QUESTION-DRIVEN EXPLANATORY REASONING ABOUT DEVICES THAT MALFUNCTION

Arthur Graesser, Brent Olde, and Shulan Lu

Department of Psychology The University of Memphis

ABSTRACT

Questions are at the heart of virtually any task an adult performs when using technological artifacts. It could be argued that any given task T can be decomposed into a set of questions that a sailor asks and answers. When a sailor encounters a device that malfunctions, the relevant questions are "What's wrong?" and "How can it be fixed?" . When an officer reads a technical document, the relevant questions are "Why is this important?" and "What should I do about it, if anything?". When a young adult reads Navy recruiting material, the relevant questions are "What's interesting? ", "Do I want to join?", and "What are the perks?". The cognitive mechanisms that trigger question asking, exploration patterns, and question answering strategies need to be understood in order to design the messages and technological artifacts effectively. In turn, these inquiry strategies map onto cognitive components that are familiar to cognitive scientists, such as search, pattern recognition, comparison, case-based analogical reasoning, knowledge construction, and structure mapping.

In projects funded by the Office of Naval Research, we have developed a cognitive computational model of question asking (called PREG) and question answering (called QUEST). Our current ONR grant is investigating relationships among a person's understanding of everyday devices (e.g., dishwasher, cylinder lock), the asking and answering of questions, and general psychometric tests of cognitive abilities. After reading about a device, the participants subsequently receive scenarios in which the device breaks down and they generate questions about the malfunction, eye tracking data are also collected at that time. Adults with high mechanical ability ask good questions that converge on faults. The quality of questions in the center of a breakdown scenario is the fastest way to find out whether an adult has a deep understanding of a device.

Introduction

Questions are at the heart of virtually any task an adult performs. It could be argued that any given task can be decomposed into a set of questions that a person asks and answers. For example, when a sailor in the Navy encounters a device that malfunctions, the relevant questions are "What's wrong?" and ""How can it be fixed?". When an officer reads a technical document, the relevant questions are "Why is this important?" and "What should I do about it, if anything?". When a young adult reads Navy recruiting material, the relevant questions are "What's interesting?", "Do I want to join?", and "What are the perks?". The cognitive mechanisms that trigger question asking and exploration patterns need to be understood in order to optimize the design of artifacts, whether they be text, visual displays, mechanical devices, electronic equipment, or telecommunication systems.

In a recent project funded by the Office of Naval Research, we developed a cognitive computational model of question asking, called PREG (Graesser, Olde, Pomeroy, Whitten, Lu, & Craig, in press; Otero & Graesser, 1999). According to the PREG model, cognitive disequilibrium drives the asking of genuine information-seeking questions (Berlyne, 1960; Chinn & Brewer, 1993; Collins, 1988; Festinger, 1957; Flammer, 1981; Graesser, Baggett, & Williams, 1996; Graesser & McMahen, 1993; Graesser & Person, 1994; Schank, 1999). 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 homeostatis. It often takes a large among of knowledge to identify such clashes in knowledge. Miyake and Norman (1979) presented the argument 20 years ago that "to ask a question, one must know enough to know what is not known." Otero and Graesser developed a set of production rules that specifies the categories of questions that are asked under particular conditions (i.e., content features of text and knowledge states of individuals).

Questions that tap explanatory reasoning are particularly diagnostic of deep comprehension. Explanations are needed when devices break down, faults are diagnosed, and devices are repaired. The person responsible for a broken piece of equipment needs to construct explanations in the form of causal networks, goal-plan-action hierarchies, and logical justifications. It is well documented that the construction of explanations is an excellent (if not the best) predictor of adults' abilities to learn technical material from written texts (Chi, deLeeuw, Chiu, & LaVancher, 1994; Cote, Goldman, & Saul, 1998).

Question asking tasks have the potential for improving the accuracy of personnel selection and classification. For example, a sailor would ideally be assigned to be a locksmith if the sailor has deep knowledge that explains lock mechanisms, but not if the sailor merely knows the jargon. But how does one know whether a sailor has the talent and the deep knowledge for a task? We know that we will not get much useful information by simply asking the sailor (e.g., "How good are you in operating a lock?"). There are serious limitations in the metacognitive abilities of adults in monitoring the accuracy of their own comprehension (Hacker, Dunlosky, & Graesser, 1998). We know that we will not get much useful information by testing the sailor on inert shallow knowledge, such as a test of vocabulary and technical jargon (e.g., "What is cam?"). We know that it would be impractical to spend several years developing a fully validated, reliable, psychometric test on each device in the military. The device would be outdated by the time the psychometric test was finished.

The present project investigated the questions that college students ask when an everyday device malfunctions. After reading about a device (e.g., cylinder lock, dishwasher), the participants subsequently received scenarios in which the device breaks down (e.g., *the key turns but the bolt doesn't move*, in the context of a cylinder lock) and they generated questions about the malfunction. Eye tracking data were also collected at that time. There were a number of straightforward predictions of the PREG model. First, those participants who have a deep understanding of the device should ask good questions that converge on faults. Second, the eye movements of deep comprehenders should quickly converge on likely faults that explain the breakdown. This paper briefly summarizes the highlights of two studies that confirm these predictions.

Question Asking and Deep Comprehension of Devices

In a recent study by Graesser et al. (in press), college students (N = 108) at the University of Memphis first read an illustrated text, then were given a breakdown scenario, and then generated questions. After completing the question asking task, they are given a comprehension test on the devices. Finally, the participants completed a battery of tests of cognitive ability and personality.

Illustrated Texts and Tasks. The participants read 6 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 book with illustrated texts, *The Way Things Work* (Macaulay, 1988). After reading about each device, the participants subsequently received scenarios in which the device breaks down. During this time, the participants were asked either to "think aloud in writing" (which we will call the "write aloud" task) or to generate questions in writing ("question asking" task) for three minutes. The present report focuses exclusively on the question asking data. The participants typically reflected on how to diagnose and repair the malfunctions during the question asking task.

Device Comprehension Test The participants completed an objective test on their deep understanding of the devices. This consisted of six 3-alternative, forced-choice questions about each device (36 total questions across the 6 devices, so scores could vary from 0 to 36). There were 4 test questions per device that tapped explicit information and 2 questions that tapped inferences. An example of an inference question is provided below.

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)

All 36 questions followed a "qualitative physics" framework that was designed to tap deep comprehension. That is, suppose there are N components in the device. If one component C1 is damaged or changed, what is the impact on another component (C2) in the device? A state associated with C2 can either increase, decrease, or stay the same. It takes deep understanding to answer such constraint propagation questions correctly. It should be noted that the scores on the device comprehension test served as the gold standard for deep comprehension in this study. If our PREG model is correct, than the quality of questions asked should predict device comprehension scores.

Battery of Tests of Individual Differences Participants completed a battery of tests that measured their cognitive abilities and personality. The tests of cognitive ability include the ASVAB (the Armed Services Vocational Aptitude Battery, Department of Defense, 1983). This test is administered to over 1 million high schools students each year. There were the following subscales on this test: Mechanical comprehension, electronics, general science, auto & shop, mathematics knowledge, arithmetic reasoning, numerical operations, word knowledge, paragraph comprehension, and coding speed. Five composite variables can be derived from the 10 measured variables on the ASVAB: technical scientific knowledge, verbal ability, numerical ability, coding speed, and general intelligence (g). 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). A number of noncognitive variables were measured. These included age, gender, and scales on a personality test. The personality test is the NEO inventory (Costa & McCrae, 1991), which measures individuals on the "big five" personality factors: neuroticism, extroversion, openness, agreeableness, and conscientiousness. It took approximately 4 hours to complete the battery of tests, which were completed in two sessions on two different days.

Results.

The most accurate measure of deep comprehension was the device comprehension score. The mean score was 23.5 out of 36 questions (SD = 5.3). According to our hypothesis, we would expect the device comprehension scores to show a high positive correlation with the questions that were asked during the breakdown scenario. We in fact did find a significant

positive correlation between device comprehension and question quality (r = .51, p < .05). Question quality was defined as the proportion of questions that referred to a plausible malfunction that explained the breakdown. Question quality had a substantially higher correlation than did the mere quantity of questions, the quantity of ideas in the write aloud task, and the quality of ideas generated in the write aloud task.

Question quality compared very well with the general measures of cognitive ability and the noncognitive factors. The bivariate correlations with device comprehension were either small or nonsignificant for the 5 personality measures, age, working memory, exposure to print, and many of the ASVAB measures. The correlations with technical scientific knowledge (r = .72, p < .05) fared better than question quality, but all other measures of ASVAB (and also spatial reasoning) had approximately the same or lower correlations with device comprehension than did question quality. Males had significantly higher device comprehension scores than females (r = .40, p < .05), but the correlation was not as high as question quality. When we performed follow-up multiple regression analyses, we found that technical knowledge was the primary predictor of both question quality and of device comprehension; all other cognitive and noncognitive measures were not significant.

Technical scientific knowledge was robustly linked to device comprehension so we examined the differences between the questions that were asked by participants with high versus low technical knowledge. The questions asked by students with high scores had two characteristics: (a) the questions converged on components in the mechanism that are plausible faults and (b) the questions had a more fine-grained elaboration of the parts, processes, and relations that specify how the breakdown occurred. Stated differently, there was high convergence on plausible faults and high mechanistic detail.

We have mapped out conceptual graph structures for the illustrated texts on devices. These structures include component hierarchies, spatial region hierarchies, causal chains/networks, goal/plan/action hierarchies, and property descriptions that are depicted in either text or picture form (Baggett & Graesser, 1995; Graesser et al., 1992). We have identified the content in the conceptual graph structures that is relevant versus irrelevant to the breakdown scenarios. The content of the 108 students' questions have been mapped onto the conceptual graph structures. This has allowed us to assess the extent to which properties of the participants' knowledge representations are predicted by cognitive abilities, personality measures, gender, and device comprehension test scores. However, it is beyond the scope of this paper to discuss what these detailed analyses have revealed.

Eye Movements during Question Asking

At this point, no one has systematically analyzed the relationships between eye movements and the cognitive components in a model of question asking. We conducted a second study that tracked eye movements on college students (N = 28) who asked questions in the context of the breakdown scenarios. The college students first read each illustrated text on everyday devices, followed by a breakdown scenario for 90 seconds (while the illustrated text remained on the screen). The participants generated questions about the breakdown scenario during the 90 seconds and eye movements were recorded by a Model 501 Applied Science Laboratory eye tracker. .

According to the PREG model of question asking, we would expect deep comprehenders to show a high density of eye fixations at words, objects, parts, and processes that are at the source of cognitive disequilibrium (e.g., anomalies, contradictions, broken parts, contrasts, missing components, and so on). It should take a sufficient amount of technical knowledge to detect such irregularities in the system. That is, there should be a correlation between technical knowledge and the proportion of fixations that are on faults that explain the breakdown. An area plot displays the amount of time that the eye fixates at each region in an N x M dimensional grid. The area of interest is the subset of the display that should theoretically receive fixations (e.g., the faults).

Our analysis of the eye tracking data confirmed our expectation. The proportion of fixation time on likely faults (that explained the breakdown) was significantly higher for participants who had a relatively high number of good questions (.13 versus .09 for high versus low), for those who had relatively high device comprehension scores (.13 versus .09), and for those who had high general science scores on ASVAB (.13 versus .08); other measures of individual differences did not significantly predict the proportion of fixation times on faults.

Precisely the same results occurred when measuring the number of fixations on faults. As predicted by PREG, participants with high technical knowledge scores had a higher proportion of good questions. In a follow-up analysis, we discovered that high ability students tended to fixate on faults during the 3-second time span that preceded the question about the fault. So they see the fault and then the question emerges from their linguistic production mechanisms.

In closing it appears that we have two quick tests of whether a sailor has deep knowledge about a particular device. In both tests, we present a breakdown scenario that puts the sailor in cognitive disequilibrium and forces a problem solving mode. One test is that they will generate good questions that tap likely causes of the breakdown. The second test is that their eyes will fixate on the faults. In contrast, the poor comprehenders have questions that are not discriminating and their eyes move all over the display. In less than 2 minutes, we can identify whether a particular sailor has the deep knowledge and talent for understanding a particular device.

Acknowledgements

This research was funded by the Office of Naval Research (N00014-98-1-0331). We thank Elisa Cooper, Victoria Pomeroy, and Shannon Whitten for collecting and analyzing data on this project.

References

Baggett, W.B., & Graesser, A.C. (1995). Question answering in the context of illustrated expository text. Proceedings of the 17th Annual Conference of the Cognitive Science Society (pp. 334-339). Hillsdale, NJ: Lawrence Erlbaum.

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 curiousity. New York: McGraw-Hill.

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.

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.

Cote, N., Goldman, S.R., & Saul, E.U. (1998). Students making sense of informational text: Relations between processing and representation. Discourse Processes, 25, 1-54.

Department of Defense (1983). Armed Services Vocational Aptitude Battery, Form 12a. Washington, D.C.: Department of Defense.

Festinger, T.J. (1957). A theory of cognitive dissonance. Evanston, IL: Row and Peterson.

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

Graesser, A.C., Baggett, W., & Williams, K. (1996). Question-driven explanatory reasoning. Applied Cognitive Psychology, 10, S17-S32.

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

Graesser, A.C., Olde, B., Pomeroy, V., Whitten, S., Lu, S., & Craig, S. (in press). Inferences and questions in science text comprehension. In book edited by J. Otero and M. Helena (Eds.), Science text comprehension.

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

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

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.
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. (1999). PREG: Elements of a model of question asking. Unpublished manuscript submitted to Cognition & Instruction.
Schank, R.C. (1999). Dynamic memory revisited. Cambridge: Cambridge University Press.
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.