THE USER'S MENTAL MODEL OF AN INFORMATION RETRIEVAL SYSTEM

Christine L. Borgman
Graduate School of Library and Information Science
University of California, Los Angeles

ABSTRACT

An empirical study was performed to train naive subjects in the use of a prototype Boolean logic-based information retrieval system on a bibliographic database. Subjects were undergraduates with little or no prior computing experience. Subjects trained with a conceptual model of the system performed better than subjects trained with procedural instructions, but only on complex, problem-solving tasks. Performance was equal on simple tasks. Differences in patterns of interaction with the system (based on a stochastic process model) showed parallel results. Most subjects were able to articulate some description of the system's operation, but few articulated a model similar to the card catalog analogy provided in training. Eleven of 43 subjects were unable to achieve minimal competency in system use. The failure rate was equal between training conditions and genders; the only differences found between those passing and failing the benchmark test were academic major and in frequency of library use.

INTRODUCTION

In the search to understand how a naive user learns to comprehend, reason about, and utilize an interactive computer system, a number of researchers have begun to explore the nature of the user's mental model of a system. Among the claims are that a mental model is useful for determining methods of interaction [1,2], problem solving [2,3], and debugging errors [4]; that model-based training is superior to procedural training [2,5,6]; that users build models spontaneously, in spite of training [1,7]; that incorrect models lead to problems in interaction [4,7]; and that interface design should be based on a mental model [8,9]. Not surprisingly, these authors use a variety of definitions for "mental model" and the term "conceptual model" is often used with the same meaning. Young [10] was able to identify eight different uses of the term "conceptual model" in the recent literature, for example. This author prefers the distinction made by Norman [7] that a <u>conceptual model</u> is a model presented to the user, usually by a designer, researcher, or trainer, which is intended to convey the workings of the system in a manner that the user can understand. A <u>mental model</u> is a model of the system that the user builds in his or her mind. The user's mental model may be based on the conceptual model provided, but is probably not identical to it.

The first research comparing conceptual models to procedural instructions for training sought only to show that the conceptual training was superior [6,11]. Other recent research [1,2] has studied the interaction between training conditions and tasks, finding that model-based training is more beneficial for complex or problem solving tasks.

The research on mental models and training has been concentrated in the domains of text editing [11,12] and calculators [1,2,4,10]; no such research has yet been done in information retrieval. Information retrieval is an interesting domain, as it is now undergoing a shift in user population. In the last ten years, a significant population of highly-trained searchers who act as intermediaries for end users on commercial systems has developed. Although end users have been reluctant to use the commercial systems, libraries are rapidly replacing their card catalogs with online catalogs intended for direct patron use. The online catalogs are typically simpler to use and have a more familiar record structure, but still have many of the difficulties associated with the use of a complex interactive system. The result is a population of naive, minimally-trained, and infrequent users of information retrieval systems [13]. The need for an efficient form of training for this population is very great and we chose it as a domain to test the advantages of model-based training.

EXPERIMENTAL METHOD

The experiment was structured as a two-by-two design, with two training conditions (model and procedural) and two genders. All subjects were undergraduates at Stanford University with two or fewer programming courses and minimal, if any, additional computer experience.

We performed the experiment on a prototype Boolean logic-based online catalog mounted on a microcomputer with online monitoring capabilities. Two bibliographic databases were mounted: a training database consisting of 50 hand-selected records on the topic of "animals" and a larger database of about 6,000 records systematically sampled from the 10-million record database of the OCLC Online Computer Library Center.

Subjects in each training condition received three training documents: an introductory narrative, a set of annotated examples of system operation, and a table of searchable fields.

The introductory narrative provided to the model group described the system using an analogical model of the card catalog. The instructions first explained the structure of a divided (author/title/subject) card catalog and then explained the system structure in terms of the ways it was similar to a card catalog and the ways in which it was different. Boolean logic was described in terms of sets of catalog cards, showing sample sets and the resulting sets after specified Boolean combinations.

The narrative introduction for the procedural group consisted of background information on information retrieval that is commonly given in system manuals. The Boolean operators were defined only by single-sentence statements.

The examples provided were the same in each condition, but the annotations for each reflected the differences in the introductory materials. The list of searchable fields (16 of 25 fields were searchable) was also identical and gave examples of the search elements for each field.

The training tasks used for the benchmark test were all classified as <u>simple</u> tasks, requiring the use of <u>only</u> one index and no more than one Boolean operator. The experiment consisted of five simple and ten <u>complex</u> tasks, the latter requiring two or more indexes and one or more Boolean operators. All tasks were presented as narrative library reference questions and were designed to be within the scope of questions that might be asked by undergraduates in performing course assignments.

Subjects were given the instructional materials to read and then performed the benchmark test, which consisted of completing 14 simple tasks on the small database in less than 30 minutes. The test was based on pilot test findings that those who took longest to complete the training tasks were least able to learn to use the system ($r=-0.83$, $p<.05$). If the subject passed the benchmark test, he or she was interviewed briefly, given the experimental tasks to perform, and then asked to perform one additional search while talking aloud for the experimenter. Subjects were interviewed again after completing the experiment.

RESULTS

Due to a high failure rate on the benchmark test (11 of 43, or 26%), we were able to gather a valid dataset of only 28 cases. The difference in time required to complete the benchmark test was significant ($p<0.0001$), with those failing averaging 39.2 minutes and those passing averaged 18.2 minutes. Subjects failed equally in the two training conditions and by gender.

Subjects who passed the benchmark test tended to be from science and engineering majors rather than social science and humanities ($p<0.0001$), and were less frequent visitors to the library (average 8.0 visits per month vs. 18.4 visits for those who passed). Major and library use were not correlated.

In task performance, we found no difference between training conditions on number of simple tasks correct ($p>0.05$). The difference on number of complex tasks correct was in the predicted direction (subjects in the model condition scored higher than those in the procedural condition) but was not significant ($p=0.08$).

The user actions and system responses captured in the monitoring data were reduced to 12 discrete states and treated as a stochastic process. The patterns of interaction were measured using the two-sample Kolomogorov-Smirnov (K-S) test. On simple tasks, we found no significant differences between training conditions on any of zero-, first-, or second-order two-sample K-S tests ($p>0.05$ for each). On complex tasks, we found significant pattern differences between training conditions on each level ($p<0.01$ for zero-order; $p<0.001$ for first- and second-order tests).

The analysis of model articulation ability was based on four measures coded from the interview data: completeness of the model, accuracy of the model, level of abstraction, and use of a model in approaching the tasks. The first three variables were highly correlated, necessitating their combination into an index. We found no difference between conditions on either the model index or on the task approach variable.

If the subjects were able to describe the system's operation at all, it was most likely in terms of an abstract model bearing little resemblance to a card catalog analogy. Of 28 subjects, 15 (5 model condition, 10 procedural) gave some form of abstract model, four (3 model, 1 procedural) articulated a card catalog-based model, only one subject (procedural condition) articulated a model based on another metaphor (robots retrieving sheets of paper from bins), and eight subjects (6 model, 2 procedural) were unable to describe the system in any model-based manner.

Only minor differences between genders were found. Men scored higher than women ($p<0.05$) on the index of describing the system, although gender explained only 14% of the variance in the model index on a linear regression. Men were found to make more errors on simple tasks than women ($p<0.05$), but the difference was not significant for errors on complex tasks. On simple tasks, men and women reflected different patterns of use at all three levels of zero-, first-, and second-order transitions ($p<0.01$, 0.01, 0.001, respectively). On complex tasks, men and women also reflected different patterns of use at all three levels ($p<0.01$, 0.05, 0.01, respectively), although less strongly.

A more complete description of the results can be found in Borgman [14].

DISCUSSION

Perhaps the most surprising (and unpredicted) finding is the degree of difficulty encountered by some of the subjects in using the system. The system was similar to those in common use in libraries and the questions were similar to those an undergraduate might ask in seeking information for a course assignment. Yet more than one-fourth of the subjects could not complete 14 simple tasks in less than 30 minutes. The tasks were not difficult; nine of them were merely replications of the examples (which included the search result).

The subjects who had the most difficulty were those majoring in the social sciences and humanities. It has frequently been conjectured that this group might have more difficulty using computing technology, but hard evidence is difficult to establish [15]. The effect is not explained by measures commonly associated with major, such as number of math and science courses or number of computing courses.

It is doubtful that academic major alone is the factor determining success or failure at the information retrieval task. It is more likely that academic major is a surrogate for some other measure. Related research in human factors of computing has begun to identify psychological and skill factors that influence computing ability, such as cognitive style [15], spatial memory, and age [16]. The pattern differences between men and women also suggest that some individual differences may be operating. The individual differences issues are of particular concern for online catalogs in library environments, most of which serve a very heterogeneous population. Given the minimal control that system administrators have over training this class of users, it is important that the system be easily accessible by a broad population.

Another factor that distinguished those who passed the benchmark test from those who failed was frequency of library usage. The result is in the opposite direction of that which would be predicted: the frequent library users failed and the infrequent ones passed. If frequency of library usage were correlated with major, this result would be easier to explain. However, we can say that frequent visits to the library (for whatever purpose) offer no advantage in learning to use an online catalog.

The performance differences were in the predicted direction, but less strong than we had hoped. However, the performance results were bolstered by the stronger pattern differences in the monitoring data: no significant differences on simple tasks but very significant differences on complex tasks. The pattern differences suggest at least a difference in method of interaction, if not a difference in cognitive processing. Given the nature of these results, the interaction effect, the small sample size, and the small number of tasks, we consider the hypothesis to be supported. We would be reluctant to generalize the findings beyond this sample, however.

The results of this research and that of Halasz & Moran [2] show that model-based training is superior only for complex or problem-solving tasks. Our next challenge is to delineate the distinction between simple and complex tasks and thereby isolate the factors that may cause such an interaction. These issues are left for future research.

The predicted differences in model articulation based on training condition were wholly unsupported. The problem may have been methdological; the questions to solicit the model appear to have been interpreted in a variety of ways. A more constructive explanation is that we may have captured the variance in who is able to <u>articulate</u> a model, rather than in who is able to <u>build</u> a model. It is possible that mental models were constructed in precisely the manner predicted, yet we were unable to capture this result. We can consider the presence of a model description sufficient to indicate that the mental model exists, but not a necessary condition. This interpretation is reinforced by the lack of correlation between task performance and model articulation.

Another interesting aspect of the model articulation results is the lack of correlation between ability to describe the approach to search tasks and ability to describe the system. Subjects were frequently able to describe their approach to performing searching tasks in terms of the system's operation, but were unable to describe the same operations when asked how the system worked. It is possible that the questions solicited two types of models. The model used in problem solving (which results in performance effects) may be different from the model used in describing the system. According to Halasz [3], these two types of models may occur in sequence: one first builds a model for problem solving and only after practice is able to explain how it works. This interpretation is reinforced by the fact that no subject was able to describe the system but not able to describe his or her approach to the tasks.

One last possibility is that the amount of time spent in training and system use was insufficient to develop the model. Models develop over time with exposure to the system. Given further practice, stronger results might have been seen.

CONCLUSIONS

The present study compared the use of conceptually-based training to that of procedurally-based training on a prototype online catalog. Although the training effects were not as strong as predicted, we did find the hypothesized interaction effect between training method and task complexity, indicating that conceptually-based training is not always superior. The challenge of delineating when it is superior remains.

As expected, we found that it is easier to measure differences in who is able to articulate a model than in who is able to build a model. Subjects in both conditions were able to develop models to some degree, indicating that people do build models even if not trained with them. The fact that no relationship was found between model articulation and performance further suggests that the measures captured articulation ability only.

Perhaps the most important finding from this experiment is not the mental models result but the likelihood of individual differences in the ability to use this particular technology. Given an equal number of math, science, and computing courses, engineering and science majors still out-performed the social science and humanities majors. This finding suggests that we may be building systems for which access is inequitable. We are particularly concerned about this result in library environments, where equal access to information for all is a primary goal of the institution. If the implementation of a new technology discriminates among our users, we must find a way to achieve equity through training, design, or additional assistance.

The research reported here is the first in what is intended to be a continuing research program. The second phase, to study the individual differences correlates of technology use, is already in progress [17]. New results from the later research will be incorporated in the conference presentation. It is our hope that this research will contribute not only to our understanding of human-computer interaction, but also to improving equity in access to information technology.

ACKNOWLEDGEMENTS

The research reported here was funded by the OCLC Online Computer Library Center, Dublin, Ohio. The interface simulator was developed and implemented by Howard Turtle and Trong Do, under the direction of Neal Kaske and W. David Penniman. The author also is grateful for the assistance of her dissertation advisor, William Paisley, and the other members of her committee, Everett M. Rogers, David A. Thompson, and Barbara Tversky, all of Stanford University.

REFERENCES

[1] Bayman, Piraye; Mayer, Richard E. 1984. Instructional manipulation of users' mental models for electronic calculators. International Journal of Man-Machine Studies, 20, 189-199.

[2] Halasz, Frank G.; Moran, Thomas P. 1983. Mental models and problem solving using a calculator. In Janda, Ann (ed.), Human factors in computing systems: Proceedings of a conference sponsored by the Association for Computing Machinery Special Interest Group on Computer and Human Interaction and the Human Factors Society. 1983 December 12-15, Boston, MA. New York, NY: Association for Computing Machinery, 212-216.

[3] Halasz, Frank G. 1984. Mental models and problem solving using a calculator. Ph.D. dissertation. Stanford, CA: Stanford University.

[4] Young, Richard M. 1981. The machine inside the machine: Users' models of pocket calculators. International Journal of Man-Machine Studies, 15, 51-85.

[5] Carroll, John M.; Thomas, John C. 1982. Metaphor and the cognitive representation of computing systems. IEEE Transactions on Systems, Man, & Cybernetics, SMC-12:2, 107-116.

[6] Foss, Donald J.; Rosson, Mary Beth; Smith, Penny L. 1982. Reducing manual labor: An experimental analysis of learning aids for a text editor. In Association for Computing Machinery, Proceedings of the human factors in computer systems conference. 1982 March 15-17, Gaithersburg, MD. New York, NY: Association for Computing Machinery, 332-336.

[7] Norman, Donald A. 1983. Some observations on mental models. In Gentner, Dedre; Stevens, Albert L. (eds.), Mental models. Hillsdale, NJ: Lawrence Erlbaum Assoc.

[8] Jagodzinski, A. P. 1983. A theoretical basis for the representation of on-line computer systems to naive users. International Journal of Man-Machine Studies, 18, 215-252.

[9] Moran, Thomas P. 1981. The command language grammar: A representation for the user interface of Interactive systems. International Journal of Man-Machine Studies, 15, 3-50.

[10] Young, Richard M. 1983. Surrogates and mappings: Two kinds of conceptual models for interactive devices. In Gentner, Dedre; Stevens, Albert L. (eds.), Mental models. Hillsdale, NJ: Lawrence Erlbaum Assoc.

[11] Mack, Robert L.; Lewis, Clayton H.; Carroll, John M. 1983. Learning to use word processors: Problems and prospects. Association for Computing Machinery Transactions on Office Information Systems, 1:3, 254-271.

[12] Douglas, Sarah A.; Moran, Thomas P. 1983. Learning text editor semantics by analogy. In Janda, Ann (ed.), Human factors in computing systems: Proceedings of a conference sponsored by the Association for Computing Machinery Special Interest Group on Computer and Human Interaction and the Human Factors Society. 1983 December 12-15, Boston, MA. New York, NY: Association for Computing Machinery, 207-211.

[13] Matthews, Joseph R.; Lawrence, Gary L.; Ferguson, Douglas K. 1983. Using online catalogs: A nationwide survey. New York, NY: Neal Schuman.

[14] Borgman, Christine L. 1984. The user's mental model of an information retrieval system: Effects on performance. Unpublished PhD dissertation, Stanford University.

[15] Coombs, J.J.; Gibson, R.; Alty, J.L. 1982. Learning a first computer language: Strategies for making sense. International Journal of Man-Machine Studies, 14:4, 449-486.

[16] Egan, Dennis E.; Gomez, Louis M. 1984. Assaying, isolating, and accommodating individual differences in learning a complex skill. In Dillon, Ronna F. (ed.), Individual differences in cognition, Vol. 2. New York, NY: Academic Press.

[17] Borgman, Christine L. 1984. Individual differences in learning to use a library online catalog: Pilot Project. Research project funded by the Spencer Foundation.