

Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down

ARTHUR C. GRAESSER

University of Memphis, Memphis, Tennessee

SHULAN LU

Texas A&M University, Commerce, Texas

BRENT A. OLDE

Naval Aerospace Medical Institute, Pensacola, Florida

ELISA COOPER-PYE

University of East London, London, England

and

SHANNON WHITTEN

University of Central Florida, Orlando, Florida

The PREG model of question asking assumes that questions emerge when there is cognitive disequilibrium, as in the case of contradictions, obstacles, and anomalies. Participants read illustrated texts about everyday devices (e.g., a cylinder lock) and then were placed in cognitive disequilibrium through a breakdown scenario (e.g., *the key turns but the bolt does not move*). The participants asked questions when given the breakdown scenario, and an eyetracker recorded their fixations. As was predicted, deep comprehenders asked better questions and fixated on device components that explained the malfunction. The eye fixations were examined before, during, and after the participants' questions in order to trace the occurrence and timing of convergence on faults, causal reasoning, and other cognitive processes.

Some individuals believe that they understand the inner workings of a mechanical device, such as a car or a computer. However, it is often not until a device breaks down or the mechanisms need to be explained that individuals realize how little they know about the inner mechanisms. Everyday knowledge about devices consists primarily of perceptual knowledge of the outer surfaces (i.e., what the device looks like) and the procedures for starting, monitoring, and terminating the device (i.e., how to use it), but

not the mental models of the device mechanisms (how it works internally; Graesser & Clark, 1985; Kieras & Bovair, 1984; Rozenblit & Keil, 2002). There is a need for mental models of the device mechanisms when devices break down, but rarely when the devices are functioning properly. Deeper mental models are needed to diagnose malfunctions and discover methods of repairing the devices.

Graesser and Olde (2003; Graesser et al., 2005) have recently reported that the depth of a person's comprehension of a device is manifested by the questions that he or she asks when the device breaks down. Deep comprehenders ask good questions about the likely faults of the breakdowns. College students read illustrated texts on everyday devices that were extracted from David Macaulay's (1988) *The Way Things Work*. After reading each illustrated text, they were given a breakdown scenario and were instructed to ask questions for 3 min. For example, consider the cylinder lock depicted in Figure 1 and the following breakdown scenario: *The key turns, but the bolt does not move*. In this breakdown scenario, the person moves the key, and it has no trouble turning, but the bolt does not move back and forth. The students' questions varied in quality dur-

This research was supported by grants awarded to the first author by the Office of Naval Research (N00014-98-1-0331, N00014-01-1-0917, and N00014-00-1-0600) and the National Science Foundation (SBR-0126265, REC 0106965, and ITR 0325428). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Office of Naval Research. We thank Xiaoshan Cai, Scotty Craig, Derek Harter, Victoria Pomeroy, Amy Vitale, and James Wallace for their help at various stages of this project and Mary Hegarty, Timothy Hodgson, Michael Spivey, and an anonymous reviewer for their helpful reviews of earlier manuscripts. Correspondence concerning this article should be addressed to A. C. Graesser, Department of Psychology, University of Memphis, 202 Psychology Building, Memphis, TN 38152-3230 (e-mail: a-graesser@memphis.edu).

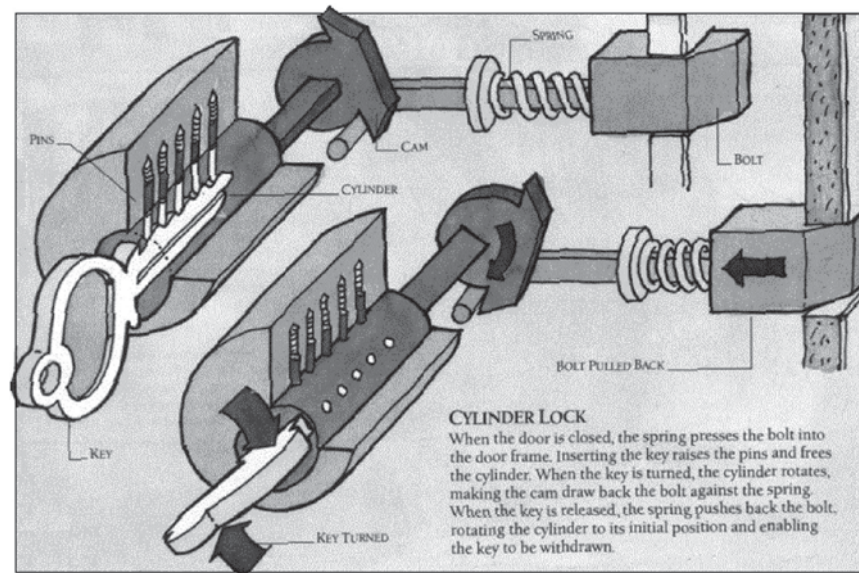


Figure 1. Example illustrated text describing a cylinder lock. From *The Way Things Work* (p. 17), by David Macaulay, 1988, Boston: Houghton Mifflin. Compilation copyright Dorling Kindersley Ltd., London. Illustration copyright 1988 David Macaulay. Text copyright David Macaulay, Neil Ardley. Reprinted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

ing the question-asking phase of the experiment. A good question was defined as any question that converged on the likely faults of the breakdown. A question about the *cam* would be considered a good question, for example, because it identified a likely fault that would explain the unmoving bolt (e.g., the cam does not rotate, the lip on the cam does not pull the rod). There are many potential bad questions that would not explain the device malfunction, such as questions about the pins rising, use of the right key, or a broken spring; these would be ruled out as plausible faults on the basis of knowledge about mechanical systems. Graesser and Olde reported that deep comprehenders did not ask a larger number of questions but that they did generate a higher proportion of good questions that pinpointed plausible faults. Depth of comprehension was measured in a number of ways, including scores on a subsequent device comprehension test and scores on psychometric tests of technical knowledge that tap mechanical reasoning, knowledge about electronics and general science, and practical knowledge about automobile repair and shop.

Why should questions be diagnostic of deep comprehension when devices break down? According to the PREG model of question asking (*Preg* is the first four letters of *pregunta*, the word for *question* in the Spanish language; Graesser & Olde, 2003; Graesser et al., 2005; Otero & Graesser, 2001), genuine information-seeking questions are very likely to be asked when individuals experience *cognitive disequilibrium*. The cognitive system is in disequilibrium when individuals are confronted with stimuli, problems, or situations that present obstacles to goals, anomalous events, contradictions, discrepancies,

expectation violations, and obvious gaps in knowledge. The mental state or process of cognitive disequilibrium drives questions and inquiry, whereas ensuing answers are expected to restore equilibrium. For example, cognitive disequilibrium typically occurs during the process of solving a novel or difficult problem, with disequilibrium not being restored until a satisfactory solution emerges.

The notion that cognitive disequilibrium drives questions and inquiry is consistent with a number of theories in cognitive, developmental, social, and educational psychology (Berlyne, 1960; Chinn & Brewer, 1993; Collins, 1988; Festinger, 1957; Flammer, 1981; Graesser, Baggett, & Williams, 1996; Graesser & McMahan, 1993; Graesser & Person, 1994; Miyake & Norman, 1979; Piaget, 1952; Schank, 1999). Graesser and McMahan reported that an increase in the number of questions (particularly the good questions) was caused directly by the cognitive disequilibrium that was triggered by transformations of mathematical word problems and of simple stories through contradictions, irrelevant information, and deletions of critical information. Otero and Graesser's (2001) PREG model differs from the other theories of question asking in its attempt to predict specific questions that will be asked under particular conditions. That is, the model attempts to specify precisely the content features of a text and the knowledge states of individuals that exist when particular questions are asked. For example, definitional questions (*what does X mean?*) are asked when a text contains a rare word. Causal antecedent questions (*why did X occur?*) are asked when the text has an Event X with no obvious cause or when some expected Event X fails to occur (*why didn't X occur?*). Readers generate chains and networks

of potential causal antecedents to X during the course of answering such questions. They also trace causal consequences that violate the expectations generated earlier during comprehension.

In the present study, we investigated eye-tracking behavior when college students were confronted with breakdown scenarios and when they asked questions. Eye-tracking behavior should help us (1) identify the cognitive processes that occur before, during, and after the questions about the faults and (2) test some predictions of the PREG model of question asking. Our focus on illustrated texts about devices allows us to relate our findings to other investigations of device comprehension and reasoning (Hegarty, 1992; Hegarty & Just, 1993; Hegarty, Narayanan, & Freitas, 2002; Mayer & Sims, 1994; Rozenblit, Spivey, & Wojaslawowicz, 2002) and, thus, refine the PREG model with respect to the underlying cognitive mechanisms.

We adopted a number of working assumptions about eye movements in the present study. One assumption was that there is a sufficiently close connection between display item viewed and the content being thought about, as well as between time spent fixating on display items and the amount of cognitive processing (Anderson, Bothell, & Douglass, 2004; Just & Carpenter, 1980; Rayner, 1998). A second assumption was that eye movements provide an important window for dissecting cognitive processes and representations with cognitive tasks that require deeper levels of processing. Eye-tracking analyses have, indeed, provided an illuminating method of investigating problem solving (Grant & Spivey, 2003; Hodgson, Bajwa, Owen, & Kennard, 2000; Knoblich, Ohlsson, & Raney, 2001), reasoning (Just & Carpenter, 1992), the evaluation of arguments (Wiley, 2001), and deeper levels of comprehension for different classes of text, such as narrative and expository text (Just & Carpenter, 1980; Magliano, Graesser, Eymard, Haberlandt, & Gholsen, 1993; O'Brien, Raney, Albrecht, & Rayner, 1997), advertisements (Fox, Krugman, Fletcher, & Fischer, 1998; Rayner, Rotello, Stewart, Keir, & Duffy, 2001), and illustrated texts on mechanical systems (Hegarty & Just, 1993; Rozenblit et al., 2002). However, eye-tracking research has not been conducted with tasks that help us understand the relationship between cognitive disequilibrium and question asking.

There are a number of hypotheses that could be derived directly from the PREG model of question asking (Graesser & Olde, 2003; Graesser et al., 2005; Otero & Graesser, 2001). Three of the predictions have already been supported by available research or are intuitively obvious (but not yet supported by empirical research). The first, obvious prediction is that comprehenders should fixate on plausible faults that explain the breakdown. The breakdown scenario presents to participants a discrepancy between the functioning system and the dysfunctional symptoms. After the breakdown scenario is received and interpreted, a state of cognitive disequilibrium should be triggered. This cognitive disequilibrium is expected to persist for a reasonably large stretch of time (i.e., minutes), because the devices are comparatively complex for most college students, the problem solving is challenging,

and the breakdowns are sufficiently novel that students are not likely to have solved them in the past. According to PREG, cognitive disequilibrium drives inquiry in the form of questions and eye movements. During the course of this inquiry, a successful problem solver will consider possible answers to the questions and will periodically converge on faults by fixating on regions of the text and illustrations that refer to the malfunctioning components. A second prediction is that the deep comprehenders with more device knowledge should be particularly prone to fixate on the fault regions. In contrast, individuals with lower knowledge should have a more scattered distribution of their eye fixations among the elements in the text and diagrams. A third prediction is that the deeper comprehenders should ask better questions that capture plausible faults, as has been reported in previous studies by Graesser (Graesser & Olde, 2003; Graesser et al., 2005).

In addition to these three straightforward predictions of the PREG model, there are two other predictions that address more specific mechanisms that relate eye tracking and question asking. The fourth prediction pertains to the time course of fixating on the faults in conjunction with the generation of the questions. If questions are triggered by cognitive disequilibrium, the comprehender should fixate on a particular fault prior to asking a question about that fault. That is, the onset of the question should occur after the questioner has fixated on the fault. We refer to this as the *look before you speak* (LBS) hypothesis. This hypothesis is compatible with Griffin and Bock's (2000) analysis of eye-tracking data that were collected when participants described events in pictures. Speakers began and typically completed their fixation on a picture element before they initiated their linguistic formulation of the sentence that articulated the element. Meyer, Sleiderink, and Levelt's (1998) analysis of object naming revealed that speakers fixate on objects long enough to recover the sounds of the words that denote them. In tests of the interactive model of spoken language comprehension proposed by Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995), participants listened to commands that directed them to manipulate objects in a toy environment (e.g., "Put the apple on the towel in the box"). The comprehenders fixated on objects in the visual environment before they heard the end of the word being spoken about the object. In this case, they looked before they finished comprehending a speech segment, and both of these were completed before they acted.

It is important to acknowledge, however, that one could conceive of alternative timings of the coordination between the eyes and the questions. According to the alternative, *look while you speak* (LWS) hypothesis, the eyes will fixate on the fault at the same time that the question emerges. The process of solving the problem and attempting to restore the disequilibrium is very effortful. The articulation of a question may not be launched until the eye fixations have moved to the relevant components. According to a *look after you speak* (LAS) hypothesis, the eyes fixate on the fault after the question emerges. Again, this is not an entirely facetious alternative. It may be the case

that the eyes need to verify what gets articulated by the question. The poor comprehenders may be susceptible to this pattern. As the poor comprehender conceptually explores what may explain the breakdown, he or she mentally stumbles on a question and then verifies its plausibility by visually inspecting the illustrated text at the region of the fault. Some of the questions that the poor comprehenders ask may be lucky guesses and on the mark, whereas many others will be irrelevant to the functional device mechanisms that explain the breakdown.

We investigated whether there is support for the LBS, LWS, and LAS hypotheses by investigating the time course of eye fixations and questions about plausible faults. Specifically, we analyzed the eye movement behavior for a 3-sec interval prior to the onset of each fault question (i.e., a good question that identified a plausible fault), for the interval between the onset and offset of the fault question, and for a 3-sec interval after the offset of a fault question. These analyses of eye movements before, during, and after the fault questions were expected to help us dissect the cognitive processes associated with cognitive disequilibrium and question asking.

The fifth prediction of PREG addresses whether eye fixations tend to sample causally related content at time slices before, during, or after the asking of a fault question. The PREG model predicts that a causal analysis of the antecedents and/or consequences of faults should precede the identification of a particular fault as a cause of the breakdown. Such causal analyses should be particularly prevalent in individuals with more extensive technical knowledge, because a deep understanding of a device requires the comprehender to trace the causal sequence of events in the device mechanism. Causal analyses are generally believed to be prevalent during the processing of illustrated texts on devices. For example, Hegarty (1992) demonstrated that there are mental animations of device causal mechanisms when people comprehend static displays of pulley systems. Rozenblit et al. (2002) reported that people fixate on the objects and parts of a diagram in an order that corresponds to the causal sequence of events. It remains a question, however, whether there are parallel eye fixation sequences that scan the causal events about the device mechanisms during the process of question asking. The PREG model would predict that these causal sequences would be comparatively high in the time slice that precedes the onset of a fault question, particularly for the deeper comprehenders with more device knowledge. It is also conceivable that the causal analysis follows the question as a verification mechanism, where the comprehender makes sure that each step in the causal chain is intact. However, the simplest prediction of PREG is that causal analyses should precede the fault question.

Aside from these tests of the PREG model, we examined two auxiliary questions about the role of texts and pictures when illustrated texts about device mechanisms are processed. The first question is whether the text or the picture dominates during comprehending, reasoning about, and asking questions about the content of illustrated texts. There is some evidence that the text domi-

nates many of the comprehension processes when an illustrated text is initially comprehended (Hegarty & Just, 1993; Rayner et al., 2001). In contrast, diagrams play a more prominent role than does text in guiding processing when the cognitive task emphasizes problem solving and encourages mental animation of the causal mechanisms (Hegarty, 1992; Rozenblit et al., 2002). The diagrams in our illustrated texts may dominate processing during the course of an individual's asking fault questions, because (1) diagrams encourage explanatory reasoning more than does text (Ainsworth & Loizou, 2003), (2) diagrams have a closer fit to the mental models that explain the inner workings of devices (Hegarty et al., 2002; Kosslyn, 1994; Rozenblit et al., 2002), and/or (3) diagrams are perceived to be more interesting and result in more active processing (Mayer, 2001). The second question pertains to the integration of text and pictures. Researchers have discussed both the virtues and the challenges involved in integrating the information conveyed in texts and pictures (Schnotz, Bannert, & Seufert, 2002). The integration of texts and pictures facilitates learning and memory according to Paivio's (1986) dual-coding theory and Mayer's (2001; Moreno & Mayer, 1999) analysis of multimedia. However, empirical data have revealed that the process of moving the eye back and forth between texts and diagrams is not particularly frequent when adults comprehend illustrated texts (Hegarty et al., 2002; Rayner et al., 2001). Each modality may take on a life of its own in guiding processing and eye movements, but with periodic, intermittent, strategic alignments between the contents of text and pictures.

METHOD

Participants

The participants were 40 college students at the University of Memphis. The students participated for course credit in an introductory psychology class.

Materials

Illustrated texts and breakdown scenarios. The participants read five illustrated texts on everyday devices: a cylinder lock, an electronic bell, a car temperature gauge, a toaster, and a dishwasher. The device mechanisms were extracted from Macaulay's (1988) book of illustrated texts, *The Way Things Work*.

A breakdown scenario was prepared for each of the five devices. The breakdown scenario consisted of one or two sentences that identified the physical symptoms of a device malfunction. In the case of the cylinder lock, the breakdown scenario was "the key turns but the lock does not move." The plausible faults are the following: The cam is broken, the rod that the cam is hooked over is broken, and the intersection between the cam and its connecting rod could be broken, slipping, or not connected in some way. There were particular regions in the illustrated text that directly manifested such faults. One very important procedure was followed when the breakdown scenarios were created: The breakdown could be explained by a very small number of components, parts, events, or processes in the device system.

Two other examples will be described to convey the nature of the breakdown scenarios and faults. The breakdown scenario for the dishwasher was the following:

A person ran the dishwasher but the results were disappointing. The dishwasher was warm, the detergent was used, and most of the dishes

were clean. However, food and dirt were at the bottom of the dishwasher and in the spray arm nozzles.

The faults for this breakdown scenario would be the water pump, the filter, and a clogged drain. The breakdown scenario for the electric bell was the following:

A person walks up to the door and presses the button. There is a short “ding” and then no sound is heard even though the person continues to press the button. When looking at the bell, one can see that the hammer is resting against the bell.

The fault for this breakdown is that the contact is still touching the spring, which has the unwanted effect of there being a continual current flow. The fault is pinpointed at the region that includes the contact, the spring, and their connections; it does not include the large set of regions that captures the current flow.

Device comprehension test. The device comprehension test was one of our measures of deep comprehension. Graesser and Olde (2003) reported that this test had a .72 correlation with the Technical Knowledge subsection of the Armed Services Vocational Aptitude Battery (ASVAB; Department of Defense, 1983), which is described below. The device comprehension test consisted of 6 three-alternative, forced-choice questions about each device (30 total questions across the five devices). There were four test questions per device that tapped explicit information and two questions that tapped inferences. Examples of such questions are the following:

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

- (a) the key is inserted (correct answer)
- (b) the key is removed
- (c) the key is turned.

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

- (a) they rise
- (b) they drop
- (c) they remain stationary (correct answer).

The device comprehension scores could vary from 0 to 30. A score of 10 would be chance performance if there was no sophisticated guessing or auxiliary background knowledge.

Tests of individual differences. Following the device comprehension test, the participants completed a battery of tests that measured their cognitive abilities and personality. We included measures of individual differences that significantly predicted question asking in the study by Graesser and Olde (2003). The tests of cognitive ability included parts of the ASVAB (Department of Defense, 1983), a test that is administered to over 1 million high school students each year. The following Technical Knowledge subscales were used from this test: Mechanical Comprehension, Electronics, General Science, and Auto & Shop. As was discussed earlier, the Technical Knowledge scale was the primary component that predicted question quality in Graesser and Olde’s study; there was no significant incremental variance explained, either individually or collectively, by over a dozen additional measures of cognitive ability: other ASVAB aptitude measures (verbal, quantitative, and coding speed), spatial reasoning, working memory span, most personality measures, age, and gender. However, two noncognitive variables were measured in the present study: gender and the Openness subscale of the NEO inventory (Costa & McCrae, 1991). The Openness subscale partially captures creativity and was modestly correlated with question quality scores and device comprehension scores in Graesser and Olde’s study. The gender variable predicted the number of questions asked, but not the quality of the questions, in Graesser and Olde.

Apparatus

Eye movements were recorded by a Model 501 Applied Science Laboratory eyetracker. There was a head-mounted recording unit so the participants could move the head during data collection. The participants were calibrated before they started the experimental session of reading the illustrated texts and asking questions. During

calibration, the participants viewed nine points on the computer display and a computer recorded the x -, y -coordinates. The equipment, computer, and focus of the eye fixation became synchronized after these recordings. The calibration process took 10–15 min, depending on individual differences. The participants were dismissed if they wore glasses, but the equipment could accommodate contact lenses. Eye-tracking data was recorded 60 times per second (once every 17 msec) and were averaged over every two data points through the ASL eyetracker interface program. The participants were seated 61 cm from a 51-cm flat screen monitor. At that distance and screen size, the level of resolution was approximately three characters (1.5 mm), or 1° of visual angle.

The experimental session was videotaped and audio recorded. The video camera recorded a TV screen that displayed the illustrated text and a superimposed image of what the left eye was fixating on. The superimposed image was generated by the eye-tracking equipment. The voice of the participant was also recorded on the video camera so that the spoken questions could be transcribed. This setup allowed us to record and review (1) the contents of the computer display, (2) the fixated region of the left eye, and (3) the voice of the student asking questions.

Computer software was available to record eye-tracking behavior. The software provided area plots and gaze traces after eye-tracking data have been collected on a display. An *area plot* displayed the amount of time that the eye fixated at each region in an $N \times M$ dimensional grid. The *area of interest* (AOI) was the subset of the display that should theoretically receive fixations (such as the faults of a malfunctioning device). The area plot is to be contrasted with a *gaze trace*, which plotted the sequence of eye fixations at x - and y -coordinates as a function of time. When examining the gaze traces, we would expect eye movements and eye fixations to drift toward a locus of disequilibrium (i.e., a fault) immediately before and/or during the articulation of a question.

Procedure

Recording of eye tracking and question asking. After the participant and the eyetracker had been calibrated, the participant received illustrated texts on five devices, one at a time. For each device, the participants read an illustrated text for 3 min, which was displayed on a computer monitor. After this *reading* phase, the breakdown phase was initiated for 90 sec. A breakdown description was presented either above or to the left of the illustrated text (e.g., “The person puts the key in the lock and turns the lock, but the bolt does not move,” in the context of the cylinder lock). The breakdown was intended to place the reader in cognitive disequilibrium. The *question-asking* phase began with the onset of the breakdown scenario. In this phase, the illustrated text remained on the screen with the breakdown sentence. The participants were asked to generate questions aloud for 90 sec during this phase, and their responses were tape recorded. Each participant read and supplied questions for all five devices. The assignment of devices to test order was counterbalanced across the 40 participants with a 5×5 Latin square, with 8 participants per ordering.

After providing the question-asking protocols for all five devices, the participants completed the device comprehension test, which measured their understanding of the devices. The order of device comprehension questions was congruent with the order in which the devices had been presented.

Following the device comprehension test, the participants completed the battery of individual-difference tests that measured cognitive abilities and personality. It took approximately 2 h to complete the battery of tests, which were completed on a different day than the eye-tracking portion of the experiment.

Scoring of questions. Two measures were scored on the verbal protocols that were collected in the question-asking task: (1) *number of questions* asked during the question-asking phase, and (2) *number of fault questions*, which was the number of questions that referred to a plausible malfunction. Two trained judges segmented the protocols

into separate questions or idea units. The proportion of questions or idea units that were the same between the two judges was .75. The same judges also determined whether a question or idea matched a fault node. A group of five judges conferred and produced a list of acceptable answers for each device. This list was consulted by the two judges when making decisions about fault questions (questions that matched a fault node). When a participant gave a uniquely creative answer, that answer was revisited by the larger group of judges and was assessed as to whether it matched a fault. If the answer was deemed acceptable, it was added to the list. The judges reached a high level of agreement in making these fault decisions (.90 or higher between any pair of judges).

RESULTS

Descriptive Statistics

Table 1 presents descriptive statistics on the measures collected in the eye-tracking experiment. The means and standard deviations for the measures of individual differences were very similar in this experiment and the question-asking study reported in Graesser and Olde (2003). The device comprehension score was 18.6 (out of 30) for the five devices in this experiment, which means that 62% of the questions were answered correctly. This is comparable to the 65% found in Graesser and Olde's study that had six devices tested. Other measures that were comparable to those in the previous study were number of questions (5.3 in the present study and 3.8 in Graesser and Olde), number of fault questions (1.4, 1.8), Openness (52.1, 51.8), and the ASVAB's Mechanical Reasoning (13.9, 14.7), Electronics (9.3, 11.4), General Science (16.9, 18.6), and Auto & Shop (10.0, 12.3) subscales.

Regarding the eye-tracking measures, the present study had a mean of 29.5 fixations on plausible faults per device, or 9.3 sec out of 90 sec. The percentage of eye fixations that were on faults was 11.5%, whereas the percentage of time on the faults was 10.4%. It should be noted that the percentage of time on faults and the total fault fixation time are functionally equivalent, because the participant

was always allocated 90 sec per device for question asking. The percentage of the screens that was taken up by AOIs designated as faults was 7%.

Correlations

Table 2 presents bivariate correlations between measures in the eye-tracking experiment. This experiment replicated Graesser and Olde's (2003) experiment in showing that device comprehension scores were significantly correlated with the ASVAB Technical Knowledge composite measure (and all of its component measures), Openness, and gender. This again demonstrates that the device comprehension score is an excellent measure of deep comprehension. As in Graesser and Olde's study, the number of questions asked about plausible faults was also significantly correlated with device comprehension scores, whereas the total volume of questions was not. This outcome therefore supports Prediction 3 of the PREG model that deeper comprehenders should ask better questions that capture plausible faults. The correlations between fault questions and measures of technical knowledge, which are not reported in Table 2, were also significantly positive, ranging from .31 and .45 when the four subscales of Technical Knowledge were considered; this outcome also replicates that in Graesser and Olde.

Table 2 revealed that all three measures of the eye-tracking performance were significantly correlated with device comprehension scores: the number of eye fixations on faults ($r = .43$), the percentage of fixations on faults (.31), and the total time fixating on faults (.50). Thus, a valid litmus test of deep comprehension is whether the participant spends a greater percentage of time focusing on plausible faults when faced with a breakdown scenario. These correlations support Predictions 1 and 2 of the PREG model.

The magnitude of the correlations supports the claim that fixating on faults is a robust and quick indicator of deep comprehension. For example, the ASVAB Technical Knowledge composite score had a .54 correlation with device comprehension scores; the proportion of time the eye fixated on faults had a .50 correlation with device comprehension scores, almost as high. One of the advantages of the eye-tracking data is that deep comprehension can be assessed for specific devices, whereas the scope of a psychometric ASVAB test is generic, rather than specific.

The question arises as to whether comprehension ability and technical knowledge had a causal impact on both question asking and eye tracking or, instead, the inquiry processes caused improved device comprehension scores and technical knowledge. The present study was not designed to resolve the direction of causality, but we are convinced that it was the ability of the participants (as measured by the device comprehension test and technical knowledge), in conjunction with the breakdown scenario, that drove inquiry in the present study, rather than inquiry having a causal impact on the measures of comprehension and technical ability on subsequent tests. It is very unlikely that the breakdown scenarios and the question-asking task would have any influence on ASVAB's psy-

Table 1
Descriptive Statistics on Measures Collected
in Eye-Tracking Experiments

Measure	Mean	Standard Deviation
Armed Services Vocational Aptitude Battery		
Mechanical Comprehension	13.9	6.0
Electronics	9.3	4.3
General Science	16.9	4.7
Auto & Shop	10.0	5.3
Openness	52.1	9.2
Gender (female = 1, male = 2)	1.25	0.44
Verbal protocol measures		
Number of questions per device	5.3	2.9
Number of fault questions	1.4	0.6
Eye-tracking measures		
Number of fault fixations per device	29.54	11.1
Percentage of fixations on faults	11.5	4.1
Total fault fixation time per device	9.3	3.9
Percentage of time on faults	10.4	4.3
Device comprehension score	18.6	4.6
Number of participants	40	

Table 2
Correlations in Eye-Tracking Experiment

Measure	Comprehension Score	Fixation on Faults		
		Number of Fixations	Percentage of Fixations	Time on Faults
ASVAB Technical Knowledge	.54*	.31*	.28	.33*
Mechanical Comprehension	.39*	.16	.14	.15
Electronics	.32*	.18	.23	.19
General Science	.60*	.49*	.43*	.51*
Auto & Shop	.58*	.25	.18	.29
Openness	.41*	.31*	.36*	.41*
Gender (female = 1, male = 2)	.45*	.27	.18	.27
Device comprehension score	—	.43*	.31*	.50*
Number of questions per device	.20	.08	.20	.21
Number of fault questions	.45*	.42*	.49*	.52*

Note—ASVAB, Armed Services Vocational Aptitude Battery. * $p < .05$.

chometric measures of technical knowledge. The majority of questions on the device comprehension test also did not correspond to the faults in the breakdown scenarios.

Eye Fixations on Faults When Fault Questions Are Asked

Follow-up analyses focused exclusively on the good questions that were asked. These were the questions that referred to faults that would plausibly explain the breakdown. The asking of a good question could be the result of a systematic causal analysis that ended up converging on a fault; alternatively, a good question could be a lucky guess that was produced while the comprehender was inspecting the display without deep understanding. Deep comprehension is presumably a product of systematic inquiry, rather than lucky guesses. The purpose of the analyses in this subsection was to investigate the coordination of eye fixations with the asking of good questions. We computed the percentage of eye fixations in a fault region as a function of (1) high versus low technical knowledge, as measured by the ASVAB, (2) device (there were five different devices), and (3) time slices (3 sec *before* the question, time *during* the question, and 3 sec *after* the question). A median split criterion was used to segregate the participants into those with high versus low technical knowledge. A follow-up analysis also segregated the participants into those with high, medium, and low scores on the Technical Knowledge scale, but the three-group split did not shed additional light on the results and will not be reported.

Figure 2 plots the percentage of eye fixations on plausible faults as a function of the three time splices and high versus low technical knowledge scores. An analysis was performed that compared eye fixations on faults with chance performance. Chance for eye fixations on faults was 7.4%, on the basis of the amount of space that the AOIs occupied on the screen. Low technical knowledge participants did not differ from chance in their percentage of eye fixations on faults [$M = 6.4\%$, $SD = 5.6$; $t(19) = 0.82$, $p = .42$]. In contrast, participants with high technical knowledge differed from chance in the percentage of eye fixations on faults [$M = 11.4\%$, $SD = 7.0$; $t(19) = 2.58$, $p < .05$]. Thus, participants with high technical knowledge scores were not randomly exploring the device, as was the case for the low technical knowledge participants; high-knowledge participants were fixating on plausible fault locations with a likelihood higher than

chance performance. Chance for eye fixations on faults was 7.4%, on the basis of the amount of space that the AOIs occupied on the screen. Low technical knowledge participants did not differ from chance in their percentage of eye fixations on faults [$M = 6.4\%$, $SD = 5.6$; $t(19) = 0.82$, $p = .42$]. In contrast, participants with high technical knowledge differed from chance in the percentage of eye fixations on faults [$M = 11.4\%$, $SD = 7.0$; $t(19) = 2.58$, $p < .05$]. Thus, participants with high technical knowledge scores were not randomly exploring the device, as was the case for the low technical knowledge participants; high-knowledge participants were fixating on plausible fault locations with a likelihood higher than

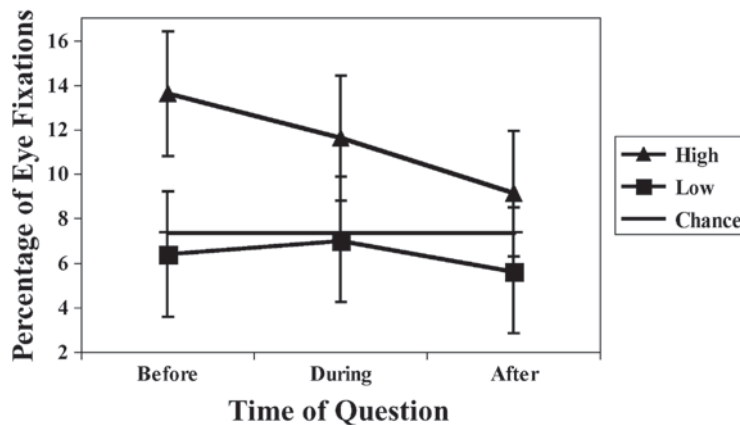


Figure 2. Percentage of eye fixations in a fault region as a function of high versus low technical knowledge and time slices.

chance. For the high-knowledge participants, the percentage of eye fixations scores were 13.6%, 11.6%, and 9.1% before, during, and after the question, respectively. The corresponding percentages for the participants with low knowledge were 6.4%, 7.0%, and 5.7%, respectively (see Figure 2). When these cell means are compared with chance performance, it becomes apparent not only that the high technical knowledge participants tended to fixate on fault locations more often than chance, but also that they did so *before* question asking [$t(19) = 3.30, p < .05$] and *during* question asking [$t(19) = 2.12, p < .05$]. However, there was no significant difference from chance *after* question asking [$t(19) = 1.06, p = .30$].

An ANOVA was performed on these eye fixation scores (which were converted to proportions), using a technical knowledge \times device \times time slice design. Knowledge was a between-subjects variable, whereas device and time slice were within-subjects variables. There were statistically significant main effects of knowledge [$F(1,38) = 6.42, MS_e = 0.006, p < .05$], device [$F(4,152) = 15.86, MS_e = 0.043, p < .05$], and time slice [$F(2,76) = 3.41, MS_e = 0.011, p < .05$]. There were no significant interactions between device and any of the other variables. Therefore, we performed a 2×3 ANOVA between knowledge and time slice that collapsed over device. The interaction between knowledge (high or low) and time slice (before, after, or during) was not quite significant ($p < .10$, one-tailed), but this 2×3 interaction was muted by the intermediate status of the eye movements during the asking of the question. We tested the predicted interaction between knowledge and time slice (before vs. after, but not during). This interaction would be predicted by the PREG model, so we treated it as an a priori planned test. There was a statistically significant interaction between knowledge and time slice [$F(1,76) = 3.29, MS_e = 0.002, p < .05$, one-tailed]. The high-knowledge participants spent significantly more time than did the low-knowledge participants looking at fault locations before asking a fault question [$F(1,38) = 9.80, MS_e = 0.005, p < .05$]. Yet there was no significant difference between the two groups of subjects for the 3 sec after asking a fault question [$F(1,38) = 2.39, MS_e = 0.005, p = .13$]. The data revealed also that there was a marginally significant difference between groups of participants during the asking of a fault question [$F(1,38) = 3.47, MS_e = 0.006, p = .07$]. These results are all consistent with the conclusion that the participants with high technical knowledge fixated on fault locations before and during fault questions (again, see Figure 2). These results support Prediction 4 of the PREG model.

An analysis of cross-panel correlations was performed in order to assess the relationship between technical knowledge, question quality, and eye fixations. When quality of the question was controlled for, the partial correlations between technical knowledge and eye fixations were .38 ($p < .05$) before the question, .16 (n.s.) during the question, and .27 (n.s.) after the question. In contrast, when fixations before, during, or after the question were controlled for, the partial correlations between technical knowledge and the quality of the question were all sig-

nificant (.39, .35, and .38, respectively; $p < .05$). When technical knowledge was controlled for, the partial correlations between the quality of the question and fixations before, during, or after the question were not significant ($-.05, .22$, and $.01$, respectively; $p > .05$). These results support a number of conclusions about the question-asking mechanism. First, as predicted by the PREG model, the amount of technical knowledge robustly predicts whether a comprehender will ask a high-quality question and fixate on a fault; without this technical knowledge, the inquiry mechanism is not initiated on a causal trajectory that will explain the breakdown. Second, because the correlations were not 1.0 and because the partial correlations were high between question quality and technical knowledge when eye fixations were controlled for, we can conclude that high-quality questions are sometimes generated without requiring fixation on the faults; these questions are possibly launched by top-down cognitive mechanisms that are not directly integrated with eye movements. Third, quality questions are sometimes preceded by the activity of the eye fixating on the fault, as would be predicted by the PREG model. Therefore, there appear to be two cognitive trajectories that launch the generation of a high-quality question: one that is instigated by a conceptual analysis that is not tightly coupled to the moment-to-moment perceptual processes (called *conceptually driven* inquiry) and one that is triggered by perceptual input (called *perceptually driven* inquiry).

There might be cause for concern that only 13.6% of the fixations were on faults before the fault questions had been asked by participants with high technical knowledge. This appears to be a low percentage, so one might be somewhat skeptical about the prevalence of the inquiry mechanisms, and one might ask what the participants were doing to account for the other 86.4% of the fixations. The 13.6% figure is not necessarily a small percentage when one takes into consideration the fact that a large proportion of the screen was taken up by white space (.333) and the breakdown scenario (.122); the remaining .545 of the screen was taken up by pictorial diagram information (.334), text (.157), arrows (.019), and labels (.035). If we take the 13.6% figure and compute a conditional percentage that takes into consideration only useful real estate on the display (e.g., $.136/.545$), the high-knowledge participants were fixating on fault regions approximately 25% of the time. Moreover, the eyes were not fixated on the display at all for approximately 17.5% of the time. If we remove the off-screen eye fixations from consideration in our estimate, approximately 30% of the fixations were on fault regions [$.250/(1.0 - .175) = .303$] before a fault question had been asked by a participant with high technical knowledge.

Eye Fixations on Causal Paths When Fault Questions Are Asked

According to Prediction 5 of the PREG model, participants should be inspecting causal chains or paths in the illustrated texts that explain the faults. This should occur particularly for participants with high technical knowl-

edge during the time slice before a fault question is asked. As in the above analyses, the means were computed for participants with high versus low scores on the Technical Knowledge scale (sometimes referred to as *good vs. poor* comprehenders), and the three time slices were segregated for the fault questions.

In order to perform a sequential causal analysis, there needed to be an operational definition for the movement of the eye along a causal path. The first step in providing such an operational definition was the adoption of a representational system that captured the content of the illustrated text. A system for representing device knowledge has been reported in previous publications by Graesser (Graesser & Clark, 1985; Graesser & Hemphill, 1991; Graesser, Wiemer-Hastings, & Wiemer-Hastings, 2001). The representational system consists of a conceptual graph structure that has a set of content nodes (referring to objects, parts, states, events, processes, and goals) and relations that connect the nodes (such as *is-a*, *has-as-part*, *is-in*, *various spatial relations*, *causes*, *enables*, *outcome*, and *implies*). The conceptual graph structure captures both text and picture content in one common representation. From the standpoint of the present analysis, the causal chains of objects, parts, states, events, and processes were identified in the structure and were presented to two research assistants. One of the research assistants created AOIs for the areas on the illustrated text that referred to the content of the causal paths. The other research assistant viewed and verified the accuracy of the AOIs that were created; changes were made to the satisfaction of both research assistants. Causal AOIs were defined as the set of AOIs that capture the causal paths of occurrences in the device system; these might or might not be related to the breakdown scenarios.

The next step in the analysis was to identify causal paths in the eye fixations that occurred before, during, or after the fault questions. Two research assistants examined each time slice associated with a fault question and decided whether there was a causal path of eye fixations within each time slice. A causal path was defined as five successive eye fixations on any causal path. Interjudge reliability data were collected for these judgments. The research assistants randomly sampled three devices for each participant and independently coded whether the sequence of eye fixations did or did not contain five continuous fixations on a causal path in the device mechanism. The interjudge reliability was high; 93% of the decisions were the same, with approximately half of the judgments being yes and half no. Given that the interjudge reliability was high, the research assistants scored the time slices for each of the fault questions. A *causal path likelihood* score was defined as the proportion of fault questions for which there was at least one causal path within a particular time slice. A *causal path density* score was the proportion of total fixations in a time slice that was part of a causal path.

The analyses revealed that a large proportion of the eye fixations were on causal paths, but the scores did not differ significantly across the three time slices and between par-

ticipants with high versus low technical knowledge. There were no significant effects in an ANOVA that crossed technical knowledge (high vs. low) and time slice (before, during, or after). The *causal path* scores were .500 for both high- and low-knowledge participants. The scores were .508, .525, and .467 for the time slices before, during, and after a fault question. The same analysis was conducted on the causal path density scores, and again there were no significant effects. The causal density scores were slightly lower for high- than for low-knowledge participants, with means of .421 and .484, respectively, but the difference was not significant. The mean causal path density scores were .447, .495, and .415 in the time slices before, during, and after a fault question. The main effects and interactions in these analyses all had *F* scores of 1.07 or lower.

Another operational definition of causal processing is to compute the likelihood that there was at least one pair of contiguous eye fixations that were in the AOIs of a causal path. Two contiguous AOIs were operationally defined as being on a causal path if the AOIs were affiliated with adjacent nodes in the causal structure. This *contiguous causal fixation* score is an objective automated method of scoring the eye-tracking data, as opposed to relying on the trained judges to decide whether there was a causal path in a sequence of eye fixations. Nevertheless, as in the analyses above, the contiguous causal fixation score showed no significant main effects and no significant interaction in an ANOVA.

In summary, our analysis of causal paths did not show strong support for Prediction 5 of the PREG model. The good news is that eye fixations frequently followed causal relations and sequences during question asking. However, these patterns did not significantly vary as a function of the participants' technical knowledge and as a function of the three time slices (before, during, and after a fault question). Apparently, participants with high versus low technical knowledge perform causal analyses while inspecting the diagrams to the same extent, whereas those with high knowledge are more likely to converge on the actual faults. It is apparently the case that causal analyses are equally prevalent before, during, and after a fault question.

Eye Fixations on Texts and Pictures When Fault Questions Are Asked

The proportion of eye fixations on the AOIs for text were computed for each of the three time slices and two groups of participants. ANOVAs showed no significant main effects or interactions ($F_s < 1$, mean proportion = .250). Also, no significant effects occurred for the proportion of time spent fixating on text AOIs ($F_s < 1$). The same set of analyses were performed on the picture AOIs, but again there were no significant main effects or interactions ($F_s < 1$, mean proportion = .483). When the AOIs of text and pictures are directly compared, .68 of the area on the display is devoted to pictures (assuming that white space is removed from consideration), whereas the remaining .32 of the area is devoted to text. These proportions are very close to the relative proportions of time

spent inspecting pictures versus text (.66 vs. .34, respectively).

Text–picture integration was defined as a contiguous pair of eye fixations that shifted from a picture to a text or vice versa. AOIs were collected for the text and picture areas of the illustrated texts. We computed the likelihood that there was at least one pair of contiguous eye fixations that were in AOIs that shifted between text and picture. This *contiguous text–picture fixation likelihood* score is an objective automated method of scoring the eye-tracking data, as opposed to relying on human judgments. An ANOVA on these scores showed a significant main effect of time slice [$F(2,76) = 3.84$, $MS_e = 0.054$, $p < .05$], but no significant main effect of technical knowledge ($F < 1$) and no significant interaction ($F < 1$). The scores were .235, .370, and .260 for the time slices before, during, and after a fault question. The scores were .317 and .260 for high- and low-knowledge participants, respectively.

DISCUSSION

This study was designed to test five predictions of the PREG model of question asking (Graesser & Olde, 2003; Graesser et al., 2005; Otero & Graesser, 2001) when participants are confronted with cognitive disequilibrium from breakdown scenarios and when eye movement data are collected. In this section, we will begin by taking stock of whether the five predictions were supported in the analyses of eye tracking and question asking. We subsequently will address some auxiliary questions about the comprehension of illustrated texts and mental models of devices.

The first prediction of the PREG model was that the comprehenders would fixate on the faults that plausibly explain the breakdowns of the devices. A breakdown scenario was expected to put the participants in a state of cognitive disequilibrium, which would encourage problem solving and inquiry. The eye-tracking results supported this first prediction. Comprehenders collectively did show a significant tendency to fixate on appropriate defective components when reflecting on the breakdown scenarios. We assumed that the cognitive disequilibrium would persist during most of the 90 sec that the participants spent reading the breakdown and generating questions, as was reported in the question-asking study of Graesser and Olde (2003). It is conceivable that the participants occasionally resolved the disequilibrium quickly, but that was not apparent when we inspected the content of the questions throughout the 90 sec. It was difficult for these college students to pinpoint the faults of the device breakdowns, because it required knowledge about mechanical and electronic systems and it was very unlikely that they had previously encountered such breakdowns prior to this experiment.

The second prediction of the PREG model is that the good comprehenders—that is, those with relatively high technical knowledge—would be particularly likely to fixate on the fault regions. A satisfactory amount of technical world knowledge and ability is required before one can

converge on likely faults and formulate the questions. This prediction was supported by the finding that high comprehenders had an above-chance incidence of fixating on the fault locations, whereas the poor comprehenders did not fixate on the faults above chance. When considering the successful detection of faults, poor comprehenders moved their eyes indiscriminately over the display. These differences are illustrated in Figures 3 and 4. Figure 3 shows the eye fixations for a participant with high technical knowledge. There is a high density of eye fixations in the regions of the cam, particularly near the lip of the cam. In contrast, Figure 4 shows the eye fixations for a participant with low technical knowledge. The eye fixations are quite scattered and not concentrated on the cam.

The third prediction of PREG is that device comprehension is manifested in the quality of the questions asked as a person reflects on a broken device. The results of the present study support this prediction and replicate the results reported in some previous studies (Graesser & Olde, 2003; Graesser et al., 2005). Good comprehenders asked more fault questions when reflecting on the breakdown scenarios. As in the previous studies, it is noteworthy that it is the number of *fault* questions, not the total number of questions, that is diagnostic of deep comprehension. The finding that quality prevails over the quantity of questions is compatible with a previous study by Fishbein, Eckart, Lauver, Van Leeuwen, and Langmeyer (1990), who 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 showed a nonsignificant negative correlation. Graesser and Person's (1994) analysis of student questions in tutoring also revealed 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, as in the present study. The fact that the same results emerged in the present study, our previous studies on everyday devices, and previous analyses of tutoring supports the general conclusion that it is quality of questions, not quantity of questions, that reflects deep comprehension.

The fourth prediction of the PREG model addressed the timing of eye movements and question asking. According to the PREG model, questions are triggered by cognitive disequilibrium, so the comprehenders should fixate on the faults prior to (or possibly during) their asking a question about that fault. That is, the eye's fixation on the fault should begin before or during the onset of the question, as stipulated in the LBS and the LWS hypotheses. The results supported this claim by showing that good comprehenders did have a significant tendency to fixate on critical fault locations before and possibly during the asking of a question about those faults; in contrast, the eyes did not significantly fixate on fault regions any more than would be expected by chance when eye movements after the question were considered.

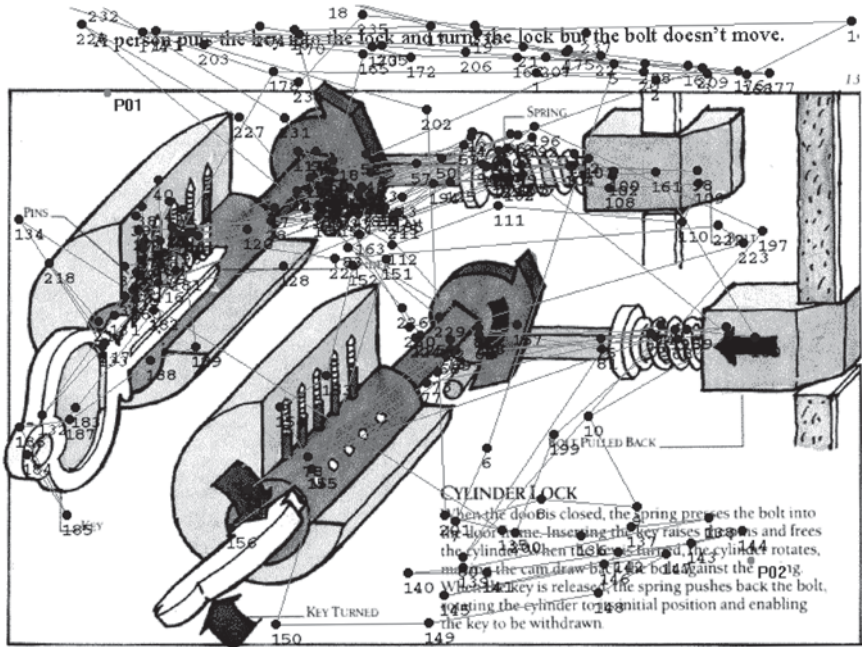


Figure 3. Eye fixations of a participant with high technical knowledge.

The pattern of partial correlations revealed that fault questions may sometimes be launched by a top-down conceptual analysis and that question delivery does not always have to be tightly coupled with corresponding eye fixation on the fault regions. The alternative processing trajectory is that questions are launched in a more bottom-up fashion when perceptual input associated with the fault is in-

terpreted and, possibly, mentally animated. Thus, there appear to be two ways of generating a high-quality question about an actual fault that explains the breakdown: one perceptual input directed, and the other conceptual knowledge directed. When the cognitive trajectory is perceptual input directed, the comprehender's eyes first fixate on the fault region, the fault is interpreted, and then

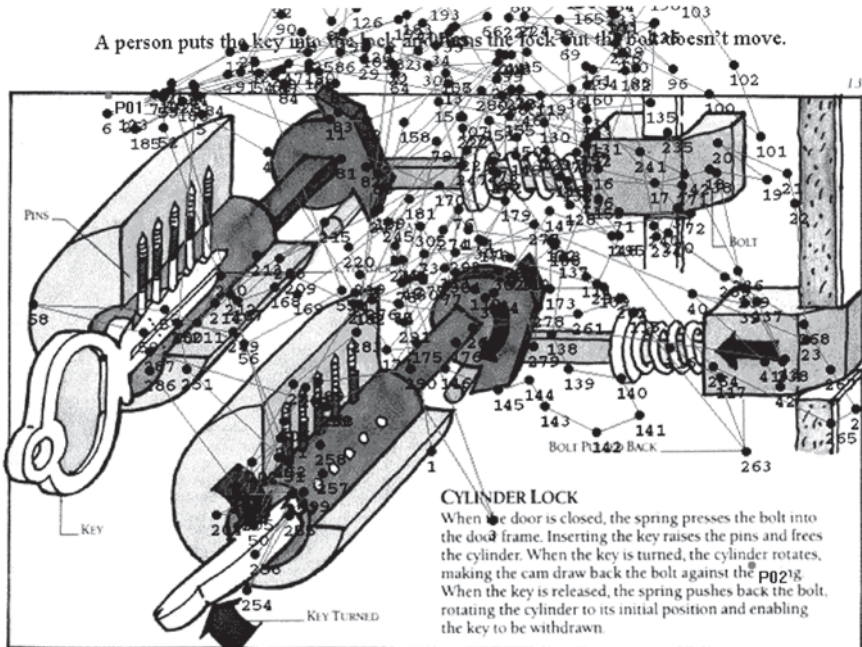


Figure 4. Eye fixations of a participant with low technical knowledge.

the question emerges. When the cognitive trajectory is conceptual knowledge directed, the question is directly launched from the cognitive system, without a systematic coordination of inquiry with eye movements.

Prediction 5 of the PREG model was not strongly supported when we examined sequences of eye fixations that followed causal paths in the device mechanism. We did find that a large proportion of eye fixation sequences followed causal paths. The high incidence of these causal fixation sequences is perfectly compatible with the kinematic-dynamic mental model of devices that has been investigated by Hegarty (Hegarty, 1992; Hegarty & Just, 1993; Hegarty et al., 2002; Hegarty & Sims, 1994) and by Rozenblit et al. (2002). Hegarty has reported that these causal paths are prevalent in mental animations of device operations, in addition to the visual inspection of illustrated texts. In the present study, the synchrony of eye movement sequences with causal paths was equally prevalent for good and poor comprehenders, rather than being diagnostic of deep comprehension. Moreover, fixation on causal sequences was not more prevalent before the fault questions, as opposed to during or after the questions, as would be predicted by the PREG model. Apparently, participants with varying technical knowledge had a strong tendency to perform causal reasoning throughout the 90-sec question-asking task, whereas it was only the participants with higher technical knowledge who could periodically zero in on the correct faults.

The finding that good comprehenders are discriminating in identifying plausible faults is compatible with the results of some previous studies on device comprehension and on problem solving. Hegarty, Just, and Morrison (1988) reported that adults with higher mechanical reasoning ability were better able to discriminate relevant from irrelevant attributes of an illustrated text on a pulley and rope system. Hodgson et al. (2000) collected eye-tracking data while college students attempted to solve a Tower of London problem. The poor problem solvers were more likely to fixate on irrelevant units and to dwell on solutions to previous problems. Deep mental models of the internal mechanisms of a system are needed before a person can converge on the plausible faults of a broken system, as is shown in the present study, or understand the functions and mechanisms of a novel piece of equipment (see Kieras & Bovair, 1984).

Good versus poor comprehenders did not differ with respect to allocating their attention to texts versus pictures or to integrating texts and pictures. Time allocation was approximately the same for text and pictures after we controlled for the size of the AOIs devoted to each of the two media. This balanced allocation of resources to the two media persisted for participants with high versus low technical knowledge. Text-picture integration is viewed as an important process for learning and memory in Paivio's (1986) dual-coding theory and Mayer's (2001) analysis of multimedia. However, it was not a frequent cognitive process in the present study when participants experienced cognitive disequilibrium during a breakdown scenario. The proportion of contiguous eye fixations that

were text-picture integrations was only .08 within the 3-sec time slides before, during, and after a fault question. Comprehenders of illustrated text tend to be absorbed in one modality at a time, with infrequent and strategic alignments between media (Carroll, Young, & Guertin, 1992; Rayner et al., 2001).

In summary, when a breakdown scenario is provided in the context of an everyday device, someone with a deep comprehension of the device asks good questions that converge on faults that are likely to explain the breakdown. We found that 30% of the deep comprehenders' time is spent fixating on faults after we remove from consideration the eye fixations on the breakdown scenario, on the white space, and off the screen. Most of the rest of the time is spent performing causal analyses of the device system. Some of the high-quality questions are systematically coordinated with eye movements. In these instances, the good comprehender visually fixates on the malfunctioning component prior to or possibly during the formulation of the question. In contrast, the poor comprehenders indiscriminately move their eyes around the visual display, fixating on the fault regions only at chance levels. Both good and poor comprehenders are equally inclined to inspect causal paths in diagrams, to read the text, to inspect the pictures, and to visually integrate text and picture during the course of delivering a good question.

REFERENCES

- AINSWORTH, S., & LOIZOU, A. T. (2003). The effects of self-explaining when learning with texts or diagrams. *Cognitive Science*, *27*, 669-681.
- ANDERSON, J. R., BOTHELL, D., & DOUGLASS, S. (2004). Eye movements do not reflect retrieval processes: Limits of the eye-mind hypothesis. *Psychological Science*, *15*, 225-231.
- BERLYNE, D. E. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill.
- CARROLL, P. J., YOUNG, J. R., & GUERTIN, M. S. (1992). Visual analysis of cartoons: A view from the far side. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 444-461). New York: Springer.
- CHINN, C. A., & BREWER, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*, 1-49.
- COLLINS, A. (1988). Different goals of inquiry teaching. *Questioning Exchange*, *2*, 39-45.
- COSTA, P. T., JR., & McCRAE, R. R. (1991). *NEO: Five-Factor Inventory*. Odessa, FL: Psychological Assessment Resources.
- DEPARTMENT OF DEFENSE (1983). *Armed Services Vocational Aptitude Battery, Form 12a*. Washington, DC: Author.
- 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.
- FOX, R. J., KRUGMAN, D. M., FLETCHER, J. E., & FISCHER, P. M. (1998). Adolescents' attention to beer and cigarette print ads and associated product warnings. *Journal of Advertising*, *27*, 57-68.
- GRAESSER, A. C., BAGGETT, W., & WILLIAMS, K. (1996). Question-driven explanatory reasoning. *Applied Cognitive Psychology*, *10*, S17-S32.
- GRAESSER, A. C., & CLARK, L. C. (1985). *Structures and procedures of implicit knowledge*. Norwood, NJ: Ablex.
- GRAESSER, A. C., & HEMPHILL, D. (1991). Question answering in the context of scientific mechanisms. *Journal of Memory & Language*, *30*, 186-209.

- GRAESSER, A. C., & McMAHEN, 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. (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., OLDE, B., POMEROY, V., WHITTEN, S., LU, S., & CRAIG, S. (2005). Inferences and questions in science text comprehension. *Tarbiya*, **36**, 103-128.
- GRAESSER, A. C., & PERSON, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, **31**, 104-137.
- GRAESSER, A. C., WIEMER-HASTINGS, P., & WIEMER-HASTINGS, K. (2001). Constructing inferences and relations during text comprehension. In T. Sanders, J. Schilperoord, & W. Spooen (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 249-271). Amsterdam: Benjamins.
- GRANT, E. R., & SPIVEY, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, **14**, 462-466.
- GRIFFIN, Z. M., & BOCK, K. (2000). What the eye says about speaking. *Psychological Science*, **11**, 273-279.
- HEGARTY, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 1084-1102.
- HEGARTY, M., & JUST, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory & Language*, **32**, 717-742.
- HEGARTY, M., JUST, M. A., & MORRISON, I. A. (1988). Mental models of mechanical systems: Individual differences in qualitative and quantitative reasoning. *Cognitive Psychology*, **20**, 191-236.
- HEGARTY, M., NARAYANAN, N. H., & FREITAS, P. (2002). Understanding machines from multimedia and hypermedia presentations. In J. Otero, J. A. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 357-384). Mahwah, NJ: Erlbaum.
- HEGARTY, M., & SIMS, V. K. (1994). Individual differences in mental animation during mechanical reasoning. *Memory & Cognition*, **22**, 411-430.
- HODGSON, T. L., BAJWA, A., OWEN, A. M., & KENNARD, C. (2000). The strategic control of gaze direction in the Tower of London task. *Journal of Cognitive Neuroscience*, **12**, 894-907.
- JUST, M. A., & CARPENTER, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, **87**, 329-354.
- JUST, M. A., & CARPENTER, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, **99**, 122-149.
- KIERAS, D. E., & BOVAIR, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, **8**, 255-274.
- KNOBlich, G., OHLSSON, S., & RANEY, R. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, **29**, 1000-1009.
- KOSSLYN, S. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.
- MACAULAY, D. (1988). *The way things work*. Boston: Houghton Mifflin.
- MAGLIANO, J. P., GRAESSER, A. C., EYMARD, L. A., HABERLANDT, K., & GHOLSON, B. (1993). The locus of interpretive and inference processes during text comprehension: A comparison of gaze durations and word reading times. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 704-709.
- MAYER, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.
- MAYER, R. E., & SIMS, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology*, **86**, 389-401.
- MEYER, A. S., SLEIDERINK, A., & LEVELT, W. J. M. (1998). Viewing and naming objects: Eye movements during noun-phrase production. *Cognition*, **66**, B25-B33.
- MIYAKE, N., & NORMAN, D. A. (1979). To ask a question one must know enough to know what is not known. *Journal of Verbal Learning & Verbal Behavior*, **18**, 357-364.
- MORENO, R., & MAYER, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, **91**, 358-368.
- O'BRIEN, E. J., RANEY, G. E., ALBRECHT, J. E., & RAYNER, K. (1997). Processes involved in the resolution of anaphors. *Discourse Processes*, **23**, 1-24.
- OTERO, J., & GRAESSER, A. C. (2001). PREG: Elements of a model of question asking. *Cognition & Instruction*, **19**, 143-175.
- PAIVIO, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.
- PIAGET, J. (1952). *The origins of intelligence*. Madison, CT: International Universities Press.
- RAYNER, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **124**, 372-422.
- RAYNER, K., ROTELLO, C. M., STEWART, A. J., KEIR, J., & DUFFY, S. A. (2001). Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, **7**, 219-226.
- ROZENBLIT, L., & KEIL, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, **26**, 521-562.
- ROZENBLIT, L., SPIVEY, M. J., & WOJSLAWOWICZ, J. (2002). Mechanical reasoning about gear-and-belt systems: Do eye movements predict performance? In M. Anderson, B. Meyer, & P. Olivier (Eds.), *Diagrammatic representation and reasoning* (pp. 223-240). Berlin: Springer.
- SCHANK, R. C. (1999). *Dynamic memory revisited*. New York: Cambridge University Press.
- SCHNOTZ, W., BANNERT, M., & SEUFERT, T. (2002). Toward an integrative view of text and picture comprehension: Visualization effects on the construction of mental models. In J. Otero, J. A. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 385-416). Mahwah, NJ: Erlbaum.
- TANENHAUS, M. K., SPIVEY-KNOWLTON, M. J., EBERHARD, K. M., & SEDIVY, J. C. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, **268**, 1632-1634.
- WILEY, J. (2001). Supporting understanding through task and browser design. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1136-1141). Mahwah, NJ: Erlbaum.

(Manuscript received November 14, 2003;
revision accepted for publication October 21, 2004.)