 104-137.

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

Graff, E. A., Steffen, M. & Lawless, R. (2004). *Statistical and cognitive analysis of quantitative item models*. Paper presented at the 30th annual conference of the International Association for Educational Assessment (IAEA), Philadelphia, PA.

Guthrie, J.T. (1988). Locating information in documents: Examination of a cognitive model. *Reading Research Quarterly, 23,* 178-199.

Hacker, D.J., Dunlosky, J., & Graesser, A.C. (Eds.) (1998). *Metacognition in educational theory and practice*. Mahwah, NJ: Erlbaum.

Hestenes, D., Wells, M. & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30,* 141-158.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

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

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal, 29,* 303-323.

King A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31*, 338-368.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kyllonen, P. C., & Roberts, R. D. (2003). Cognitive process assessment. In R. Fernandez-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 200-203). London: Sage.

Landauer, T., McNamara, D.S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook on Latent Semantic Analysis.* Mahwah, NJ: Erlbaum.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37,* 389-405.

Lehmann, F (1992). *Semantic Networks in Artificial Intelligence*. New York: Elsevier Science Inc.

Lehnert, W. G. (1978). *The Process of Question Answering: a computer simulation of cognition.* Lawrence Erlbaum Associates.

Lenat. D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*, 1995.

McKeown, M. G., Beck, I. L., Hamilton, R., & Kucan, L. (1999). *Questioning the Author Accessibles: Easy access resources for classroom challenges*. Bothell, WA: The Wright Group.

McNamara, D.S., Levinstein, I.B. & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers, 36*, 222-233.

Moreno, R., & Mayer, R. E. (2004). Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology, 96*, 165-173.

Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88*, 314-332.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.

Otero, J., & Graesser, A.C. (2001). PREG: Elements of a model of question asking. *Cognition & Instruction, 19,* 143-175.

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

Perfetti, C.A., Britt, M.A., & Georgi, M. (1995). *Text-based learning and reasoning: Studies in history*. Hillsdale, NJ: Erlbaum.

Pressley, M., & Forrest-Pressley, D. (1985). Questions and children's cognitive processing. In A.C. Graesser, & J.B. Black (Eds.), *The psychology of questions* (pp. 277-296). Hillsdale, NJ: Erlbaum.

Reder, L. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*, 90-138.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, televisions, and new media like real people and places*. Cambridge, U.K.: Cambridge University Press.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research, 66,* 181-221.

Rouet, J. (2006). *The skills of document use: From text comprehension to web-based learning.* Mahwah, NJ: Erlbaum.

Schank, R. (1999). *Dynamic memory revisited*. New York: Cambridge University Press.

Sheehan, K.M. (2003). Tree-based regression: A new tool for understanding cognitive skill requirements. In H. F. O'Neal & R.S. Perez (Eds.), *Technology and applications in education* (pp.222-227).Mahwah, NJ: Lawrence Erlbaum Associates.

Singer, M. (2003). Processes of question answering. In G. Rickheit, T. Hermann, & W. Deutsch (Eds.), *Psycholinguistics* (pp.422-431). Berlin: Walter de Gruyter.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND Corporation.

Sowa, J. F. (1983). *Conceptual structures: Information processing in mind and machine*. Addison Wesley.

Sternberg, R. J., & Perez, J. E. (Eds.). (2005). *Cognition and intelligence: Identifying mechanisms of the mind.* New York: Cambridge.

Van der Meij, H. (1987). Assumptions if information-seeking questions. *Questioning Exchange, 1*, 111-118.

Van der Meij, H. (1994). Student questioning: A componential analysis. *Learning and Individual Differences, 6*, 137-161.

VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., & Rose, C.P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 3-62.

Wisher, R.A., & Graesser, A.C. (2007). Question asking in advanced distributed learning environments. In S.M. Fiore and E. Salas (Eds.), *Toward a science of distributed learning and training* (pp. 209-234). Washington, D.C.: American Psychological Association.

Wong, B.Y.L. (1985). Self-questioning instructional research: A review. *Review of Educational Research, 55*, 227-268.

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

Zechner, K., Bejar, I. I., & Hemat, R. (2005, January). Towards an understanding of the role of speech recognition in non-native speech assessment. A paper presented at Review of Assessment Design Capabilities Initiative. Educational Testing Service.

Author Notes

The research on AutoTutor was supported by the National Science Foundation (SBR 9720314, REC 0106965, REC 0126265, ITR 0325428, REESE 0633918), the Institute of Education Sciences (R305H050169, R305B070349), and the Department of Defense Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of researchers from psychology, computer science, physics, and education (visit http://www.autotutor.org). The research on QUAID was supported by the National Science Foundation (SES 9977969) and the research on Coh-Metrix was supported by Institute for Education Sciences (IES R3056020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or DoD. Requests for reprints should be sent to Art Graesser, Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152-3230, a-graesser@memphis.edu.

Table 1

Question taxonomy proposed by Graesser and Person (1994)

QUESTION CATEGORY	GENERIC QUESTION FRAMES AND EXAMPLES
1. Verification	Is X true or false? Did an event occur? Does a state exist?
2. Disjunctive	Is X, Y, or Z the case?
3. Concept completion	Who? What? When? Where?
4. Example	What is an example or instance of a category?.
5. Feature specification	What qualitative properties does entity X have?
6. Quantification	What is the value of a quantitative variable? How much? How many?
7. Definition	What does X mean?
8. Comparison	How is X similar to Y? How is X different from Y?
9. Interpretation	What concept or claim can be inferred from a static or active pattern of data?
10. Causal antecedent	What state or event causally led to an event or state? Why did an event occur? Why does a state exist? How did an event occur? How did a state come to exist?
11. Causal consequence	What are the consequences of an event or state? What if X occurred? What if X did not occur?
12. Goal orientation	What are the motives or goals behind an agent's action? Why did an agent do some action?
13. Instrumental/procedural	What plan or instrument allows an agent to accomplish a goal? How did agent do some action?
14. Enablement	What object or resource allows an agent to accomplish a

	goal?
15. Expectation	Why did some expected event not occur?
	Why does some expected state not exist?
16. Judgmental	What value does the answerer place on an idea or advice?
	What do you think of X? How would you rate X?

Figure 1. *Landscape of questions.*