

# How Does One Know Whether a Person Understands a Device? The Quality of the Questions the Person Asks When the Device Breaks Down

Arthur C. Graesser and Brent A. Olde  
University of Memphis

Models of question asking predict that questions are asked when comprehenders experience cognitive disequilibrium, which is triggered by contradictions, anomalies, obstacles, salient contrasts, and uncertainty. Questions should emerge when a person studies a device (e.g., a lock) and encounters a breakdown scenario (“the key turns but the bolt doesn’t move”). Participants read illustrated texts and breakdown scenarios, with instructions to ask questions or think aloud. Participants subsequently completed a device-comprehension test, and tests of cognitive ability and personality. Deep comprehenders did not ask more questions, but did generate a higher proportion of good questions about plausible faults that explained the breakdowns. An excellent litmus test of deep comprehension is the quality of questions asked when confronted with breakdown scenarios.

It could be argued that questions are at the heart of virtually any complex task that an adult performs. That is, any given task can be decomposed into a set of questions that a person asks and answers. For example, when a person encounters a device that malfunctions, the relevant questions are “What’s wrong?” and “How can it be fixed?” When a person reads an office memo, the relevant questions are “Why is this important?” and “What should I do about it, if anything?” When an adult reads a job ad, the relevant questions are “Am I a good match for the job?” “How much would I make?” and “What are the perks?” The cognitive mechanisms that trigger question-asking and exploration patterns need to be understood in a query-based (or inquiry-based) theory of task analysis, comprehension, and learning.

The purpose of this study was to test some predictions of a cognitive model of question asking, called PREG (Graesser et al., in press; Otero & Graesser, 2001), and of an earlier model proposed by Graesser and McMahan (1993). According to PREG and the Graesser-McMahan model, cognitive disequilibrium drives the asking of sincere information-seeking questions (as opposed to questions that merely are part of monitoring the flow of conversation among speech participants). Questions are asked when

individuals are confronted with obstacles to goals, anomalous events, contradictions, discrepancies, salient contrasts, obvious gaps in knowledge, expectation violations, and decisions that require discrimination among equally attractive alternatives. The answers to such questions are expected to restore equilibrium and homeostasis.

The notion that questions are driven by cognitive disequilibrium (or dissonance) has a long history in cognitive, developmental, social, and educational psychology (Berlyne, 1960; Chinn & Brewer, 1993; Collins, 1988; Dillon, 1988; Festinger, 1957; Flammer, 1981; Graesser, Baggett, & Williams, 1996; Miyake & Norman, 1979; Piaget, 1952; Pressley, Symons, McDaniel, Snyder, & Turner, 1988; Schank, 1999). Surprisingly, however, there have been few systematic tests of the relationship between cognitive disequilibrium and question asking (Graesser & McMahan, 1993). What is distinctive about the PREG model is that it attempts to predict what questions are asked and when they are asked. That is, Otero and Graesser (2001) developed a set of production rules that specify the categories of questions that are asked under particular conditions (i.e., content features of text and knowledge states of individuals). Some of the production rules are very sensitive to the amount of domain knowledge that the learner has achieved because it often takes a large amount of knowledge to identify some clashes and gaps in knowledge. Miyake and Norman (1979) presented the argument over 20 years ago that “to ask a question, one must know enough to know what is not known” (p. 357).

Questions that tap explanatory reasoning are particularly diagnostic of deep comprehension (Graesser et al., 1996). The construction of explanations is a robust predictor of an adult’s ability to learn technical material from written texts and other forms of verbal material (Chi, de Leeuw, Chiu, & LaVancher, 1994; Coté, Goldman, & Saul, 1998; Graesser, VanLehn, Rose, Jordan, & Harter, 2001; VanLehn, Jones, & Chi, 1992). Explanation-based questions are frequently (but not always) signaled by the following question stems: *why*, *how* (did X occur), *what are the consequences*, *what if*, and *what if not*. Answers to explanation-based questions tap causal chains and networks, goal-plan-action hierarchies, and logical justifications. The derivation of the answers

---

Arthur C. Graesser and Brent A. Olde, Department of Psychology, University of Memphis.

Brent A. Olde is now at the Operational Psychology Department, Naval Aerospace Medical Institute, Pensacola, Florida.

This research was supported by Office of Naval Research (ONR) Grants N00014-98-1-0331, N00014-00-1-0600, and N00014-00-1-0917 and National Science Foundation (NSF) Grants REC 0126265 and REC 0106965 awarded to Arthur C. Graesser. Any opinions, findings, conclusions, or recommendations expressed in this article do not necessarily reflect the views of ONR or NSF. We thank Scotty Craig, Shulan Lu, Victoria Pomeroy, and Shannon Whitten for their help in data collection, data coding, and statistical analyses at various stages of this project.

Correspondence concerning this article should be addressed to Arthur C. Graesser, Department of Psychology, University of Memphis, 202 Psychology Building, Memphis, Tennessee 38152-3230. E-mail: a-graesser@memphis.edu

normally requires inferences, hypothetical reasoning, and other processes at the higher levels of Bloom's (1956) taxonomy of cognitive objectives. When one considers equipment, explanations are needed when devices break down, faults are diagnosed, and devices are repaired (e.g., Why-how did the device break? Is this part stuck? How can I fix it?). It follows, therefore, that an individual who comprehends a device at a deep level should produce questions that identify those faults that explain equipment breakdowns and how they might be repaired.

In this study, we investigated the questions that college students ask when an everyday device malfunctions. One of the six illustrated texts that were used in the present study, a text about a cylinder lock, is presented in Figure 1. Our texts were directly extracted from David Macaulay's *The Way Things Work* (1988), whose illustrated texts succinctly capture the components and mechanisms of devices. After reading the illustrated text about the cylinder lock for 5 min, the participant subsequently received a scenario in which the device breaks down (e.g., "the key turns but the bolt doesn't move"). As soon as the breakdown scenario appeared (on a page without the illustrated text), the participant generated questions about the malfunction. The breakdown presumably placed the comprehenders in cognitive disequilibrium; once comprehenders identified the fault, equilibrium would be

restored. In this case, the most likely faults are the cam's failure to rotate or the lip of the cam failing to catch the rod (which in turn pulls back the bolt). The fault would not reside in a problem with the pins rising because the key is inserted and successfully turns. The fault also does not reside in a broken spring, even though it is spatially close to the malfunctioning bolt; the spring can assist the bolt in moving, but it cannot prevent the bolt from moving. A question was defined as a high-quality question if it referred to a component, event, or process that was a plausible cause of the breakdown.

The PREG model and the Graesser-McMahan model predict that participants who have a deep understanding of the device should ask good questions that converge on faults that explain the breakdown. To test this prediction, we performed a correlational study. We measured the volume and quality of questions that participants ask when confronted with a breakdown scenario. We correlated these measures with an objective test of device comprehension and with a battery of measures of cognitive ability and personality. There should be a high positive correlation between measures of question answering and (a) the scores on an objective comprehension test on the devices and (b) measures of cognitive ability that capture mechanical or electronic reasoning.

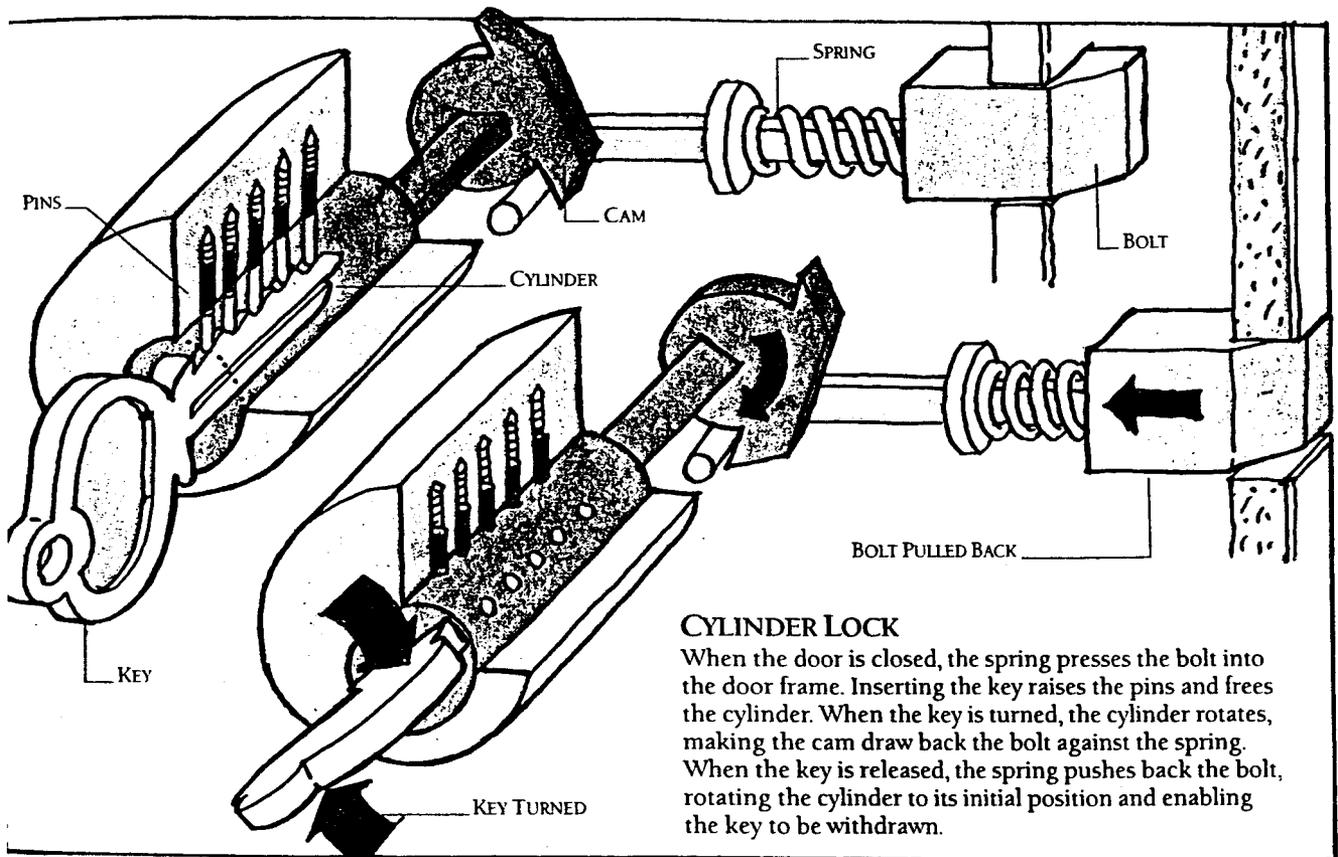


Figure 1. Example of illustrated text describing a cylinder lock. From *The Way Things Work* (p. 17) by David Macaulay, 1988, Boston: Houghton Mifflin. Compilation copyright Dorling Kindersley Ltd., London. Illustration copyright 1988 David Macaulay. Text copyright David Macaulay, Neil Ardley. Reprinted by permission of Houghton Mifflin Company. All rights reserved.

### Previous Research on Question Asking

Researchers in cognitive science and education have routinely advocated learning environments that encourage students to generate questions (Beck, McKeown, Hamilton, & Kucan, 1997; Bransford, Goldman, & Vye, 1991; Brown, 1988; Dillon, 1988; Edelson, Gordin, & Pea, 1999; Guthrie & McCann, 1997; Palincsar & Brown, 1984; Papert, 1980; Piaget, 1952; Pressley & Forrester-Pressley, 1985; Scardamalia & Bereiter, 1985; Schank, 1999; van der Meij, 1994; Zimmerman, 1989). There are several reasons why the process of question generation might play a central role in learning. The most frequently articulated rationale is that it promotes active learning and construction of knowledge. The learner actively constructs knowledge during learning rather than passively receives information. Another reason is that question asking has the potential for enhancing motivation because information acquisition is student centered. Yet another reason is that question asking encourages the learners to develop sophisticated metacognitive skills. Learners identify their own knowledge deficits, ask question that focus on these deficits, and answer the questions by exploring reliable information sources.

The assumption that learners are vigorous question generators who actively self-regulate their learning is now known to be idealistic. The truth is that the vast majority of learners have trouble identifying their own knowledge deficits (Baker, 1985; Hacker, Dunlosky, & Graesser, 1998) and ask very few questions (Dillon, 1988; Good, Slavings, Harel, & Emerson, 1987; Graesser & Person, 1994). The typical student asks .17 question per hour in a classroom, and the poverty of classroom questions is a general phenomenon across cultures (Graesser & Person, 1994). The fact that it takes about 6–7 hr 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. An average student asks 26.5 question per hour in one-on-one human tutoring sessions (Graesser & Person, 1994). Thus, when there is an attentive question answerer, the rate of student question asking goes up over 200-fold. However, Graesser and Person (1994) also reported that very few of the students' questions reflect deep comprehension, for example, questions such as "why," "why not," "how," "what if," or "what if not." This low incidence of deep questions by students reflects the classroom environment. Only about 4% of the questions asked by teachers are deep questions (Dillon, 1988; Kerry, 1987).

Given that student questions are infrequent, one might expect significant learning gains by teaching students how to ask questions. Indeed, it is well documented that improvements in the comprehension, learning, and memory of technical material can be achieved by training students to ask questions during comprehension (Ciardiello, 1998; Craig et al., 2000; Davey & McBride, 1986; King, 1989, 1992, 1994; Palincsar & Brown, 1984; Rosenshine, Meister, & Chapman, 1996; Singer & Donlan, 1982; van der Meij, 1994). The process of question generation accounts for a significant amount of these improvements in comprehension, over and above the information supplied by answers. Moreover, question generation learning is most effective when students are trained to ask good questions. Rosenshine et al. (1996) provided the most comprehensive analysis of the impact of question asking on learning in their meta-analysis of 26 empirical studies that compared

question generation 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. The fact that question-generation training is effective in promoting learning is very well documented but is not the focus of the present study. Instead, in the present study we investigated the conditions that elicit questions, the extent to which such questions reflect deep comprehension, and theoretical components in models of question asking.

Much of the research reviewed in this section has glossed over the potentially important relationship between cognitive disequilibrium and question asking, which is the hallmark of the PREG model (Graesser et al., in press; Otero & Graesser, 2001) and the model of Graesser and McMahan (1993). Student questions are rare in those classroom environments in which teachers are prone to deliver organized curricula of shallow knowledge (Dillon, 1988; Graesser & Person, 1994), rather than having students experience cognitive disequilibrium through challenges, obstacles, contradictions, and other conditions that require deep knowledge. The volume and quality of student questions may substantially increase when students experience cognitive disequilibrium, as in the case of the breakdown scenarios in the present study. However, a sufficient amount of domain knowledge and reasoning ability is apparently necessary for individuals to generate questions in the face of cognitive disequilibrium (Miyake & Norman, 1979). This motivated us to explore the relationship between question asking and individual differences when learners are experiencing cognitive disequilibrium.

The detection of cognitive disequilibrium is alone not sufficient for question generation. According to Graesser and McMahan (1993), the potential question asker must pass two additional hurdles after *disequilibrium detection*: the articulation of the disequilibrium in words (called *verbal coding*) and the initiative to express the question in a social setting (called *social editing*). Thus, three stages need to be intact for a sincere information-seeking question to be produced (disequilibrium detection, verbal coding, and social editing). Graesser and McMahan investigated this by having college students read different versions of stories and mathematical word problems in a laboratory environment: a version with contradictions between text statements, a version that had the deletion of critical information, a version that inserted irrelevant information, and a control version (no anomalies). The likelihood of generating questions was higher in the anomaly conditions than in the control condition when participants were instructed to generate questions. It is informative to note, however, that the likelihood of students asking questions was extremely low under *self-induced* question asking (.04) compared with *task-induced* question asking (.50). *Task-induced* question asking occurs when participants are directly instructed to ask questions. *Self-induced* question asking occurs when participants spontaneously ask questions without being instructed to do so; the participants in the Graesser and McMahan study were free to ask questions to an experimenter in an adjacent room, but they had to take the initiative to do so. The extremely low likelihood of question asking under self-induced conditions suggests that questions do not surface when it is physically effortful or socially awkward to ask them. The social constraints no doubt partly

explain why questions are so rare in classroom settings. If learners are not instructed to generate questions under task-induced conditions, the incidence of questions is extremely low.

In the present study we attempted to maximize the incidence of questions in two ways, in light of the extensive evidence that it is difficult to get students to ask questions. In the question-asking condition, students were explicitly instructed to generate questions when they were presented the breakdown scenario. This was a task-induced condition that encouraged them to express their ideas in the form of questions (verbal coding) and that should remove many of the social barriers to question asking (social editing). In a write-aloud condition, the students were merely asked to think aloud in writing when given the breakdown scenario, without any explicit encouragement for them to ask questions per se. This unconstrained-writing condition should remove potential barriers that might exist if students have trouble formulating their thoughts into an interrogative linguistic form.

The PREG model (Graesser et al., in press; Otero & Graesser, 2001) provides a more detailed specification of the cognitive, metacognitive, and pragmatic mechanisms that elicit cognitive disequilibrium and questions. PREG identifies specific types of questions that are asked when there is a discrepancy between world knowledge and text information at different levels of representations: words, statements (e.g., propositions, clauses, sentences), and links between statements. However, we did not attempt to test these specific predictions in the study reported here. The direct focus in the present study was to test whether there is a significant relationship between (a) the cognitive disequilibrium that is triggered by breakdown scenarios and (b) student questions that elaborate or potentially explain the breakdowns.

### Individual Differences

The volume and quality of learner questions could be influenced by a host of individual differences, such as subject-matter knowledge, comprehension skill, and creativity. However, researchers have not yet investigated the relationship between question asking and such potential differences, so the present study was designed to close this gap. This is the first study to focus on individual-differences assessments under conditions when comprehenders are directly facing cognitive disequilibrium. Moreover, we would anticipate that particular measures of individual differences would map onto theoretical components of question asking.

The measures of individual differences in this study included an objective test on the participants' comprehension of the six devices, a battery of cognitive-ability measures, and a number of noncognitive measures. The objective device-comprehension score served as an index of subject-matter knowledge for the devices investigated. The cognitive-ability measures included the Armed Services Vocational Aptitude Battery (ASVAB; Department of Defense, 1983). This psychometric test is administered to over 1 million high schools students each year. ASVAB is particularly relevant to the present investigation because it has subscales tailored to mechanical and electronic systems. ASVAB has 10 subscales altogether, plus four composite variables derived from the 10 subscales (technical knowledge, verbal ability, quantitative ability, and speed) and a measure of general intelligence. We would expect the technical knowledge to be correlated with cognitive disequilibrium and the detection of discrepancies, as pre-

dicted in our models of question asking. Verbal ability should tap the verbal-coding component (Stage 2) of Graesser and McMahan's (1993) three-stage model of question generation. Additional tests of cognitive ability were included because they are not directly captured in ASVAB but are frequently incorporated in individual-differences research on cognitive abilities.

A number of noncognitive variables were also measured because they tap the social editing and pragmatic components of our models of question asking. These included gender and scales on a personality test. The personality test was the NEO Five Factor Inventory (NEO; Costa & McCrae, 1991), which measures individuals on the "big five" personality factors: neuroticism, extroversion, openness, agreeableness, and conscientiousness. The NEO was designed to provide a general description of the personality of adults in clinical, counseling, and educational situations. The test is routinely adopted in research in social and personality psychology because it can be administered in only 45 min, has high reliability, and has been validated with other personality inventories and projective techniques. The internal consistency coefficients range from .86 to .95 for the five-factor scales; stability coefficients range from .51 to .83 for 3-year, 6-year, and 7-year longitudinal studies. The Openness Scale partially captures creativity, which we anticipated might be correlated with question asking.

In summary, the primary prediction of the PREG model is that deep comprehenders of the device should generate high-quality questions when confronted with a breakdown scenario. Thus, there should be a robust positive correlation between measures of question asking and (a) device-comprehension scores and (b) more generic measures of individual differences that are directly relevant to device comprehension, such as technical knowledge (i.e., mechanical comprehension, electronics, general science). Noncognitive variables might also predict question quality because pragmatics and social editing are among the components of the PREG model. When considering Graesser and McMahan's (1993) three-stage process of generating questions, one should tap discrepancy detection by ASVAB's technical knowledge scale and other measures of deep comprehension; verbal coding should be tapped by ASVAB's Verbal Ability scale; and social editing should be tapped by noncognitive variables.

### Method

#### *Participants*

The participants were 108 college students enrolled in an introductory psychology course at the University of Memphis. This sample included 41 men and 67 women. The participants volunteered to participate in exchange for extra credit in the course as well as educational benefits from learning about the research being conducted.

#### *Illustrated Texts and Breakdown Scenarios*

The participants read six illustrated texts on everyday devices: a cylinder lock, an electronic bell, a car temperature gauge, a clutch, a toaster, and a dishwasher. The device mechanisms were extracted from Macaulay's (1988) book with illustrated texts, *The Way Things Work*. The illustrated texts contained sections in printed text, visual diagrams of the components of the device, labels of major components, and directional arrows that convey motion or temporal changes. Two of the devices were selected to be stereotypically male (temperature gauge, clutch), two were stereotypi-

cally female (toaster, dishwasher), and two were gender neutral (cylinder lock, bell). This classification of devices on gender stereotypes was confirmed when we collected ratings from a sample of 10 college students in a normative pretest. The students rated the six devices on a 6-point scale that ranged from 1 (*definitely male*) to 6 (*definitely female*). The mean ratings were lowest for the temperature gauge (2.30) and clutch (2.20), highest for the toaster (4.60) and dishwasher (4.50), and intermediate for the cylinder lock (2.70) and bell (3.00). However, it is important to acknowledge that stereotypes do not necessarily reflect actual usage of devices by men versus women. Our purpose was merely to select devices to study that are not obviously skewed to a particular gender. Also, the fact that we are investigating the mechanisms of devices may be inherently biased against women because women allegedly show less interest than men in mechanics and electronics.

A breakdown scenario was prepared for each of the six devices. The breakdown scenario consisted of one or two sentences that identified physical symptoms of a device malfunction. For example, in the case of the cylinder lock, the breakdown scenario was "the key turns, but the bolt does not move." The breakdown scenarios were selected to fulfill one important criterion: The breakdown could be explained by a very small number of components, parts, events, or processes in the device system. In the case of the cylinder lock, for example, the fault would converge on the cam, parts of the cam, parts that interact with the cam, and events that move the cam. There are a host of other components, parts, events, and processes in Figure 1 that would not be plausible explanations of the breakdown.

### Experimental Tasks

Each participant ended up reading and receiving breakdown scenarios on all six devices. The trial for each device consisted of a study phase and a verbal-protocol phase. During the study phase, the participant read the illustrated text for 5 min. The illustrated text was presented on a page in a booklet. The experimenter signaled the participant when to start studying by saying "turn the page and start reading" and when to stop studying by saying "turn the page." When the participant turned the page, the breakdown scenario appeared on the page. At that point the participant provided verbal protocols by writing down content on the page, following instructions presented earlier in the booklet. During this time, the participants provided either think-aloud-in-writing protocols (which we call the write-aloud task) or generate-questions-in-writing protocols (question-asking task) for 3 min. The participants typically reflected on how to diagnose and repair the malfunctions during the write-aloud and question-asking tasks. The instructions in the question-asking task encouraged the participants to write down whatever questions came to mind in the context of the breakdown scenario. The instructions did not encourage them to think about the mental process of generating questions; rather the focus was on the question content.

The write-aloud task was completed for three devices prior to the question-asking task, which in turn was completed for the three remaining devices. The assignment of devices to conditions and test order was counterbalanced across 108 college students according to a  $6 \times 6$  Latin square. The participants completed the write-aloud task first because we did not want to bias them with the more directed question-asking task; writing aloud was supposed to be as unconstrained as possible, following the normal instructions to participants in studies that collect think-aloud protocols (Ericsson & Simon, 1980; Pressley & Afflerbach, 1995). Of course, it should be acknowledged that the question-asking task was more susceptible to practice or fatigue effects. However, our analyses of the data consisted of internal analyses within each of these tasks, namely, the correlations between the indices of the verbal protocols and the measures in our inventory of individual differences; our goal was not to directly compare the verbal output of write aloud versus question asking. We could have used an independent-groups design that had different participants assigned to the question-asking and write-aloud tasks. However, we de-

cidated on assigning all participants to both tasks in order to maximize the number of data points in the multiple regression analyses and to ensure that the profile of predictor variables was equivalent in the analyses of both tasks.

The booklets that the participants received for the experimental task had 14 pages altogether. There were 7 pages for the write-aloud task and 7 pages for the question-asking task. There was an instruction page for the write-aloud task, followed by three pairs of pages for the three devices. One page of each pair was the illustrated text. The other page was the breakdown scenario printed at the top of the page, followed by space for the participant to write down the verbal protocol. The participants were instructed that they should write down whatever comes to mind when they read the breakdown scenario and that they would have 3 min to complete the writing task. The 7 pages devoted to the question-asking task followed the same procedure. First, there was an instruction page that instructed the participant to write down as many questions as they could think of when they receive the breakdown scenario. This instruction page was followed by three pairs of pages corresponding to the three devices assigned to the question-asking task. The experimenter monitored the time course of the experiment by signaling when the participant was to turn pages. Thus, there was a fixed amount of time (5 min) to read each illustrated text and a fixed amount of time (3 min) to provide the written protocols. The entire experimental task lasted approximately 1 hr, after considering the time for participants to read the instructions and for the experimenter to answer any clarification questions about the nature of the tasks.

### Device-Comprehension Test

The participants were given an objective test on their understanding of the devices after they completed the write-aloud and question-asking protocols for all six devices. The device-comprehension test consisted of 6 three-alternative forced-choice (3AFC) questions about each device (36 total questions across the six devices). There were 4 test questions per device that tapped explicit information, and 2 questions that tapped inferences. Examples of such questions are as follows:

EXPLICIT: What action by a person causes the pins to rise?

The key is inserted. (correct answer)

The key is removed.

The key is turned.

INFERENCE: What happens to the pins when the key is turned to unlock the door?

They rise.

They drop.

They remain stationary. (correct answer)

The device-comprehension scores could vary from 0 to 36. A score of 12 would be chance performance if there were no sophisticated guessing or auxiliary background knowledge.

The device-comprehension test was defined as the primary gold standard for deep comprehension. The questions were generated systematically by adopting a theoretical foundation in qualitative physics (Forbus, 1984). Suppose there is a set of  $N$  "nodes" in a system that accounts for the device mechanism; a node refers to a physical component, a part of a component, a system state, an event, a process, or some other class of epistemological content (Baggett & Graesser, 1995; Graesser, Gordon, & Brainerd, 1992). The set of  $N$  nodes are connected by a network of negative, positive, and zero causal relations. If Node  $C$  is disturbed or changed in some fashion (e.g., broken, initiated, rotated, increased input), how would it propagate its effects on the other nodes in the system (e.g., Nodes  $X$ ,  $Y$ , and  $Z$ )? There were always 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.

A deep comprehender is able to trace the causal antecedents and causal consequences of an event (Graesser & Bertus, 1998; Hegarty & Just, 1993; Kieras & Bovair, 1984; Mayer, 1997), whereas poor comprehenders are undiscriminating in tracking the impact of one event on other events in the device system.

It should be noted that *conceptual graph structures* were prepared for each of the six devices, following an analytical scheme that captures both text and pictures in one common representation (Baggett & Graesser, 1995; Graesser et al., 1992). The conceptual graph structures were very helpful guides for generating the questions on the device-comprehension test. A conceptual graph structure contains a set of categorized *nodes* (concept, state, process, event, goal) that are interrelated by relational *arcs*. For everyday devices, the conceptual graph structures include the following: the components of the electronic or mechanical system, the spatial arrangement of components, the causal chain of events when the system successfully unfolds, the processes and enabling states that explain causal steps, and the goals or plans of agents who manipulate the system for various purposes. It is beyond the scope of this article to describe this representational system and report how the data collected in this study map onto the structures. However, we do want to point out the use of this detailed, systematic approach to analyzing device knowledge.

### Battery of Tests of Individual Differences

Following the objective test of device comprehension, participants completed a battery of tests that measured their cognitive abilities and personality. The primary test of cognitive ability was the ASVAB (Department of Defense, 1983). There were the following 10 subscales on this psychometric test: Mechanical Comprehension, Electronics, General Science, Auto and Shop, Mathematics Knowledge, Arithmetic Reasoning, Numerical Operations, Word Knowledge, Paragraph Comprehension, and Coding Speed. Composite variables are derived from the 10 measured variables on the ASVAB: technical knowledge, verbal ability, quantitative ability, speed, and general intelligence. Additional tests of cognitive ability included working-memory span (LaPointe & Engle, 1990), spatial reasoning (Bennet, Seashore, & Wesman, 1972), and exposure to print (the Author Recognition Test; Stanovich & Cunningham, 1992).

The noncognitive variables included age, gender, and five scales on a personality test. The personality test was the NEO (Costa & McCrae, 1991), which measures individuals on the big five personality factors: neuroticism, extroversion, openness, agreeableness, and conscientiousness. It took approximately 4 hr to complete the battery of tests, which were completed in two sessions on 2 different days.

### Scoring of Question-Asking and Write-Aloud Protocols

Four measures were scored on the verbal protocols that were collected in the question-asking and write-aloud tasks. These are listed and defined as follows.

*Volume of questions.* The number of questions that were asked in the question-answering task.

*Question quality.* The proportion of questions that referred to a plausible fault that explained the breakdown.

*Volume of ideas.* The number of ideas expressed in the write-aloud task.

*Idea quality.* The proportion of ideas that referred to a plausible fault explaining the breakdown.

These four scores were computed for each device that a particular participant completed. An average score over the three devices was then assigned to the participant for each of these four measures. We also computed measures of the volume of quality questions and the volume of quality ideas; however, these measures produced virtually identical results as question quality and idea quality, so we decided to report the latter proportion measures.

Three trained judges segmented the protocols into separate questions or idea units. There was a high reliability in such judgments (.93 or higher between any given pair of judges). A question or idea unit was counted if two out of three judges scored it as a unit. Trained judges also determined whether a question or idea matched a fault node. The judges were initially given a list of possible faults that might explain the breakdown for each device; these lists were provided by us. The judges were instructed to count a question or idea unit as high in quality if it matched one of the faults on the list. The judges based this judgment on semantic criteria (meaning, gist), rather than exact wording. This judgment required more training, but pairs of judges did eventually reach an acceptable level of agreement (80% or higher in common decisions).

## Results

### Descriptive Statistics

In Table 1, we present means and standard deviations for the measures that were collected in this study. This includes the ASVAB scores, spatial reasoning, working memory span, exposure to print, the personality measures (NEO), gender, age, measures of the verbal protocols, and device-comprehension scores.

The measures of individual differences presented no surprises when compared with available published studies and norms. As would be expected, the general intelligence scores and other ASVAB subscales for these lower division college students were

Table 1  
*Descriptive Statistics*

Variable	<i>M</i>	<i>SD</i>
Cognitive measures		
ASVAB (g)	125.3	19.4
Mechanical Reasoning	14.7	5.0
Electronics	11.4	4.0
General Science	18.6	4.3
Auto and Shop	12.3	5.5
Mathematics Knowledge	18.3	4.7
Arithmetic Reasoning	22.4	5.5
Numerical Operations	39.4	7.9
Word Knowledge	29.9	5.1
Paragraph Comprehension	12.5	3.3
Coding Speed	58.8	12.3
Spatial reasoning	27.3	14.4
Working-memory span	33.4	8.6
Exposure to print	9.1	6.9
Personality measures		
Neuroticism	49.9	12.3
Extroversion	52.7	12.3
Openness	51.8	11.6
Agreeableness	46.3	13.5
Conscientiousness	47.2	12.3
Demographics measures		
Gender (female = 1, male = 2)	1.4	0.5
Age	24.7	7.7
Verbal-protocol measures (per device)		
Volume of questions	3.8	2.0
Quality of questions <sup>a</sup>	22.4	14.7
Volume of ideas	5.4	1.8
Quality of ideas <sup>a</sup>	17.3	13.4
Device-comprehension score	23.5	5.3

*Note.* *N* = 108. ASVAB = Armed Services Vocational Aptitude Battery; g = measure of general intelligence.

<sup>a</sup> Numbers are in percentages.

above average, compared with the population of high school students who take the ASVAB. The other cognitive measures are not significantly different from the scores for college students that are reported in Bennet et al. (1972) for spatial reasoning, in LaPointe and Engle (1990) for working-memory span, and in Stanovich and Cunningham (1992) for exposure to print. It should be noted that the working-memory-span measure is not a measure of the number of chunks that can be held in working memory at any point in time (approximately 2–7 items). Instead, it is the number of items that were correctly recalled throughout several trials that had 2–7 items per trial; thus, the scores are considerably higher, with a mean of 33.4 items correctly recalled over several trials with different set sizes. The five personality subscales on the NEO (Costa & McCrae, 1991) were all hovering around the population mean of 50. There were more women (62%) than men in the sample, which is consistent with the estimates of college populations in the year 2000 (59% being women).

A number of observations can be made about the measures of the verbal protocols in the question-asking and write-aloud tasks. The mean was somewhat higher for the write-aloud task than the question-asking task, but the standard deviations were very close. The write-aloud task is less constrained than the question-asking task because the participants could write down whatever comes to mind. This added flexibility ended up producing more written units, as would be expected. However, the write-aloud task also was completed before the question-asking task, so assessment order is another possible explanation of the difference. Nevertheless, direct quantitative comparisons of measures between the write-aloud and question-asking tasks is not at all relevant to the present study. Instead, the relevant concern is the magnitude of the correlations between measures of the verbal protocols and (a) individual differences and (b) device-comprehension scores. The fact that the standard deviations are virtually the same for write-aloud and question-asking tasks allows us to meaningfully compare the profile of correlations associated with the two forms of verbal protocols.

The quality of questions and ideas was measured by computing the percentage of verbal units (questions or ideas) that matched one of the faults on the fault list that potentially explains the breakdown. Some examples of high-quality questions and ideas associated with the cylinder lock are “Is the cam broken?” “The cam may be disconnected from the cylinder,” and “Is the bar that fits under the cam broken?” According to the data in Table 1, 22% of the questions and 17% of the ideas were high in quality. Standard deviations did not significantly differ for the two tasks.

The primary criterion for measuring deep comprehension was the device-comprehension score. The mean score was 23.5 out of 36 questions, so 65% of the 3AFC questions were answered correctly. The questions that tapped explicit information in the illustrated texts were answered correctly more often than those that required inferences, 71% versus 54%, respectively. We do not distinguish these subclasses of comprehension questions in the remainder of the article, however, because the conclusions that could be drawn from separate measures were redundant with the conclusions derived from an overall measure of device comprehension.

## Correlations

Table 2 contains correlation coefficients that are relevant to assessments of individual differences that predict device-comprehension scores and the measures of verbal protocols. There also is a composite measure of general intelligence and the following four composite measures provided by ASVAB: (a) *Technical Knowledge*: Mechanical Reasoning, Electronics, General Science, and Auto and Shop; (b) *Verbal*: Word Knowledge, Paragraph Comprehension; (c) *Quantitative*: Mathematics Knowledge, Arithmetic Reasoning; (d) *Speed*: Numerical Operations, Coding Speed.

Consider first the prediction of device-comprehension scores, which are shown in the left column of numbers. Device-comprehension scores were significantly predicted by question quality, idea quality, spatial reasoning, exposure to print, the technical-knowledge composite (and each of its component measures), the quantitative composite (and each of its component measures), the verbal-composite measure (and each of its component measures), ASVAB general intelligence, gender, and openness. The other measures were nonsignificant predictors of device-comprehension scores ( $r < .20$ ): volume of questions, volume of ideas, age, most personality measures (neuroticism, extroversion, openness, agreeableness, and conscientiousness), working memory, and the speed-composite measure (and each of its component measures).

Question quality was a robust predictor of device-comprehension scores, on par with the psychometric measures that were expected to predict device comprehension. Indeed, the .51 correlation between question quality and device-comprehension scores was on par with spatial reasoning (.54) and the noncomposite measures of ASVAB that measured technical knowledge (which varied from .52 to .63). A set of partial correlations was computed between question quality and comprehension scores, holding each one of the other particular measures of individual differences constant. All 25 of these partial correlations were statistically significant in two-tailed tests ( $.25 < r < .51$ ), with one exception. The exception was the partial correlation between question quality and comprehension scores while holding technical knowledge constant ( $r = .18$ ); this partial correlation was significant in a one-tailed test but not in a two-tailed test.

It was the quality of the questions that predicted device comprehension, not the quantity. The correlation was nonsignificant ( $r = -.01$ ) between device-comprehension scores and the volume of questions. Similarly, the correlation was nonsignificant between device-comprehension scores and the volume of ideas in the write-aloud task ( $r = .08$ ). These results support the general conclusion that the amount of information produced in the written verbal protocols was not significantly correlated with deep comprehension of the devices.

The quality of ideas in the write-aloud task did significantly correlate with device-comprehension scores ( $r = .39$ ), but this correlation was not quite as high as the .51 correlation between the comprehension scores and question quality. This difference in predicting device-comprehension scores was significant in a one-tailed test but not in a two-tailed test. More generally, we found the quality of ideas in the write-aloud task was not as highly correlated with measures of individual differences as was the quality of questions in the question-asking task. This is apparent when we

Table 2  
*Correlations Among Measures of Individual Differences, Device-Comprehension Scores, and Verbal Protocols*

Variable	Device-comprehension score	Question asking		Write aloud	
		Quality	Volume	Quality	Volume
Cognitive measures					
ASVAB (g)	.59	.41	.05	.18	.12
Mechanical Reasoning	.63	.56	-.11	.30	.09
Electronics	.56	.52	-.01	.36	-.02
General Science	.60	.48	-.10	.24	.05
Auto and Shop	.52	.40	-.05	.32	-.05
Mathematics Knowledge	.56	.45	-.04	.18	.12
Arithmetic Reasoning	.52	.37	-.05	.11	.04
Numerical Operations	.12	.10	.13	.10	-.05
Word Knowledge	.42	.29	.00	.09	.10
Paragraph Comprehension	.27	.11	-.02	.21	.08
Coding Speed	.08	-.08	.21	.07	.02
Technical Knowledge Composite	.72	.55	-.08	.43	-.02
Verbal Composite	.49	.28	-.01	.26	.10
Mathematical Composite	.59	.44	-.02	.23	.08
Coding and Speed Composite	.09	.03	.20	.06	-.04
Spatial reasoning	.54	.44	.00	.22	.26
Working memory	.08	.17	.05	-.08	.12
Exposure to print	.25	.18	-.11	-.04	.10
Personality measures					
Neuroticism	-.08	.04	-.07	-.07	.01
Extroversion	-.01	-.03	.06	-.03	.02
Openness	.31	.21	-.09	.08	.03
Agreeableness	-.03	-.12	-.02	-.04	-.18
Conscientiousness	-.01	-.04	.18	.01	.13
Demographics measures					
Gender (female = 1, male = 2)	.41	.33	-.28	.30	-.24
Age	.01	.01	-.01	-.08	-.19
Verbal-protocol measures					
Quality of questions	.51				
Volume of questions	-.01	-.34			
Quality of ideas	.39	.35	-.08		
Volume of ideas	.08	-.14	.46	-.34	

Note. ASVAB = Armed Services Vocational Aptitude Battery; g = measure of general intelligence.

compared the columns in Table 2. The pattern of correlations in the second column in Table 2 (the column for question quality) was extremely similar to the first column (the column for device-comprehension scores). In contrast, the correlations in columns 3 (volume of questions), 4 (quality of ideas), and 5 (volume of ideas) were substantially different from column 1. Of the 25 measures of individual differences (not including the measures of verbal protocols), 16 measures significantly correlated with device-comprehension scores. For these 16 measures, we compared correlations associated with the quality of the questions (column 2) and the quality of the ideas (column 4); 15 of these correlations were higher in magnitude for quality of questions than quality of the ideas, which is significant according to a Wilcoxon sign test. Therefore, question quality was a more robust predictor than idea quality of both (a) the comprehension scores and (b) the other measures of individual differences that are significantly correlated with the comprehension scores. The quality of questions asked in the context of a breakdown scenario was an excellent quick litmus test of deep comprehension and had an acceptable level of criterion validity.

There was a modest correlation between the quality of questions and ideas, and also between the volume of questions and ideas. In

contrast, there was a modest negative correlation between these volume measures and the two quality measures. Thus, those participants who generated more content tended to produce a lower percentage of quality content. There were several potential interpretations of this result. Perhaps some students were extremely compliant in producing a large amount of content, even after the high-quality content was tapped out. Alternatively, perhaps deep comprehenders were more succinct and discriminating. The correlations in Table 2 cannot discriminate between these alternatives.

Our operational definition of a good question was whether it matched a plausible fault. We found this measure to be superior to an alternative operational definition that was based on the syntactic or semantic form of the question. According to the question taxonomy of Graesser and Person (1994), deep questions ("why," "how," "what if," "what if not") tap causal, goal-oriented, and logical reasoning so they are higher quality than the more shallow short-answer questions ("who," "what," "when," "where," "how much"). Graesser and Person's scale of question quality correlated significantly with Bloom's (1956) taxonomy of cognitive objectives. This definition of question quality did not significantly correlate with device comprehension ( $r = .11$ ). Therefore, a good question was best defined according to whether it referred

to a plausible fault, not whether it was worded in a sophisticated form.

The fact that the technical-knowledge and spatial-reasoning measures were among the better predictors of device comprehension was quite expected and indirectly confirms the construct validity of the device-comprehension measure. However, several of the measures of ASVAB were intercorrelated, so additional analyses were needed to tease apart the contributions of these processes. We did so in the multiple regression analyses reported subsequently.

Most of the noncognitive measures did not significantly correlate with device comprehension and question quality. Openness was the only significant personality measure to have significant correlations. This result can perhaps be attributed to the creativity component that is linked to the openness measure. Men had deeper comprehension of devices than did the women. One interpretation of this outcome was that it reflected gender stereotypes in the United States. There was another interpretation, however. Women produced a greater volume of verbal content perhaps because they were more compliant or had higher verbal aptitude. Women may have run out of the quality questions and ideas but continued producing content. The additional content would have diluted (lowered) the quality measures, which consisted of proportion scores. This possibility also was assessed in the subsequent multiple regression analyses.

### Multiple Regression Analyses

We performed some multiple regression analyses to dissect the contributions of the various cognitive and noncognitive components discussed previously. Five multiple regression analyses were conducted, one for each of the following five dependent measures: device-comprehension scores, question quality, volume of questions, idea quality, and volume of ideas. There were nine predictor variables in each of these analyses: the four ASVAB composite variables, spatial reasoning, working memory, exposure to print, openness, and gender. All nine predictors were included in the regression analysis, with forced entry rather than a hierarchical

design. These predictors were included because they either had a significant correlation with one or more of the five dependent measures or they were a cognitive measure. The results of the multiple regression analyses are in Table 3.

Technical knowledge was the primary predictor variable in the multiple regression analyses for the device-comprehension scores. The multiple regression equation with nine predictors accounted for 57% of the variance; technical knowledge alone accounted for 52% of the variance (i.e.,  $.72^2 = .52$ ). None of the other eight predictors were significant. We also assessed interactions between pairs of predictor variables, but these were rarely significant (less than 5%, readily attributable to a Type I error). So it was technical knowledge that reigned supreme in predicting device-comprehension scores, an outcome that confirmed the construct validity of our primary criterion measure for deep comprehension.

The multiple regression analysis for question quality perfectly mirrored the results of the device-comprehension scores. The multiple regression equation with nine predictors accounted for 37% of the variance. Technical knowledge alone accounted for 26% of the variance (i.e.,  $.51^2 = .26$ ), and none of the other eight predictors were significant. Once again, pairwise interaction components also were rarely significant. Question quality was an excellent index of deep comprehension and had a satisfactory degree of construct validity.

The remaining three measures of the verbal protocols did not fare as well. The volume of questions was not significantly predicted by the nine variables. The volume of ideas in the write-aloud task was significantly predicted by spatial reasoning and gender, with women articulating more information; however, the amount of variance explained was a modest 19%. The quality of ideas was significantly predicted by technical knowledge, speed, and exposure to print (the latter in the opposite direction, a likely suppression effect); 26% of the variance was predicted. The fact that the quality of ideas was significantly predicted by technical knowledge suggests that the write-aloud task is another way to find out whether a person has deep knowledge about a device. However, the measures of individual differences accounted for signif-

Table 3  
*Beta Weights in Multiple Regression Analyses*

Variable	Device-comprehension Score	Question asking		Write aloud	
		Quality	Volume	Quality	Volume
Cognitive measures					
ASVAB					
Technical Knowledge	.51*	.37*	.24	.47*	.15
Verbal	.10	-.16	.01	.19	.00
Quantitative	.14	.13	-.02	-.22	.03
Speed	.06	.01	.09	.20*	.17
Spatial reasoning	.11	.15	.00	-.01	.39*
Working memory	-.05	-.02	.00	-.11	.08
Exposure to print	-.12	.16	-.17	-.27*	.13
Other measures					
Openness	-.02	-.06	-.08	-.08	-.03
Gender (female = 1, male = 2)	.11	.13	-.37*	.19	-.29*
Variance predicted ( $R^2$ )	.57*	.37*	.13	.26*	.19*

Note. ASVAB = Armed Services Vocational Aptitude Battery.  
\*  $p < .05$ .

icantly more of the variance of the question-asking task (37%) than the write-aloud task (26%).

We performed a factor analysis that included the four variables that comprised technical knowledge and the remaining three cognitive measures (spatial reasoning, working-memory span, and exposure to print). The results revealed that there was only one significant underlying factor, with an eigenvalue of 3.76 (chance being 1.38). The four ASVAB measures loaded most heavily on the factor, with the highest being technical knowledge. We interpreted this underlying factor to be deep comprehension of technical knowledge. Device-comprehension scores and question quality were believed to be two excellent correlates of this factor.

### *Phases of Output in the Question-Asking and Write-Aloud Tasks*

Perhaps the initial questions or ideas produced in the question-asking and write-aloud tasks were systematically different from subsequent content. For example, the major faults of the device breakdowns may have been identified quickly in the initial output, whereas the subsequent output may have had lower quality content that filled up the 3 min of the verbal protocol. Perhaps the more persistent participants were more diligent in asking additional questions after the initial insights were expressed; this may have applied to the women, who had a higher volume of output but a lower proportion of quality output, as reported earlier. In order to assess these possibilities, we segregated each protocol into the initial output versus the later output. The initial output consisted of the first 2 questions or thoughts expressed in the protocols. Later output consisted of Questions-thoughts 3 through N.

We found support for the notion that the initial output was higher quality. The quality of the questions was .28 versus .18 for initial versus later questions, respectively. The quality of the ideas in the write-aloud task was .22 and .08, respectively. We performed four multiple regression analyses that assessed the extent to which these four quality scores could be predicted by the nine measures of individual differences that were reported in Table 3. Question quality for the initial questions was significantly predicted by the multiple regression equation with nine predictors,  $F(9, 98) = 3.59, p < .05, R^2 = .25$ ; technical knowledge was the only significant predictor ( $\beta = .39, p < .05$ ). Idea quality for the initial ideas expressed in the write-aloud tasks had exactly the same result, although it was not quite as robust in variance predicted. The quality of the initial ideas was significantly predicted by the multiple regression equation with nine predictors,  $F(9, 98) = 2.18, p < .05, R^2 = .16$ , with technical knowledge being the only significant predictor ( $\beta = .47, p < .05$ ). These results matched well the analyses of all questions (see Table 3).

In contrast to the initial output, the later output was less predictable or was explained by different measures of individual differences. Question quality for the later questions was significantly predicted by the multiple regression equation with nine predictors,  $F(9, 98) = 3.54, p < .05, R^2 = .25$ , but this time the significant predictors were spatial reasoning and working memory ( $\beta_s = .25, p_s < .05$ ). The visual-spatial sketchpad component of working memory (Baddeley, 1986) may play a more prominent role in the question-generation mechanism during the later stages of output. The quality of the subsequent ideas in write-aloud was not significantly predicted by the multiple regression equation with

nine predictors,  $F(9, 98) = 1.39, p > .20, R^2 = .11$ , and none of the individual predictors were significant.

Gender was not a significant predictor in any of the multiple regression analyses on question or idea quality. Bivariate correlations on the total set of questions or ideas would lead one to believe that women had a higher volume of output, with the output being of lower quality than men. However, the multiple regression analyses revealed that women were more compliant and generated more questions; that reduced the overall quality scores, as measures by the proportion of output that was high quality (i.e., referring to faults that potentially explained breakdowns). Gender did not predict question-idea quality over and above technical knowledge. So any differences between men and women can be attributed to the expected gender differences that are embodied in technical knowledge.

### Discussion

We confirmed the primary prediction of the PREG model of question asking (Graesser et al., in press; Otero & Graesser, 2001). Specifically, deep comprehenders of a device ask good questions when they are confronted with a breakdown scenario. Good questions are those that tap plausible faults that explain the symptoms of the breakdown. In contrast, shallow comprehenders are not as discriminating when they ask questions; their questions are not as likely to converge on plausible faults. We found a robust correlation between question quality and both device-comprehension scores and generic technical knowledge (as measured by ASVAB). In contrast, the mere volume of questions was not a significant reflection of deep comprehension. Moreover, question asking produced content that was more diagnostic of deep comprehension than did the write-aloud task.

There is some other evidence that it is question quality, not number of questions, that is diagnostic of deep comprehension. Fishbein, Eckart, Lauver, Van Leeuwen, and Langmeyer (1990) investigated the questions asked by students in a tutoring session. They reported that question quality was positively correlated with subject-matter knowledge, whereas number of questions was not significantly positively correlated. In an analysis of student questions in tutoring, Graesser and Person (1994) found that question quality (i.e., deep reasoning questions corresponding to "why," "why not," "how," "what if," and "what if not") was positively correlated with exam scores, whereas the quantity of questions yielded a significant negative correlation. These studies of naturalistic tutoring did not specifically focus on conditions of cognitive disequilibrium, however, as we did in the present study. That we obtained the same results in the present study as in the previous analyses of tutoring permits the general conclusion that is the quality of questions, not the quantity of questions, that reflects deep comprehension.

Graesser et al. (in press) reported some qualitative analyses of a subset of the questions in the present study. They compared the distributions of questions that were asked by 11 students with high mechanical comprehension scores versus 11 students with low scores. Presented following are some questions that were asked by students with high versus low scores; the numbers in parentheses refer to the number of participants out of 11 who generated the questions, with high scorers preceding low scorers.

Is the cam broken? (6, 2)

- Is the cam moving? (2, 2)
- Is the cam moving back the bolt? (2, 0)
- Is the bar that fits under cam broken? (2, 0)
- Is the cam disconnected or out of synch with the cylinder? (2, 0)
- What kind of lock is it? (0, 3)
- What are the pins used for? (0, 2)
- Is the bolt stuck in the slot? (3, 3)

Three points can be illustrated with these examples. First, students with high technical knowledge identified the fault in the region of the cam more often than the low technical students. In contrast, students with low technical knowledge were more prone to ask about components that cannot explain the breakdown (the pins). Stated differently, the questions of deep comprehenders had a convergence on likely faults, whereas the questions of shallow comprehenders were diffuse. Second, the questions asked by students with high scores had a more fine-grained elaboration of the parts, processes, and relations that specify how the breakdown occurred; there was more mechanistic detail. Third, the students with low scores had more questions that were shallow and irrelevant (e.g., "What kind of lock is it?").

In a recent study, Graesser, Olde, and Lu (2001) collected eye-tracking data while college students generated questions about the breakdown scenarios associated with the devices. The participants first studied the illustrated text for a device for 3 min and then were presented the breakdown sentences beside the illustrated text for 90 s. The participants were instructed to generate questions aloud during the 90-s period. Eye fixations and eye movements were recorded by an eye tracker with a head-mounted recording unit; this allowed the participants to ask questions and move their head freely. According to the PREG model, there should be a high density of eye fixations on words, objects, parts, and processes that are at the source of cognitive disequilibrium (e.g., anomalies, contradictions, broken parts, contrasts, missing components, etc.). The results of the eye-tracking study were consistent with several predictions of PREG. The eye fixations indeed drifted toward the faults that explain the breakdown. However, the readers needed to have a sufficient amount of technical knowledge to detect such irregularities in the system. Students with high technical knowledge focused on the causes of the device breakdown (the cam), whereas students with low technical knowledge indiscriminately scanned the various regions of the illustrated text. Technical knowledge, device-comprehension scores, and question quality were all positively correlated with the percentage of fixations and the percentage of time that the comprehender focused on the fault area. The fact that question quality was once again a significant index of deep comprehension replicates the present study.

There are more fine-grained predictions of PREG that call for further research. In particular, the model generates predictions about the particular questions that should be asked under particular conditions when expository texts are read. One of the long-term goals of the PREG model is to serve as a computational model, in which an expository text is input to a computer and the output is a family of particular questions that are likely to be asked by

humans with different levels of knowledge, skill, and abilities. The predicted questions are sensitive to (a) the explicit text (words, propositions, discourse constituents), (b) the reader's world knowledge about the topics in the text, (c) the reader's metacognitive skills, and (d) the reader's knowledge about the pragmatics of communication. Questions are triggered when there are discrepancies between text and world knowledge, when there are discrepancies within the text, and when there are discrepancies within world knowledge. The hope is that the PREG model has a sufficient degree of complexity and analytical precision that it will provide a host of testable predictions for different theories of inquiry learning.

It is conceivable that other variables would have predicted question quality other than technical knowledge. Graesser and McMahan's (1993) three-stage mechanism of question asking includes disequilibrium detection, verbal coding, and social editing. Whereas technical knowledge tapped the disequilibrium detection stage, the verbal-ability scores presumably tapped the verbal-coding stage and the noncognitive variables tapped social editing. We found, however, that the verbal-ability scores and personality scores did not discriminatively predict question quality, over and above technical knowledge. The only noncognitive factor that predicted verbal protocols was the finding that female students produced a higher volume of questions in the question-asking tasks and ideas in the write-aloud task. The present study investigated question asking and think aloud under task-induced conditions, where participants were directly instructed to ask questions or think aloud. This was designed to remove potential barriers of social editing and pragmatic constraints and to optimize conditions that trigger questions. Our attempt to remove the constraints of social editing and pragmatics was apparently quite successful. We suspect that the noncognitive variables would play a major role in self-induced conditions, when social editing and pragmatics are more prominent. Nearly everyone has had the experience of wanting to ask a question but turning down the opportunity in order to avoid embarrassment or ridicule. Social editing and pragmatics explain, to a large extent, why the incidence of student questions are rare in classroom environments (Dillon, 1988; Good et al., 1987; Graesser & Person, 1994). A more systematic investigation of these social and pragmatic components on question asking is one important direction for future research.

## References

- Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.
- Baggett, W. B., & Graesser, A. C. (1995). Question answering in the context of illustrated expository text. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 334–339). Hillsdale, NJ: Erlbaum.
- 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, & G. T. Waller (Eds.), *Metacognition, cognition, and human performance* (pp. 155–205). New York: Academic Press.
- Beck, I. L., McKeown, M. G., Hamilton, R. L., & Kucan, L. (1997). *Questioning the author: An approach for enhancing student engagement with text*. Newark, DE: International Reading Association.
- Bennet, G. K., Seashore, H. G., & Wesman, A. G. (1972). *Differential Aptitude Test: Spatial Relations, Form T*. New York: Psychological Corporation.

- Berlyne, D. E. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill.
- Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook I. Cognitive domain*. New York: McKay.
- Bransford, J. D., Goldman, S. R., & Vye, N. J. (1991). Making a difference in people's ability to think: Reflections on a decade of work and some hopes for the future. In R. J. Sternberg & L. Okagaki (Eds.), *Influences on children* (pp. 147–180). Hillsdale, NJ: Erlbaum.
- Brown, A. L. (1988). Motivation to learn and understand: On taking charge of one's own learning. *Cognition and Instruction*, 5, 311–321.
- Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.
- 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.
- Ciardiello, A. V. (1998). Did you ask a good question today? Alternative cognitive and metacognitive strategies. *Journal of Adolescent & Adult Literacy*, 42, 210–219.
- Collins, A. (1988). Different goals of inquiry teaching. *Questioning Exchange*, 2, 39–45.
- Costa, P. T., & McCrae, R. R. (1991). *NEO: Five Factor Inventory*. Odessa, FL: Psychological Assessment Resources.
- Coté, N., Goldman, S., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, 25, 1–53.
- 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.
- Davey, B., & McBride, S. (1986). Effects of question generation on reading comprehension. *Journal of Educational Psychology*, 78, 256–262.
- Department of Defense. (1983). *Armed Services Vocational Aptitude Battery, Form 12a*. Washington, DC: Author.
- Dillon, T. J. (1988). *Questioning and teaching: A manual of practice*. New York: Teachers College Press.
- 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.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–251.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.
- Fishbein, H. D., Eckart, T., Lauver, E., Van Leeuwen, R., & Langmeyer, D. (1990). Learners' questions and comprehension in a tutoring setting. *Journal of Educational Psychology*, 82, 163–170.
- Flammer, A. (1981). Towards a theory of question asking. *Psychological Research*, 43, 407–420.
- Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence*, 24, 85–168.
- 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., Baggett, W., & Williams, K. (1996). Question-driven explanatory reasoning. *Applied Cognitive Psychology*, 10, S17–S32.
- Graesser, A. C., & Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading*, 2, 247–269.
- 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., & McMahan, C. L. (1993). Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. *Journal of Educational Psychology*, 85, 136–151.
- Graesser, A. C., Olde, B., & Lu, S. (2001). Question-driven explanatory reasoning about devices that malfunction. In T. Filjak (Ed.), *Proceedings of the 36th International Applied Military Psychology Symposium* (pp. 114–119). Zagreb, Croatia: Ministry of Defense of the Republic of Croatia.
- Graesser, A. C., Olde, B., Pomeroy, V., Whitten, S., Lu, S., & Craig, S. (in press). Inferences and questions in science text comprehension. In J. Otero & M. Helena (Eds.), *Comprension de los libros de texto de ciencias* [Science text comprehension]. Madrid, Spain: El Paidós.
- Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, 31, 104–137.
- Graesser, A. C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22, 39–51.
- Guthrie, J., & McCann, D. A. (1997). Characteristics of classrooms that promote motivations and strategies for learning. In J. Guthrie & A. Wigfield (Eds.), *Reading engagement: Motivating readers through integrated instruction* (pp. 128–148). Newark, DE: International Reading Association.
- Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (1998). *Metacognition in educational theory and practice*. Mahwah, NJ: Erlbaum.
- Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language*, 32, 717–742.
- Kerry, T. (1987). Classroom questions in England. *Questioning Exchange*, 1, 32–33.
- Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, 8, 255–274.
- King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology*, 14, 366–381.
- King, A. (1992). Comparison of self-questioning, summarizing, and note-taking 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.
- LaPointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: General*, 64, 1118–1133.
- Macaulay, D. (1988). *The way things work*. Boston: Houghton Mifflin.
- Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist*, 32, 1–19.
- Miyake, N., & Norman, D. A. (1979). To ask a question one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior*, 18, 357–364.
- Otero, J., & Graesser, A. C. (2001). PREG: Elements of a model of question asking. *Cognition and Instruction*, 19, 143–175.
- Palincsar, A. S., & Brown, A. (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.
- Piaget, J. (1952). *The origins of intelligence*. Madison, CT: International Universities Press.
- Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Mahwah, 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.
- Pressley, M., Symons, S., McDaniel, M. A., Snyder, B. L., & Turner, J. E.

- (1988). Elaborative interrogation facilitates in the acquisition of confusing facts. *Journal of Educational Psychology*, 80, 301–342.
- 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.
- Scardamalia, M., & Bereiter, C. (1985). Fostering the development of self-regulation in children's knowledge processing. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills* (Vol. 2, pp. 563–577). Hillsdale, NJ: Erlbaum.
- Schank, R. C. (1999). *Dynamic memory revisited*. Cambridge, England: Cambridge University Press.
- Singer, H., & Donlan, D. (1982). Active comprehension: Problem-solving schema with question generation for comprehension of complex stories. *Reading Research Quarterly*, 17, 166–186.
- Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition*, 20, 51–68.
- van der Meij, H. (1994). Student questioning: A componential analysis. *Learning and Individual Differences*, 6, 137–161.
- VanLehn, K., Jones, R. M., & Chi, M. T. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, 2, 1–59.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81, 329–339.

Received August 28, 2001

Revision received June 22, 2002

Accepted July 15, 2002 ■

### Call for Nominations

The Publications and Communications (P&C) Board has opened nominations for the editorships of *Comparative Psychology*, *Experimental and Clinical Psychopharmacology*, *Journal of Abnormal Psychology*, *Journal of Counseling Psychology*, and *JEP: Human Perception and Performance* for the years 2006–2011. Meredith J. West, PhD, Warren K. Bickel, PhD, Timothy B. Baker, PhD, Jo-Ida C. Hansen, PhD, and David A. Rosenbaum, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2005 to prepare for issues published in 2006. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations also are encouraged.

Search chairs have been appointed as follows:

- *Comparative Psychology*, Joseph J. Campos, PhD
- *Experimental and Clinical Psychopharmacology*, Linda P. Spear, PhD
- *Journal of Abnormal Psychology*, Mark Appelbaum, PhD, and David C. Funder, PhD
- *Journal of Counseling Psychology*, Susan H. McDaniel, PhD, and William C. Howell, PhD
- *JEP: Human Perception and Performance*, Randi C. Martin, PhD

To nominate candidates, prepare a statement of one page or less in support of each candidate. Address all nominations to the appropriate search committee at the following address:

Karen Sellman, P&C Board Search Liaison  
 Room 2004  
 American Psychological Association  
 750 First Street, NE  
 Washington, DC 20002-4242

The first review of nominations will begin December 8, 2003. The deadline for accepting nominations is **December 15, 2003**.