Government Information Expert Systems: A Quantitative Evaluation

John V. Richardson Jr. and Rex B. Reyes

In this article—the first published quantitative evaluation of knowledge-based systems (KBS) or so-called expert systems—the authors quantitatively compare and contrast two systems: POINTER and Government Documents Reference Aid (GDRA). In a test based on fifteen typical U.S. government document reference questions about the federal level of government, POINTER answered 65 percent of the questions correctly while GDRA answered only 37 percent correctly. An analysis of keystroke efficiency revealed that POINTER required 120 strokes in the reference interview and 60 for the question negotiation phase while GDRA needed 120 keystrokes in the reference interview but only 45 during its question negotiation. The discussion and implication section should help developers of knowledge-based computer systems focus their future activities in this area and reassure human reference librarians who work with government information that these systems still have a way to go before they are truly competent systems. Nonetheless, the first generation of expert systems for depository libraries could already be playing a widespread, if modest, role in assisting with federal level reference questions.

riting in 1964, Jesse Shera argued that the "fullest utilization of the potential of automation [such as expert systems in reference work] necessitates a thorough study of the total reference process—from the problems that prompt the asking of a question to the evaluation of the response."[1] Hence, the overarching goal of the following study is to contribute to the profession's understanding of the total process by evaluating reference question responses in the field of government information.

There are several microcomputer knowledge-based or so-called expert systems whose coverage includes the field of government information. Of the systems that specifically emphasize this area of specialization, the best known is POINTER, which was developed by Karen F. Smith at SUNY, Buffalo, in 1984. Four years later on the West Coast, Bruce Harley and Patricia Knobloch developed Government Documents Reference Aid (a.k.a. GDRA) at Stanford University. In each case, the computer system attempts to answer reference questions much the way a government documents specialist

---

John V. Richardson Jr. is Associate Professor at the Graduate School of Library and Information Science at the University of California, Los Angeles, 300 Circle Drive North, Suite 204, 405 Hilgard Avenue, Los Angeles, California 90024. He can be reached at (310) 206-9369 or via Internet at IBQ1JVR@MVS.OAC.UCLA.EDU. Rex B. Reyes is Reference Librarian at Western State University, College of Law, Fullerton, California 92631.
The authors wish to thank Terry Crowley of San Jose State University for informally discussing the methods, results, and implications of this study as well as reviewing an early draft of the article; Zorana Ercegovac of UCLA for her discussion of response scoring; and Matthew Schall of Tulane University's Department of Psychology (formerly of UCLA's Office of Academic Computing) for statistical consulting.

might—by referring the user to a single source or even several sources that are thought likely to contain the answer. To the best of the authors' knowledge, these are the only available systems in government information.[2]

Although the first of these computer systems has been extant for ten years, no one has examined systematically the quality or accuracy of these systems. Harley and Knobloch infer that their system is expert while Smith is careful to qualify user expectations of her system: "POINTER is not an expert system. POINTER is a computer-assisted reference program—inspired by expert system developments of the recent past, and aspiring to be upgraded to a real expert system in the future."[3] Nonetheless, it is not clear how much assistance users can expect from these systems nor how much future development work may be necessary for these systems to be truly expert in the human sense.

Hence, the authors believe that a quantitative evaluation of the quality of these extant systems needs to be undertaken. To the best of their knowledge, no such published study exists; hence, this article is an original contribution to understanding the nature of expertise in these systems. As a result, the authors have established a method for benchmarking these systems for the first time. Using this methodology, readers can judge for themselves the technological promise of expert systems.

SYSTEMS, EVALUATION ISSUES, AND GENERAL REFERENCE STUDIES

To keep this research project within manageable limits, the scope of this study involves the following three dimensions: (1) the extant microcomputer systems, (2) evaluation issues, and (3) thirty years of reference quality studies.

Extant Microcomputer Systems

An expert system can be defined as: "a program that relies on a body of knowledge to perform a somewhat difficult task usually performed only by a human expert. The principal power of an expert system is derived from the knowledge the system embodies rather than from search algorithms and specific reasoning methods. An expert system successfully deals with problems for which clear algorithmic solutions do not exist."[4] The basic assumption underlying expert systems is the idea that "Knowledge Is Power."[5]

The quality of a knowledge-based system, much like human reference work, can be measured by a variety of factors, such as speed of response, subjective short- or long-term user satisfaction, the interface design (paralleling the question negotiation phase of the reference transaction), and, of course, the accuracy of responses.

As mentioned above, there are two expert systems that focus on the field of government information: (1) POINTER and (2) Government Documents Reference Aid (GDRA). For the purposes of subsequent discussion, the authors prefer the phrase *knowledge-based systems* because it more accurately describes the state of the art at this point.

POINTER.[6] Developed between 1984 and 1987 by Karen F. Smith, documents librarian at the Lockwood Library of SUNY, Buffalo; Stuart Shapiro, a SUNY Buffalo faculty member; and Sandra Peters, a computer science student, this system was originally written in LISP for a VAX minicomputer. It now runs on an IBM PC with a minimum of 256K RAM and a disk drive; there are 6,000 lines of BASIC code. The program is menu-driven and includes about 130 screens of text.

The work of the program developers is based on an analysis of 1,071 queries in the university library's documents department, and took four months to develop with a $6,000 investment, including a $3,000 Council on Library Resources' Faculty and Librarian Cooperative Research Grant.[7] The perceived benefits of this particular system are twofold: (1) a solution to lack of staff, and (2) a training tool for student assistants and clerical

staff. User interaction (i.e., the reference interview) with POINTER begins with a welcome screen, asking whether the user wants to continue. If the user types yes or y, the next screen describes the organization of their collection according to the SuDoc classification scheme. The system asks the user if s/he has a SuDoc number; the three acceptable responses are: yes, no, or unsure. "Yes" refers the user directly to the collection. "No" or "Unsure" provides more information about SuDoc numbers to help the user decide if s/he has a SuDoc number. Strictly speaking, the interaction thus far is not part of question negotiation. Nonetheless, the system forces the user to answer these questions as part of the preliminary phase of the reference interview. Next, the user progresses to a menu screen with four options. This screen is the first in the real question negotiation phase. For the purposes of this study, the authors assume that the user does not have a SuDoc classification number. Question negotiation ends when the last screen of sources appears. The reference interview concludes after the user responds to the prompt "Do you have another question?" To exit the program, the user must press "control-break."

Government Documents Reference Aid (GDRA).[8] Created in 1988 by Bruce Harley and Patricia Knobloch, then of Stanford University Libraries (SUL), GDRA was developed on an IBM AT using Level 5 shell software, which employs production rules and backward chaining logic. The Payson J. Treat Fund provided $1,265 and its development required about four weeks. Its library environment is SUL and covers U.S. federal, state, and local as well as foreign and international (including United Nations) government publications.[9] Special features provide that "both ASCII text files and an external program are directly accessed. The external program, Samson, provides the telecommunications link to Socrates, SUL's online catalog, activated from with GDRA's rule structure."[10] There are four perceived benefits of GDRA: it "solve[s] the problem of increasing workload; contributes to the mainstreaming of government documents with SUL; helps train staff providing government documents reference service; and supplements existing government documents reference service."[11]

User interaction with GDRA progresses according to the following pattern: the Level 5 shell screen appears; six options, selected by using the arrow keys, are presented. Novices will not know where to start; however, Info, Intro, or Main Menu are the most obvious choices. The correct place is the introduction module; the authors consider the reference interview to start here. The Main Menu is where question negotiation starts. During the question negotiation phase, there is a compulsory information screen about "U.S. Federal Documents" that discusses the SuDoc shelving arrangement at Stanford. Question negotiation ends when all the sources appear (i.e., the screen labeled "Subject, Author, Title"). The reference interview ends when the user presses the "F2" function key to return to the welcome screen. The "F10" function key allows the user to exit GDRA.

Evaluation Issues

Naturally, the question arises: What constitutes a good system? The quality of a knowledge-based system, much like human reference work, can be measured by a variety of factors, such as speed of response, subjective short- or long-term user satisfaction, the interface design (paralleling the question negotiation phase of the reference transaction), and, of course, the accuracy of responses. In this study, the authors investigate two aspects of quality: efficiency and accuracy. Further, the authors defined efficiency as the number of keystrokes that the user has to type. Operationally, the authors defined accuracy as the percentage of questions correctly answered out of a set of fifteen test questions. In this respect, since the authors presume that these systems are serving as surrogates for real reference librarians, it seems reasonable that the competence of such systems should be addressed in this manner.

Thirty Years of Reference Quality Studies: The Theoretical Bridge

The authors knew only what the general nature and extent of the extant microcomputer-based systems in government information were, and they wanted to know more about their quality. What the authors needed was a link between the known and the unknown. Thus, they propose to model this study of machine-based reference work on the prior thirty years of human reference quality studies. Of course, there have been some difficulties in undertaking such studies of accuracy; notably, the literature does not report the most frequently asked government information questions.[12] Rather than undertake that subject as the focus of their work, the authors will adopt those questions that have already been worked out by other researchers. They assume that there is a kind of comparability with the studies of general reference quality because imbedded in many of their test questions are questions that, in fact, can be answered in documents departments and are documents-type questions.

KEY OBJECTIVES AND RESEARCH QUESTIONS

To be explicit, the three key research objectives of this article are: (1) to depict how well each knowledge-based system performs; (2) to compare and contrast each system; and (3) to test the null hypotheses laid out below. Logically, three research questions flow from these objectives: first, can the user get a correct answer from either POINTER or GDRA? Second, compared to each other, how well do POINTER and GDRA perform in percentage terms? Lastly, and most importantly, how well do they perform against reports of human reference experts?

Answers to these questions can help knowledge-based system developers focus their activities and provide a method of benchmarking the state of the art in knowledge-based systems for government information.

PROVISIONAL HYPOTHESES

The authors propose the following two hypotheses, one about the system's accuracy and the other which addresses its efficiency. Together these hypotheses address the quality issue of a knowledge-based system for answering requests for government information.

The Accuracy Hypothesis

There is no difference between the performance of these two knowledge-based systems and the reported literature accuracy rate of 52 to 65 percent success in real reference settings. The 13 percent variability occurs because unobtrusive studies have reported lower levels of success than obtrusive ones. Given the existence of this range, the authors contemplated establishing a similar confidence interval for these two knowledge-based systems under study by using strict or more liberal responses during each system's question negotiation session (see method section below).

More fundamentally, the authors believe that the present state of the art in this new technology is still first generation. While the authors are optimistic about the long-term future of this technology, they suspect that, at present, there is a serious need for further development work (essentially, more time and money needs to be spent in this area) for real results in knowledge-based systems that can deal with question answering.

The Efficiency Hypothesis

There is no difference in the efficiency between the two systems during the reference interview or question negotiation phase. As mentioned above, the authors defined *efficiency* as the number of keystrokes that the user had to type. By *reference interview* the authors mean the entire interaction with the knowledge-based system. The *question negotiation* phase is just the interaction which addresses the inquiry (i.e., not, as in POINTER, the background information on the SuDoc classification scheme or, as in GDRA, the description of the Stanford collection).

METHOD

This section addresses four concerns: (1) defining the population of test questions, (2) selecting a training set of three

questions and drawing a representative, random sample of test questions, (3) modeling the obtrusive nature of prior studies, and (4) evaluating system response and scoring.

Population of Test Questions

As mentioned above, the authors based their own evaluation of these two knowledge-based systems upon the more than thirty years of general reference quality studies.[13] From these numerous studies, the authors selected those that actually reported the real questions they used in measuring the quality of human reference service: Charles Bunge (1968), Thomas Childers (1971), Terry Crowley (1971), Jassim Jirjees (1983), Marcia Myers (1983), Charles McClure and Peter Hernon (1983 and 1987), and Kathy Way (1987).[14] Interestingly, only the reported studies of McClure and Hernon focused solely upon federal government publication–type questions. From the remaining studies, the authors identified just those questions which could be answered using federal level government publications. The final pool consisted of eighty questions.

Training Set and Random Sample

The authors trained together on a set of three randomly selected test questions (i.e., Hernon and McClure's number 1; Jirjees' number 3; and Myers' number 2; see appendix A). The design was to act as two independent judges, reviewing the quality of each system. The authors worked independently with each system and then came back together to compare their findings. When differences emerged in recording the results, the authors reached a consensus by discussing how they interpreted each system's prompts and then the authors agreed on the proper path (those questions are marked in appendix B with an asterisk to indicate their initial disagreement).

Next, the authors randomly selected a total of fifteen test questions devoted to the U.S. federal level based on each of the seven studies with each study proportionately represented.[15,16] In terms of difficulty (i.e., time to answer a question), the authors assumed that each question was of equal difficulty.[17]

Modelling Prior Studies: Liberal versus Conservative Approach

The authors considered the obtrusive versus unobtrusive nature of the previous reference studies. They finally adopted one approach to the knowledge-based system interface and its question negotiation. Using the set of fifteen test questions, the authors were generous in their analysis. This liberal approach would be similar to the way a familiar user of government publications would respond to the environment. Such a user is willing to use a computer, read an entire screen full of information, and thoughtfully select menu items after considering all the options. This approach is the best case scenario. It more closely models the obtrusive nature of the previous research on reference quality. The authors wanted to see how capable these knowledge-based systems are in answering questions accurately.

Evaluating System Response and Scoring

At the outset the authors reviewed the accuracy scoring methods that have traditionally been used. Historically, many of the previous studies of reference quality have scored the results as a dichotomous variable—either the question was answered or not (i.e., most report the percentage of correct answers).[18] Arguably, the ideal response for a fact-type question is a single source which contains the complete and correct answer. In this case, previous investigators often gave one point for the correct answer and no points for an incorrect one. Further, some used a test set of ten questions to make the math involved more straightforward. Obviously though, the real world of reference work is more complex than that—a range of responses is possible and extreme values can occasionally occur.[19] So more recent investigators such as Cheryl Elzy, Alan Nourie, Wilf Lancaster, and Kurt Joseph (1991)

have reconsidered this response variable; they implicitly recognize it as continuous.[20] In this study, the authors explicitly recognized the response range as continuous in developing their own scoring method (see table 1).

Next, the authors assigned point values, creating an eight-point response scheme and added qualitative judgments related to the level of service provided. In the authors' estimation, this scheme more adequately reflects reality. In fact, the above-named investigators agree with the authors that it would be appropriate to "give minus values to inappropriate referrals. . . ,"[21] but they did not do so in their particular study. The present authors do so because they believe that wrong answers significantly penalize users and create ill will. Hence, the authors' method does not artificially restrict the range of responses and takes into consideration the possibility of extreme values as well.

Finally, to measure efficiency, the authors counted keystrokes for both systems. They counted the total number of keystrokes from the beginning to the end of the interaction as the "reference interview." They counted the prompt "Do you have another question?" as the end of the reference interview for POINTER; for GDRA, the reference interview ended when the F2/F3 option appears allowing the user to start at the beginning or just at the Main Menu. For question negotiation, the authors started from the numbered menu option in POINTER and from the Main Menu in GDRA. They did not count the compulsory information screen in GDRA nor did the authors count the offer of help with the SuDoc classification scheme in POINTER. Statistical analysis was supported by SAS, Version 6.08, running on an IBM Series 9000/900 mainframe. During data screening, a univariate analysis confirmed: (1) data points are not missing, (2) data are not demonstrably nonnormal (as measured by skewness and kurtosis of less than two for the experimental variables), and (3) no data outliers except as discussed below.[22]

FINDINGS

The authors can confidently answer their first research question straight away—yes, the user can get a correct answer some of the time. However, the systems vary in their ability to do so.

TABLE 1
TAXONOMY OF SYSTEM'S POTENTIAL RESPONSES

Score	Range of System's Response	Service Quality
5.0	Referred to a single source, complete and correct answer	Excellent
4.0	Referred to several sources, one of which gave complete and correct answer	Very good
3.0	Referred to a single source, none of which leads directly to an answer but one of which serves as a preliminary source	Good
2.0	Referred to several sources, none of which leads directly to an answer but one of which serves as a preliminary source	Satisfactory
1.0	No direct answer; referred to specific person/institution	Fair/poor
0.0	No answer; no referral (e.g., I don't know)	Failure
-1.0	Refered to a single inappropriate source	Unsatisfactory
-2.0	Referred to several sources, none of which answers	Most unsatisfactory

Source: Suggested by Gers and Seward (1985) and Elzy, Nourie, Lancaster, and Joseph (1991).

TABLE 2
SCORING OF POINTER AND GDRA ON THE FIFTEEN TEST QUESTIONS

Question	Pointer's Score	GDRA's Score	Total Possible
1.	4	2	5
2.	2	2	5
3.	2	2	5
4.	4	2	5
5.	4	2	5
6.	3	3	5
7.	2	2	5
8.	2	2	5
9.	4	2	5
10.	4	2	5
11.	2	2	5
12.	4	-1	5
13.	4	2	5
14.	4	2	5
15.	4	2	5
Grand total	49 (65.33%)	28 (37.33%)	75 (100%)
Mean score	3.266	1.866	Per question

POINTER Does a Better than Satisfactory Job

Overall, POINTER scored a total of 49 out of 75 possible points (or 65 percent of the federal level fact-type questions asked of it). The average score was 2.3 points per question. Based on table 1, that means POINTER is doing a good job in the authors' qualitative judgment. Parenthetically, see table 2 for the actual scores on each question. An analysis of efficiency (defined as the number of keystrokes) reveals that POINTER required 120 strokes during the reference interview and 60 for the question negotiation phase (see table 3). A Pearsonian correlation between POINTER's accuracy score and the total number of keystrokes for POINTER's question negotiation was -.237 ($t = .88$, $df = 13$, and $p = .39$). In other words, there is no significant correlation between more extensive question negotiation and higher accuracy in this knowledge-based system.

TABLE 3
KEYSTROKE EFFICIENCY OF POINTER AND GDRA ON THE FIFTEEN TEST QUESTIONS

	POINTER		GDRA	
Question	RI	QN	RI	QN
1.	8	4	8	3
2.	9	5	8	3
3.	6	2	8	3
4.	7	3	8	3
5.	8	4	8	3
6.	8	4	7	2
7.	11	7	8	3
8.	8	4	8	3
9.	8	4	8	3
10.	7	3	8	3
11.	8	4	8	3
12.	8	4	9	4
13.	7	3	8	3
14.	7	3	8	3
15.	10	6	8	3
Grand total	120	60	120	45
Keystrokes				
Mean	8.0	4.0	8.0	3.0
Median	8.0	4.0	8.0	3.0
Standard deviation	1.25	1.25	.37	.37

Note: RI = reference interview; QN = question negotiation

GDRA Is Doing an Almost Satisfactory Job

GDRA scored a total of 28 out of 75 possible points (or 37 percent of the federal level fact-type questions asked of it). The average score was 1.9 points per question. Based on table 1, that means that GDRA is doing a nearly satisfactory job in the authors' qualitative judgment. For a detailed analysis of scoring by question, see table 2. GDRA needed 120 keystrokes in the reference interview but only 45 during its question negotiation. A Pearsonian correlation between GDRA's accuracy score and the total number of keystrokes for GDRA's question negotiation was -.91 ($t = 7.7$, $df = 13$, and $p = .0001$). This time, there is significant correlation between more question negotiation and a lower score.

Comparison of the Two Systems

The second research question asked how these systems compared or contrasted. Neither system does an excellent job (i.e., earning five points in the scoring system), meaning that the user was referred to a single source that provided the complete and correct answer. Overall, though, POINTER is a better system for answering federal-level, fact-type government publication questions.

It may be useful to discuss particular questions where one system did much better or worse than the other. GDRA scored very poorly on question 12 (see Appendix B) because it recommended an inappropriate source and took more keystrokes in the reference interview as well as the question negotiation to achieve the wrong answer. The reason for this situation appears to be that the designers of GDRA did not anticipate users asking retrospective questions, specifically historical ones from the nineteenth century.

Hypotheses Testing

The first hypothesis proposed that there was no difference between the performance of these two knowledge-based systems and the reported literature rate of 52 to 65 percent success in real reference settings. The authors rejected the first part of this hypothesis. POINTER answered 65 percent of the test questions completely and accurately while GDRA answered only 37 percent of them. The second part of the hypothesis related their findings to the reported literature. POINTER matched the higher end of the reference studies while GDRA happened to match McClure and Hernon's 1983 reported findings about the performance of documents librarians.

Arguably, the ideal response for a fact-type question is a single source which contains the complete and correct answer.

Similarly, the authors rejected the second hypothesis that there is no difference in the efficiency between the two systems during the reference interview or question negotiation phase. POINTER required a total of 120 keystrokes (or 60 in the question negotiation phase) before recommending a source(s). On the other hand, GDRA also required 120 total keystrokes to answer the 15 test questions but only 45 in the question negotiation phase. In addition, there is an annoying inconsistency in the use of keystrokes during GDRA's interaction (e.g., sometimes one uses the function key while at other times it is the enter key that is used).

To test their qualitative observation that a modest increase in question negotiation doubles accuracy (i.e., POINTER scores 65 percent accuracy with 60 keystrokes versus GDRA's 37 percent with 45), the authors ran a logistic regression to model accuracy being equal to each knowledge-based system and question negotiation.[23] The chi-square for model fit with 2 degrees of freedom is 13.24, $p = .001$. The association of predicted probabilities and observed responses is concordant 86.6 percent, discordant 8.6 percent, and ties 4.8 percent. The chi-square suggests the model does not fit the data very well while the association of predicted probabilities suggests it

does. However, the power to detect significant differences is low and a larger *N* of test questions would be desirable in the future.

DISCUSSION AND IMPLICATIONS

Much of the preceding section treats the two knowledge-based systems (KBS) systems as a black box—i.e., mere input and output. More attention needs to be focused on the diagnostic issues; for example, why do these systems fail to perform at higher levels? Either system could score higher if it recommended fewer titles at the end of question negotiation. In an extreme case, POINTER recommended nine potentially relevant sources (for question numbers 1 and 15). The authors speculate that the naive user's confidence in the system's knowledge may be lessened by the large number of recommended titles. The authors' scoring system explicitly assumed that users want the single best source which completely and accurately answers their fact-type question.

Obviously, the two systems are still performing at a modest level, that is, they serve as reference systems (i.e., only referrals are given) rather than information systems (i.e., direct answers to the specific questions are given). Ideally, these systems should be able to give the user a direct answer to their question; this situation will most likely occur when these systems have a knowledge base similar to that of humans.

For the moment POINTER has a greater depth of knowledge about the federal level than does GDRA. To be a fully comprehensive system, POINTER ought to have GDRA's greater breadth of coverage. And, of course, in both of the systems under review, there is a substantial burden on the user rather than on the system.

Future Work

Subsequent investigations could take several directions in the future. One possibility is to make a more user-oriented evaluation of the knowledge-based systems. By that the authors mean that the typical user's accuracy as well as satisfaction with the interaction could be measured, either immediately or for the longer term; the authors hypothesize that it would be more in line with what the authors called a conservative approach (see above discussion).

Second, other useful work might involve the identification of the user's model of government information seeking or simply the user's model of the knowledge-based system. Then, one could compare and contrast their model with others such as the one presented by the government information textbook authors.[24]

Third, Cherie Weil's pioneering work at the University of Chicago also raises questions about the relationship of a knowledge-based system and the human reference expert.[25] Using 234 biographical sources, Weil found that while her knowledge-based system answered 10 out of 14 questions (71 percent) correctly and the human expert answered 11 out of 14 (79 percent) correctly, working together the human expert and the knowledge-based system could answer more questions correctly than either one working independently. Could the two KBS systems in this study serve a similar complementary support role for practitioners, especially general reference librarians who only occasionally answer government-publication-type questions?

A narrowly conceived line of future work would be a second pass through the fifteen test questions, taking a more strict or conservative approach, much as a naive user might. A naive user (i.e., one who knows relatively little about government publications or computer systems generally) might be willing to use a computer, but may not understand technical terms related to government information. Hence, the naive user might select, from a long menu, the first item that even looks applicable. In other words, s/he may not be willing to read an entire screen full of information. Such an approach may be said to emulate the unobtrusive approach.

Finally, the scope of analysis could be extended to other levels of government such as state, local, foreign, and interna-

tional/UN. At the present state of development, GDRA would excel POINTER at these other levels of government since POINTER only addresses the federal level.

CONCLUSIONS

This study has demonstrated that there is a need for improvement of knowledge-based systems in the government information field. For the purposes of subsequent research and discussion, the phrase *knowledge-based systems* should be used because it more accurately describes the present state of the art. The question of what role these systems should play needs to be examined in greater detail. Will knowledge-based systems be expected to serve the user in place of the reference librarian, or will they merely be used as supplementary help? The answer will depend on future study. Whatever the case may be, there is certainly a need to improve aspects of these systems, such as the breadth and depth of the knowledge base.

The authors' method of evaluating GDRA and POINTER can be replicated to judge the effectiveness of other knowledge-based systems, either in government information or in general question answering. The authors realize that there is still more research to be done regarding scoring techniques because quality and effectiveness may mean different things to different people. Because this study builds on the definitive studies of reference work, the authors believe their scoring method is a move in the right direction.

The authors believe that these knowledge-based systems have a place in the reference environment, especially in a time of budgetary constraints and staff shortages. In addition, at least one previous study demonstrates that the combination of a reference librarian and a KBS results in more accurate answers than either by themselves. When an overwhelming number of studies reveal that reference accuracy rates fall between 52 percent and 65 percent, automated solutions for the improvement of reference service certainly deserve further exploration.

REFERENCES AND NOTES

1. Jesse Shera, "Automation and the Reference Librarian," *RQ* 3 (July 1964): 3.
2. John Richardson, *Knowledge-based Systems for General Reference Work: Applications, Problems, and Progress* (San Diego: Academic, 1995).
3. Karen F. Smith, "POINTER: The Microcomputer Reference Program for Federal Documents," in *Expert Systems in Libraries*, ed. Rao Aluri and Donald E. Riggs (Norwood, N.J.: Ablex, 1990), 41.
4. Kamran Parsaye and Mark Chignell, *Expert Systems for Experts* (New York: Wiley, 1988), 1.
5. See Eliot Freidson, *Professional Powers: A Study of the Institutionalization of Formal Knowledge* (Chicago: Univ. of Chicago Pr., 1986) as well as Dennis H. Wrong, *Power: Its Forms, Bases, and Uses* (New York: Harper, 1979).
6. Karen F. Smith, Stuart Shapiro, and Sandra Peters, *Final Report on the Development of a Computer Assisted Government Documents Reference Capability: First Phase* (Buffalo: SUNY at Buffalo, 1984) ; Smith, "Robot at the Reference Desk?" *College and Research Libraries* 47 (Sept. 1986): 486–90; and "POINTER vs. *Using Government Publications*: Where's the Advantage?" *Reference Librarian* 23 (1988): 191–205. It is available for $30 from Karen Smith.
7. According to the sixth edition of the *Directory of Government Documents Collections and Librarians* (Bethesda, Md.: CIS, 1991), SUNY reports an extensive collection of federal and state materials and limited collections of local, international, and foreign documents.
8. Bruce L. Harley and Patricia J. Knobloch, "Government Documents Reference Aid: An Expert Systems Development Project," *Government Publications Review* 19 (Jan./Feb. 1991): 15–33.
9. According to the sixth edition of the *Directory of Government Documents Collections and Librarians* (Bethesda, Md.: CIS, 1991), Stanford holds an extensive collection of federal, international, and foreign documents and a moderate collection of state and local materials.

10. John Richardson, *Knowledge-based Systems for General Reference Work* (San Diego: Academic, 1995).
11. Ibid.
12. The first author consulted with Peter Hernon who has worked extensively in this area as well. He agreed that "there has never been a reported study done on the types of questions asked. There have been studies—with no reliability and validity indicators—of the questions asked at a general reference desk," correspondence dated Apr. 30, 1993.
13. Kenny Crews, "The Accuracy of Reference Services: Variables for Research and Implementation," *Library and Information Science Research* 10 (July/Sept. 1988): 331–56.
14. Charles A. Bunge, *Professional Education and Reference Efficiency*, Research Series No. 11 (Springfield, Ill.: Illinois State Library, 1968; abridged version of "Professional Education and Reference Efficiency" (Ph.D. diss., University of Illinois, 1967); Terence Crowley and Thomas Childers, *Information Service in Public Libraries: Two Studies* (Metuchen, N.J.: Scarecrow, 1971); Jassim M. Jirjees, "Telephone Reference/Information Services in Selected Northeastern College Libraries," in *The Accuracy of Telephone Reference/Information Services in Academic Libraries: Two Studies* (Metuchen, N.J.: Scarecrow, 1983); Peter Hernon and Charles R. McClure, *Improving the Quality of Reference Service for Government Publications*, ALA Studies in Librarianship, No. 10 (Chicago: ALA, 1983) and *Unobtrusive Testing and Library Reference Services* (Norwood, N.J.: Ablex, 1987); Marcia J. Myers, "Telephone Reference/Information Services in Selected Northeastern College Libraries," in *The Accuracy of Telephone Reference/Information Services in Academic Libraries: Two Studies* (Metuchen, N.J.: Scarecrow, 1983); and Kathy A. Way, "Quality Reference Service in Law School Depository Libraries: A Cause for Action," *Government Publications Review* 14 (1987): 207–19.
15. In the history of reference quality, most studies have asked as few as ten questions while only a few have asked as many as twenty. Future work should consider the implication of small *Ns*; generally as *N* increases, so does sensitivity.
16. Because almost any question could be answered using a government publication, we tried to select only those obviously requiring such a source: i.e., questions requiring an official version, an authoritative source, or reliable statistical information. From Bunge's Appendix C, which lists eight government documents questions (i.e., 1, 2, 4, 7, 18, 20, 23, and 28), we randomly selected numbers 8 and 9; Childers had eight (i.e., 2, 5, 9, 11, 18, 22, 25, and 26), we selected number 4 and 5; Crowley had four (i.e., 1, 2–4, 7, and 8), we selected number 2; Jirjees' nine (i.e., 1, 6, 7, 12, 16, 17, 24, 28, and 34), we selected number 8 and 9; Way's twelve (i.e., 1, 3, 4, 5, 6, 8, 11, 12, 17, 18, 19, and 20), we took number 3 and 6; Myers' four (i.e., 1, 5, 7, and 13), number 7; McClure and Hernon's (1983, appendix A) listed twenty (i.e., all of them), we selected number 3, 9 and 17; and from McClure and Hernon (1987, appendix B), fifteen (i.e., all of them), we selected number 1 and 12.
17. We need more studies on the degree of difficulty issue. In 1967 Bunge asked 47 librarians, of whom 37 responded, to rate questions as "easier, average, or harder" than normal (see *Professional Education*, appendix B).
18. Crews, "The Accuracy of Reference Services," 331–56.
19. The issue of multiple sources is vexing. A user validation of the response scheme is highly desirable. For instance, we need to know the answer to the following questions: (1) Is the user more confident when he has more sources in hand, or (2) Is the user more satisfied when he has more sources in hand?
20. Cheryl Elzy, Alan Nourie, F. W. Lancaster, and Kurt M. Joseph, "Evaluating Reference Service in a Large Academic Library," *College and Research Libraries* 52 (Sept. 1991): 454–65.
21. Ibid., p. 458. One consequence of negative values is that if the data screening reveals that the distribution of this variable is not normal, then a constant may be added before undertaking logarithmic transformations. A. A. Afifi and Virginia Clark provide a clear discussion of this point as well as the "effect on the statistical properties of the transformed variable" in their *Computer-Aided Multivariate Analysis*, 2d ed. (New York: Van Nostrand Reinhold, 1990), 53.
22. The standard discussion of such statistical matters is covered in Vic Barnett and Toby Lewis, *Outliers in Statistical Data*, 2d ed. (New York: Wiley, 1984) or R. D. Cook, "Influential Observations in Linear Regression," *Journal of the American Statistical Association* 74 (1979): 169–74.

23. When the dependent variable is dichotomous (i.e., high accuracy versus low accuracy), a logistic regression is appropriate; see David W. Hosmer Jr. and Stanley Lemeshow, *Applied Logistic Regression* (New York: John Wiley and Sons, Inc., 1989). For our analysis, SCR=MACHINE QN where AVGSCR=actual score/potential score and if AVGSCR .5 then SCR =1 and ELSE = 0. Machine is dummy coded 1 for POINTER and 0 for GDRA.
24. John Richardson Jr., "Paradigmatic Shifts in the Teaching of Government Publications, 1895–1985," *Journal of Education for Library and Information Science* 26 (Spring 1986): 249–66; reprint ed., *Encyclopedia of Library and Information Science*, vol. 44: 242–58.
25. Cherie B. Weil, "Classification and Automatic Retrieval of Biographical Reference Books" (master's thesis, University of Chicago, 1967; idem, "Automatic Retrieval of Biographical Reference Books," *Journal of Library Automation* 1 (Dec. 1968): 239–49. In fact, Weil said that she could not answer three of the four questions because she had exhausted her resources; her knowledge-based system found answers to those three questions in the same sources to which Weil had access.

APPENDIX A
Set of Three Training Questions

1. For a term paper in history, I am studying the Army's use of camels in the nineteenth century. It is my understanding that there is a government document, from the 1850s, on the topic. Please help me find it. (Hernon and McClure, 1983, #1)
2. I would like to know the name of a general who was forced to retire from the Army after twice publicly criticizing President Carter's military policies. I think the incident took place sometime around the middle of 1977. (Jirjees, #3)
3. When was George Washington given the title of General of the Armies of the United States? (Myers, #2)

APPENDIX B
Fifteen Test Questions

1. I would like the names and office addresses of the senators and representatives representing me in the federal legislature. I live in the downtown area of this city. (Bunge, #1)
   POINTER: Y, N, N, 3, 3 = *Government Manual, Official Congressional Directory,* FED, *Congressional Staff Directory,* and Government Documents Catalog.
   GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index.*

*2. How much more or less expensive is it for an average family to live in Chicago than it is in Atlanta? (Bunge, #18)
   POINTER: Y, N, N, 3, 2, N, 2 = *American Statistics Index* and *Statistical Abstract.*
   GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index.*

*3. Where is the nearest commercial airport to Rio Grande, Ohio? (Childers, #11)
   POINTER: Y, N, N, 4 = Maps (referral to same institution, but different department) plus *Using Government Publications* and *Monthly Catalog* and Government Documents Catalog.
   GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index.*

4. What is the salary of the President of the United States? (Childers, # 22; assumptions: federal law)
   POINTER: Y, N, N, 3, 5 = *United States Code.*
   GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index.*

5. What is the name of the secretary of commerce? (Crowley, #2-4)
   POINTER: Y, N, N, 3, 3 = *United States Government Manual.*
   GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index.*

6. I need to know the percentage of persons below the poverty line in Colorado for the year 1975. (Jirjees, # 28)
   POINTER: Y, N, N, 3, 2, Y = 1980 Census.
   GDRA: INTRO, F2, F2, STATS, US STATS, F2 = *American Statistics Index.*

*7. In 1977 the U.S. Commission on Civil Rights released a report called *Window Dressing on the Set*. It's about the treatment of women and minorities on TV. Has the commission published any study to update that report since then? (Jirjees, #34; assumptions: subject approach; report, when do you stop—after checking every year since 1977)
POINTER: Y, N, N, 3, 14, 7, Y, and 8 = *Monthly Catalog* and Cumulative Index 1981–85, 1976–1980.
GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index*.

*8. I understand that the Caffeine Study Review Panel submitted its final report to the Food and Drug Administration on May 15, 1981. The report contains information pertinent to the FDA's review of the safety of added caffeine. I would like to know if the final report is available. (McClure and Hernon, 1983, # 3; the authors deleted the remainder of this question.)
POINTER: Y, N, N, 3, 14, 7 = *Monthly Catalog* and *Cumulative Index*.
GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index*.

9. In February 1978, there was an FTC (Federal Trade Commission) staff report on television advertising to children, by Ellis M. Ratner and others. It recommended the elimination of "harms arising out of television advertising to children." Is it still in print? What is the cost? (McClure and Hernon, 1983, #9; assumptions: current date)
POINTER: Y, N, N, 3, 14, 8 = Government Documents Catalog; *Publication Reference File*.
GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index*.

10. Where can I get a detailed breakdown of the distribution of federal funds for research and development by agency? (McClure and Hernon, 1983, #17)
POINTER: Y, N, N, 3, 4 = *Catalog of Federal Domestic Assistance*.
GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index*.

11. What is the zip code for Behrend College in Erie, Pennsylvania? (Myers, #7; assumptions: inquirer does not want address and the college is not a government organization)
POINTER: Y, N, N, 3, 14, 8 = Government Documents Catalog; Publications Reference File; Cumulative Index; *Monthly Catalog*.
GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index*.

12. For a term paper in history, I am studying the laws on the imprisonment of free black seamen in the South prior to the Civil War. It is my understanding that the government published a report on the topic in the 1840s. (McClure and Hernon, 1987, #1; assumptions: laws = Congress)
POINTER: Y, N, N, 3, 14, 1 = *CIS Index* or *CIS US Serial Set Index*.
GDRA: INTRO, F2, F2, INFO, US, F2, F, CONG = *CIS Index*.

13. In 1980 a public law was enacted that it provided universities and small business with the right to obtain patents for inventions which their faculties and staff created with the use of Federal funds. Please help me locate a copy of the law. (McClure and Hernon, 1987, #12)
POINTER: Y, N, N, 3, 5 = *U.S. Code* and other titles.
GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index*.

14. How many years must a U.S. magistrate have been a member of the bar prior to appointment? (Way, #4; assumptions: federal law)
POINTER: Y, N, N, 3, 5 = United States Code or 6 = *Code of Federal Regulations*.
GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index*.

15. Did former President Ford appoint Barbara Walters and Katherine Hepburn to the National Commission on the Observance of International Women's Year? (Way, #17; assumptions: done by Executive Order)
POINTER: Y, N, N, 3, 8 = *Weekly Compilation of Presidential Documents or Public Papers of the Presidents*.
GDRA: INTRO, F2, F2, INFO, US, F2, T = *Monthly Catalog* or *CIS Index*.

Note: For POINTER, N = No, Y = Yes, numbers are responses required at menu options. For GDRA, INTRO = Introduction, F2 Continue, Info = Information, US = United States, STAT = Statistics, and T = True.

* Indicates initial disagreement in interpreting appropriate response to system's question. Consensus, as reported in appendix, was achieved after discussion.