

ARDA Challenge Workshop 2004

An Investigation of Evaluation Metrics for Analytic Question Answering

Emile L. Morse, PhD and Jean Scholtz, PhD, PIs
Paul Kantor Ph.D. and Diane Kelly Ph.D. Consultants
With Ying Sun, Rutgers University Graduate student

EXECUTIVE SUMMARY	3
1 INTRODUCTION	4
2 DESCRIPTION OF THE EXPERIMENTAL METHOD	6
2.1 EXPERIMENTAL SET-UP	6
2.1.1 <i>Location</i>	6
2.1.2 <i>Systems</i>	6
2.1.3 <i>Hardware</i>	6
2.1.4 <i>Infrastructure</i>	6
2.1.5 <i>Observers</i>	7
2.1.5 <i>Analysts</i>	8
2.1.6 <i>Scenarios</i>	8
2.1.7 <i>Data Corpus</i>	9
2.2 SCHEDULE, INSTRUMENTS AND ACTIVITIES	10
2.2.1 <i>Introductory Plenary Session</i>	10
2.2.2 <i>Experimental Design</i>	10
2.2.3 <i>Experimental Blocks and Sessions</i>	11
2.3 SYSTEM LOGS AND GLASS BOX INFRASTRUCTURE	16
2.4 END-OF-WORKSHOP ACTIVITIES	17
2.4.1 <i>Scenario Difficulty Assessment</i>	17
2.4.2 <i>System Discussions</i>	17
2.4.3 <i>Focus Group on Systems</i>	17
2.4.4 <i>Discussion of the Entire Workshop</i>	17
3 SOURCES OF ERROR IN THE ANALYSIS REPORTED HERE	18
3.1 SAMPLING ERRORS.....	18
3.2 STATISTICAL ERROR.....	18
3.3 SYSTEMATIC ERROR.....	19
4 PRINCIPAL FINDINGS	20
4.1 CROSS EVALUATION	20
4.1.1 <i>Factor Analysis</i>	21
4.1.2 <i>The System Effect</i>	21
4.1.3 <i>Individual Aspects of the Product</i>	24
4.1.4 <i>Cross-evaluation Validation</i>	25
4.2 POST-SESSION AND POST-SYSTEM QUESTIONNAIRES	26
4.3 SMIFRO CONSOLE AND STATUS QUESTIONNAIRES.....	31
4.4 SYSTEM LOGS.....	32
4.5 GLASS BOX DATA.....	34
4.5.1 <i>Allocation of time</i>	34
4.5.2 <i>Activity trails</i>	35
4.6 QUERY TRAILS	36
4.7 NASA TLX.....	37
4.8 SCENARIO ASSESSMENT.....	37
4.8.1 <i>Post-Scenario Questionnaires</i>	37
4.8.2 <i>Scenario Difficulty Assessment</i>	38
5 SUMMARY	40
5.1 KEY FINDINGS	41
5.2 LESSONS LEARNED.....	41
5.3 AVAILABLE RESOURCES.....	42

Executive Summary

Evaluation of interactive systems is a complex task because it concerns two entities: the system and the individual and the interaction of the two. While technical evaluations give information on system performance (time, precision, and recall), we still need to determine the utility of the system – its value to the user. There are two aspects of interaction to use for evaluation: the analysts' process and the analysts' products. The objective measures of efficiency and effectiveness are measures of the process – the more efficient the analyst is in finding information with the system, the more time she will have for analysis. The more effective the analyst is also saves iterations of queries and hence also time. The more subjective rating of the analysts' reports will assess if better quality information is being located and going into the final product.

This workshop addressed the challenge of determining metrics that could distinguish between end to end question answering systems and to determine how these metrics could be implemented. The ultimate goal is to identify an evaluation scheme for the AQUAINT end to end systems.

The workshop team decided that to run an empirical study using a number of different question answering systems and a baseline system for comparison. Analysts would use all the systems on real-world analytic tasks. The Glass Box infrastructure from the Novel Intelligence from Massive Data (NIMD) program along with system logs would be used to collect analysts' interactions with the software. In addition, analyst would be observed and would be given various questionnaires. The first need was to identify possible metrics and data collections to support those metrics. This was necessary so that the systems would know what to log and that the appropriate questionnaires could be developed. This does not rule out the possibility that more metrics might be discovered during the workshop, but ensures that a number of potential metrics are examined.

The approach of this investigation was based on using hypotheses about question-answering systems to drive the development of appropriate methods and metrics for measuring system effectiveness. Fifteen hypotheses were developed, of which 13 were operationalized using a variety of methods. Cross-evaluation was used as the primary product method. System logs and Glass Box data were the core logging methods to collect process data. In addition, post-scenario, post-session, post-system and cognitive workload questionnaires, interviews, focus groups, and other user-centered methods were applied.

The study employed eight analysts, eight scenarios in the chembio WMD domain, and four system – 3 QA systems and a Google baseline system. Each analyst used each system to analyze two scenarios and was tasked to write a pseudo-report containing enough structure and content to allow it to be judged by his peers during cross-evaluation.

The results show that of the 14 hypotheses that were tested, 13 of them could be addressed by at least one of the methods. That is, a difference across systems could be demonstrated with respect to 13 of 14 hypotheses. Overall, questionnaires gave the greatest coverage – 10 hypotheses. Cross-evaluation, a product quality method, was sensitive to the different systems. The NASA TLX cognitive workload method succeeded in distinguishing systems that had a high and low workload; the mental and temporal components were significantly affected by the system be used. Most of the methods gave evidence that could distinguish systems from one another.

It was not the charter of the Workshop to identify a 'best' QA system from the participating systems. We hoped that by including systems that were sufficiently diverse that we had our best chance of identifying metrics that could be applied across the range of QA systems that exist. We believe that we succeeded in this goal. Although one QA system was deemed superior across many of the measured aspects, it is worth noting that every system was 'best' in at least one aspect. Even the Google system was best in one area – ease of learning. This leads us to believe that we have identified a strategy for testing systems that is likely to be sensitive to differences between systems.

The real message of this study goes beyond QA systems. We believe that measurement in and of itself is a worthy effort. Only when systems prepare for and undergo rigorous testing will they know if they have succeeded in their original design goals.

1 Introduction

This workshop addressed the challenge of determining metrics that could distinguish between end to end question answering systems and to determine how these metrics could be implemented. The ultimate goal is to identify an evaluation scheme for the AQUAINT end to end systems.

The workshop team decided that to run an empirical study using a number of different question answering systems and a baseline system for comparison. Analysts would use all the systems on real-world analytic tasks. The Glass Box infrastructure from the Novel Intelligence from Massive Data (NIMD) program along with system logs would be used to collect analysts' interactions with the software. In addition, analyst would be observed and would be given various questionnaires. The first need was to identify possible metrics and data collections to support those metrics. This was necessary so that the systems would know what to log and that the appropriate questionnaires could be developed. This does not rule out the possibility that more metrics might be discovered during the workshop, but ensures that a number of potential metrics are examined.

The workshop team started by listing a number of hypotheses about the benefits of question answering systems. For each hypothesis, the team identified potential metrics and ways to operationalize these metrics. These hypotheses are shown in table 1.

Table 1: Initial hypotheses for benefits of QA systems

	Hypothesis	Metrics	Measurement vehicles
H1	Support information gathering with lower cognitive workload	Workload [Cognitive load, number of questions asked, percent of interactions where system takes the initiative, percent of non-content input (clarifications), analysts' ratings]	NASA TLX System logs Questionnaires
H2	Assist in exploring more paths/hypotheses	Number of paths/hypotheses explored Aspects covered in report Confidence in report	System log Product evaluation Questionnaires
H3	Enable production of higher quality reports	Quality of report [completeness, clarity, coverage]	Product evaluation Questionnaires
H4	Provide useful suggestions to the analyst	Percentage of suggestions useful to analyst [Questions input by analyst Question suggested by system that analysts use]	System log Questionnaire
H5	Provide more good surprises than bad	Percentage of surprises rated good [# good things, # bad things]	Real-time input from analyst
H6	Enable more focus on analysis than data collection	Time spent in different stages of analysis [foraging, synthesizing]	Glass box records Questionnaire
H7	Enable analysts to collect more data in less time	Growth of shoebox	Glass box records Questionnaire
H8	Reduce the time spent reading	Time spent reading documents	Glass box records
H9	Identify gaps in the knowledge base	Number of times no response was given to a query or question	Glass box records System logs
H10	Help the analysts recognize	Agreed not to measure	

	Hypothesis	Metrics	Measurement vehicles
	gaps in their thinking		
H11	Provide relevant context for information	Quality of justifications, number of times analyst asks for justification	System logs Questionnaires
H12	Provide context, continuity, and coherence of dialogue	Shifts in dialogue focus, number of content dialogues vs number of clarification dialogues Systems' redundancy detection of documents	System logs Questionnaires
H13	Let analysts relocated previously seen materials	Agreed to postpone for longer term analysis sessions	
H14	Be easy to use	Usability metrics: efficiency, effectiveness, user satisfaction	Proficiency in the usability test (ease of learning), NASA TLX (performance, frustration) compared between task 1 and task 2
H15	Increase an analyst's confidence in exploration and quality of report	Confidence ratings	Questionnaire

The next step was to operationalize the metrics in column three. Multiple implementation schemes for each metric were necessary to determine if the hypothesis was supported. The next step was to add the fourth column to table 1 identifying possible ways to obtain the necessary measures for each metric. For many of the metrics, it was suggested to use questionnaire or interview data as well as quantitative data collected from system logs or from the glass box infrastructure.

We identified a number of instances where we wanted to collect data from the users. In column 4 of table 1 we have simply labeled these as Questionnaires. However, there are a number of techniques such as mini-focus groups, full (all participants) focus groups, observations of users, interviews with users. For the purposes of this experiment, we did not necessarily pay attention to the cost of obtaining the data for the metrics, although this will be important once program wide evaluations are created. Table 2 shows a comparison of methods and the associated costs of data collection and analysis in addition to the validity and discriminatory power of the implementation methods.

Table 2: Cost of various methods

Method	Collection cost	Analysis Cost
System logs (SL)	.	\$\$
Glass Box (GB)	.	\$\$\$
Questionnaire	\$	\$
NASA TLX	.	\$
Cross-Evaluation	\$\$	\$
Mini-focus	\$\$	
Full Focus	\$\$	
Observation	\$\$	\$\$
Interview	\$	\$\$\$
Email Follow-up	.	,

Cost is regarded as very low (.), low (\$) moderate (\$\$) or high (\$\$\$). This cost represents primarily human effort required to repeat the analysis after it has been done once.

For the purposes of this workshop, we used as many multiple methods as possible. Our analysis will look at each hypothesis and determine which were supported using the various implementations. Recommendations for future evaluations will consider the costs associated with the various implementations.

It is important to note that the goal of this workshop was to develop metrics that could discriminate among various question answering systems and was not intended to evaluate the systems participating. As a benefit of participating in the workshop, systems were able to get feedback that could be incorporated. Participating systems were in various stages of development and the researchers had varying amounts of time to develop the systems specifically for this workshop.

2 Description of the Experimental Method

2.1 Experimental Set-up

2.1.1 Location

The experiment was conducted at the PNNL in Richland, Washington. The space used was one room with support servers, four rooms in which systems were installed, and a conference room seating 20, which was used for general meetings, focus group discussions, meetings among observers, meetings among developers, etc. For part of the workshop, this room also housed two laptop computers that were plugged into the PNNL LAN, one of which was networked to two local printers. The conference room was pleasant, with excellent light and air, and nearby snacks and facilities. The space also served, with some awkwardness, as a "break room" for analysts who completed specific tasks in less than the allotted time.

2.1.2 Systems

The systems were FERRET (developed by LCC); GINKO (Developed by Cycorp); Google (operated by NIST); HITIQA (developed by SUNY-Albany/Rutgers). Two of the experimental system developers, along with the Google baseline team, began work on Wednesday. System developers determined which system rooms they would occupy by blind selection. The fourth developer, who arrived on Thursday, was given the final system room that had not been blindly selected. Each developer completed different activities to set up their systems. Two developers ran their systems from remote locations, while two developers, including the Google developer, did local installs. One developer, who did a local install, installed the system on a PNNL server, at his own site, and estimates that one week is required for burn-in, and ensuring there are no hardware, third-party application or OS incompatibilities.

2.1.3 Hardware

The workstations used during the experiment were set up at the start of the first week. Each of the Dell workstations were configured with Windows XP Professional with updated OS, Intel Pentium IV processor 3.40 Ghz 512 K/800 Mhz, 2 GB DDR 400 SD RAM, 120 GB SATA 7200 RPM hard drive with Data Burst Cache, video card, floppy drive, 16 DVD ROM, and 48/32/48 CDRW. Two workstations were installed in each system room.

2.1.4 Infrastructure

The HITIQA and Google servers were housed on different machines at PNNL, while LCC and Cycorp connected to computers at their home office to effect server support.

Glass Box v2.3 software (PNNL) was set up on each of the 8 computers used in this test. Internet Explorer was the only browser installed. Microsoft Word, Excel and PowerPoint were also

available on each computer. Each computer was attached to the PNNL network so that they could contact the appropriate QA server. In addition, each local system was configured to access on-line questionnaires from a variety of sources. Refer to section 2.3 where what glass box captures will be described

2.1.5 Observers

There were four observers in this study. Two observers were graduate students who were awarded competitive PNNL Fellowships to participate in the project. These students were, in part, selected based on their previous experiences conducting user studies and system evaluations. Two other observers had additional roles in the project. One of these observers was a Metrics Consultant to the project, and also served as Lead Observer. This observer was responsible for, among other things, creating and administering training materials to the other three observers on the interviewing and observation methods. The final observer was a senior member of the research team and served many roles in the project, foremost among them was Co-PI.

As with the system developers, observers began work on Wednesday, before the arrival of the analysts, and had three days to prepare. During this time, observers finalized instruments, prepared experimental packets, trained on the use of the various systems, trained on the interviewing and observation methods, and set up the system rooms. Observers also did various other tasks like photocopying, and securing office supplies.

The Lead Observer conducted an observer training session on the interviewing and observation methods. The training materials used were a set of guidelines and notes about each procedure created by the Lead Observer. Training consisted of discussing these materials in an informal classroom-type setting. The Lead Observer had planned for some mock practice sessions of each technique, but time constraints prevented these from occurring. The Lead Observer was also responsible for preparing the Observer Protocol (to be discussed in more detail in Section 2.2.3.1), and introducing this instrument to other observers.

During the three-day preparation period, observers visited each system room to insure that the desktops of each system were uniform in appearance and to pilot the various applications that would be used during the experiment. Appropriate desktop shortcuts were created for all relevant applications. A desktop shortcut to a web page that contained links to all of the various online instruments was also created at this time. During these visits, observers also organized the physical space to allow for an optimal environment for conducting observations of two analysts simultaneously, and for having two analysts work in the same room. Efforts were made to insure that the physical setup of all rooms was uniform, but some small variations still existed.

The amount of preparation time that observers had before the arrival of the analysts was reported as barely adequate to complete all of the necessary activities. The observers estimated that five full days would have been adequate. Observers recommend that more time be dedicated to their own training on the use of the various systems. This additional training would also provide developers with an opportunity to practice their training materials on real users. Observers also desired more time for training on the various methods of the study, and in particular, the naturalistic observation and interviewing methods. Observers suggest dedicating time to conduct "mock" system use/observation sessions, which would allow them to both practice using the systems and practice the particular methods of the study. This longer time period would also allow for a more thorough development and piloting of the Observer Protocol, which instructed each observer on how to conduct the experimental sessions.

After the analysts' arrival, observers' duties primarily consisted of observing and administering the experimental sessions. At the end of each day of the study, observers met in the conference room to discuss their experiences observing and administering the experiment. Observers also presented problems they encountered during the experimental sessions, and suggested specific

solutions to apply to these problems when, and if, they occurred again. More details of observers' experimental duties are presented in Section 2.2.

2.1.5 Analysts

Analysts were recruited for the study through NIST's usual contact in the Navy Reserves. A short description of the study was posted by the contact to an online list of upcoming opportunities. Responses to the ad were vetted by the POC and a list of names was returned to NIST. Participation in this study fulfilled the reservists' yearly two-week service requirement. Eight analysts were recruited for the study, but only seven analysts were present for the entire experiment. The eighth analyst arrived during the second week of the study, and completed only one of four experimental blocks. This analyst's data is excluded from the analysis presented in this report unless otherwise noted; his ID in tables is ch6. The other seven analysts and one consultant arrived on Monday following, and were badged within 2.5 hours. This Monday was the first day of the experiment, when analysts would begin work as experimental subjects.

All analysts, with the exception of the late arrival, were naval reservists. Analysts had a mean age of 40.8 years. The education levels of the analysts ranged from high school diploma to Ph.D.; over half of the analysts had Master's degrees. Analysts served in a range of civilian jobs including hospital administrator, portfolio manager, and homemaker. Several analysts worked for the military full-time, including one intelligence officer and one intelligence analyst. Analysts had a mean of 18.3 years of military service, and a mean of 10.4 years doing analysis work. With the exception of one person, all had some experience doing analysis work.

All analysts had a computer either at home or at work; most had computers at both locations. Five of the eight analysts used a computer to conduct analysis tasks, and all analysts but one characterized their level of computer expertise as "medium." The remaining analyst characterized his level of computer expertise as "novice." All analysts, with the exception of the novice user, indicated that they had experience querying systems. The self-assessed level of expertise for querying ranged from "none" (1 analyst), to "novice" (3 analysts), to "medium" (4 analysts). A full description of the analysts can be viewed in Appendix A.

2.1.6 Scenarios

The scenarios for this study were developed by the AFRL team and were reviewed by a panel of experts at the Kick-Off meeting. The scenarios contained a short title and detailed description of the task. Fourteen tasks were developed by AFRL but only eight were used during the study. The full text of all 14 scenarios is included in Appendix B. It is important to note that the tasks were developed while the document collection was still being finalized. Once the full set of tasks was delivered, NIST characterized each of the scenarios with respect to how well the final, indexed collection could support it. As shown in Figure 1, several tasks had less than 100 documents in the CNS collection to support them and the next section describes how we dealt with this issue by manipulating the document corpus.

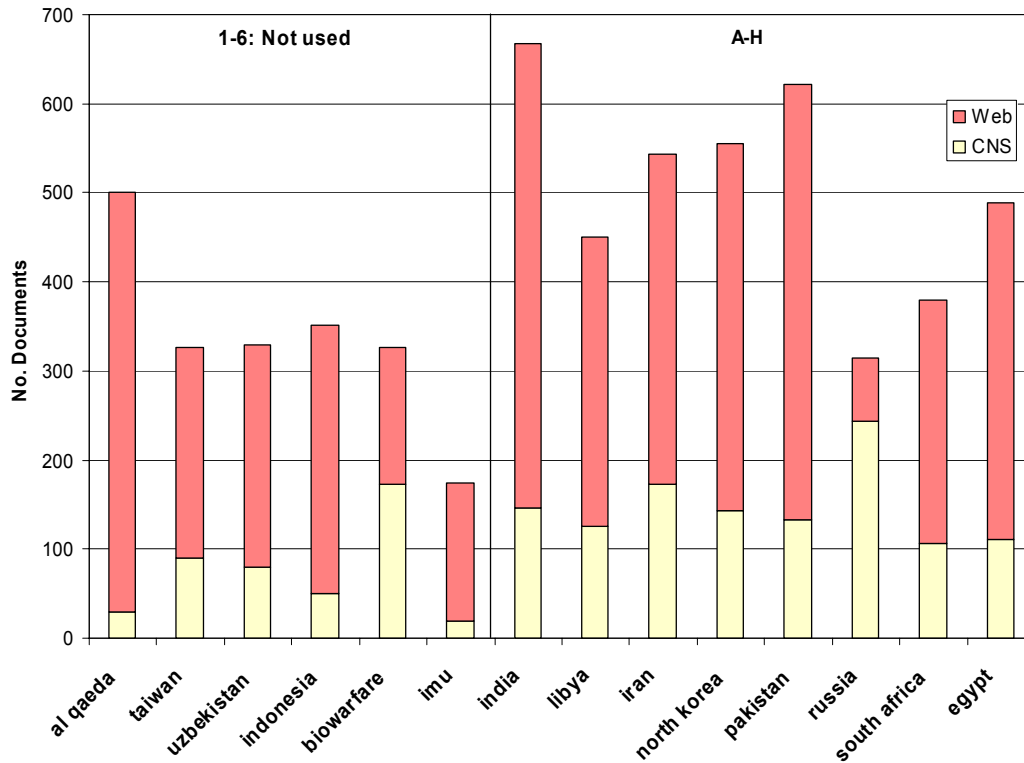


Figure 1: Coverage of scenarios within document collection

At the end of the first block of the experiment, it was apparent that even though the scenario descriptions were sufficient in describing the content of the report, important information regarding context of the description and the format of the report, such as customer and length, were omitted from the descriptions. This omission left room for ambiguity during report creation, and caused some uncertainty for the analysts with how to proceed with the task. Thus, thereafter, at the beginning of each block, analysts met as a group in the conference room to discuss the additional specifications for each scenario. In addition to this information, the project director and one analyst worked together to design a template for the report (after hours), which established a uniform report structure, and specified formatting guidelines such as headings and length. The added contextual details for each scenario also appear in Appendix B, along with the document describing the report structure Appendix C.

2.1.7 Data Corpus

We did not want to use the live world wide web, as this would make it impossible to replicate the experiment. Hence we developed a large collection of documents using a repository from the Center for Non-Proliferation Studies (CNS) and additional documents mined from the web. The document collection consists of a special distribution of "disaggregated" data from CNS and supplemental documents automatically collected from the Internet. The CNS data consists of the January 2004 distribution of the Eye on Proliferation CD, which has been "disaggregated" by CNS into about 40,000 documents. The other documents were collected by semi-automated querying of Google and retrieval of the documents listed in the results. A few unusually large and useless items, like CD images and word lists, were deleted, but the supplemental collection is still more diffuse than the native CNS documents.

The approximate counts of different kinds of files, determined by the extensions, are summarized in the table below.

Table 3: Size characteristics of Document Corpus

Source	All Files	Documents	Images
CNS	40192	39932	945
Other	261590	48035	188729

2.2 Schedule, Instruments and Activities

I would actually rearrange this as Instruments – describe everything – not JUST the user stuff – but also the NASA TLX, the system logs, the smiley-frowny. And then show the schedule

2.2.1 Introductory Plenary Session

After the badging, the director of the project gave an opening plenary in the conference room. In this plenary, the workshop goals and processes were explained to the analysts. This was a talk given from notes (see Appendix D). During this session, analysts completed a consent form. It should be emphasized to those not familiar with conducting studies with human subjects that it is necessary to obtain approval from the Institutional Review Board (IRB) for research with human subject before the start of the study. Persons interested in using the methodology described in this report should make inquiries about the IRB at their host institution as soon as possible, as this process can take several months. The IRB review packet for this experiment appears in Appendix E, along with the Consent Form.

During the plenary session, analysts were provided with a schedule that presented an overview of the study procedures. This schedule is displayed in Appendix F. Analysts were also introduced to many of the instruments and activities of the study. In one case, a handout about the instrument was circulated to aid in its explanation. However, for most instruments, such as the questionnaires and interviews, only brief overviews were provided during this opening plenary; instead, observers were charged with providing more specific instructions to analysts during the first experimental block when the various instruments were encountered. The Cross-Evaluation interface was not introduced during this opening plenary, but rather it was introduced directly before it was used for the first time. This was done to avoid biasing analysts' behavior during report creation.

2.2.2 Experimental Design

The workshop was comprised of four, two-day blocks. In each block, two analysts were assigned to each system, and a single observer was assigned to the pair of analysts. After each block, analysts and observers rotated to different system rooms, so that analysts were paired together only once and observers observed different analysts during each block. The goal in using designed experiments is to ignore the second order interactions, so that estimates of the main effects can be obtained from a much smaller set of observations than is required for a full factorial design. For instance, one might imagine potential interaction effects of system and scenario (some systems are better for certain scenarios); system and analysts (some analysts might adapt more quickly to a system); and analyst and scenario (some analysts may be more expert for certain scenarios). To minimize these potential interactions, we used a modified Greco-Latin 4x4 design. This design is shown in Table 4.

Table 4: Experimental Design (O=observer; A=analyst)

	Dates	June 14/15	June 16/17	June 18/21	June 22/23
	Scenarios	A, B	C, D	E, F	G, H
System	System 1	O1	O2	O3	O4
		A1	A2	A3	A4
		A5	A6	A7	A8
	System 2	O2	O1	O4	O3
		A4	A3	A2	A1
		A7	A8	A5	A6
	System 3	O3	O4	O1	O2
		A2	A1	A4	A3
		A8	A7	A6	A5
	System 4	O4	O3	O2	O1
		A3	A4	A1	A2
		A6	A5	A8	A7

This design ensured that each analyst was observed by each of the four observers, and used each of the four systems. This design also ensured that each system was, for some analyst, the first, second, third or last to be encountered, and that no analyst did the same pair of scenarios twice. Since we anticipated that analysts would interact with one another over the course of the workshop, we paired each of the analysts so that they were only paired together once. It should be noted that both analysts and scenarios were randomly assigned codenames (e.g. A1, and Scenario A), and that the systems were randomly assigned to the rows of Table 4. It is also worth noting that because of this random assignment, the analyst who only participated in the final block was A6, not A8. Although observers were simply rotated across the system rows, the assignment of human individuals to code number was random.

At the start of each block, observers met with the analysts that they would be observing, in the conference room and escorted them to the appropriate system room. Observers were trained to situate analysts with code numbers 1-4 at the workstation on the left-hand side of the room and analysts with code numbers 5-8 at the workstation on the right-hand side of the room. This decision was related to the operation of the Glass Box software. Once analysts were situated, the observer left the room while the training session occurred.

At the start of each block, there was a training session conducted by the system developer, followed by a skills check test, and free experimentation. The methods of training varied from a tight script during which trainees carried out steps on their own workstation, to a slide presentation with scripted trainee activities, to a presentation from a printed manual, to presentation, orally and with participation, guided by a checklist.

2.2.3 Experimental Blocks and Sessions

Each experimental block was composed of two experimental sessions, which corresponded to the two unique scenarios that were conducted during the block. The instruments described below were either administered at the end of the experimental session, in which case the analyst would complete two of these instruments during the block (i.e. one for each scenario), or once, at the end of the experimental block.

2.2.3.1 Observer Protocol

After the training sessions, observers re-entered the system rooms with the analysts. At the start of the first block, observers carefully introduced (and reintroduced in some cases), the various instruments that would be used to collect data during the experimental session such as the Glass Box logging tool, the SmiFro Console, the Status Questionnaires, and various other questionnaires. These instruments are described more fully in Section 2.2.3. As each instrument

was encountered *during* the block, observers explained its purpose and verbally described its use. Although most analysts quickly learned about all of the instruments during the first block, observers continued to briefly re-introduce each instrument at the start of each session as specified in the Observer Protocol.

The Observer Protocol provides a detailed list of instructions to the observer about how to conduct the experiment. The Observer Protocol is a set of step-by-step instructions about how to administer the experimental session. The goal of this Protocol is to insure uniformity in how each experimental session was conducted, to remind the observers of all of the necessary steps to take during the experimental session and the particular order in which these steps should be taken, and to provide potential solutions to problems that might arise. This Protocol was developed during the three-day preparation period and refined over the first few days of the study. Observers began the session by reading a set of instructions to the analysts, which described the various activities and parameters of the session. The Observer Protocol can be viewed in Appendix H.

2.2.3.2 Scenario Distribution & Instructions

In the first block of the study, the observers distributed a scenario description to each analyst after reciting the introductory set of instructions referred to the previous section. In all other blocks, scenarios were distributed to the analysts in the conference room during a 15-minute Scenario Briefing. As described in Section 2.1.6, the short briefing was added to the schedule to promote a common context for the analysts. The analysts worked to build a “draft report” on that topic using Microsoft Word. Each scenario description contained a unique and random identification number that analyst used to name their file and title their report. Analysts were given 2.5 hours to complete this task.

2.2.3.3 Observation Worksheet

During the experimental session, observers used an Observation Worksheet to record activities and behaviors that were believed to be indicative of analysts’ level of comfort, and feelings of satisfaction or dissatisfaction. Observers further recorded information about analysts’ apparent patterns of activities. Observers were trained on how to use the Worksheet before the experiment. The Worksheet clearly enforced a separation between the behavior observed, and the observers’ possible inferences or questions about the behavior. Observers further used the Worksheet to identify behaviors about which to follow-up during subsequent interviews about the session. The Observation Worksheet primarily produced qualitative data; this data is not included in this report. The Observation Worksheet can be viewed in Appendix I.

2.2.3.4 Spontaneous Self-Reports

Analysts were asked to report on their experiences during the experimental session in three ways: spontaneous comments into a lapel microphone, spontaneous use of the SmiFro Console comment device, and periodic (30 min. intervals) completion of a three-item online Status Questionnaire.

The Glass Box software recorded comments that analysts made into the lapel microphone. However, during the experiment, there was *no* use of the lapel microphone for making spontaneous comments.

The SmiFro Console provided analysts with a persistent tool for commenting on their experiences using the system. This tool could be used at any moment during the interaction. The SmiFro Console was rendered in a small display window, and analysts were asked to leave this window open on their desktops at all times. The SmiFro Console displayed both a smiley and a frowny face, either of which analysts could select using radio buttons. This Console also displayed a text box, which analysts could use to elaborate on their smiley or frowny, or write additional

comments. These two items (the faces and the text box) could be submitted independently of one another or in conjunction with one another. The SmiFro Console can be viewed in Appendix J.

The SmiFro Console also contained links to the Status Questionnaires. The Status Questionnaires were designed to solicit analysts' opinions and feedback about the progress of their work during the session. Each questionnaire contained the same three questions, which were worded differently to reflect different moments in time. There were four total Status Questionnaires, each corresponding to a 30-minute interval during the session: 30, 60, 90, 120 minutes. Observers tracked the session time and instructed analysts on when to complete the Status Questionnaires. The Status Questionnaires can be viewed in Appendix K.

2.2.3.5 NASA TLX Questionnaire

After the session, analysts first completed the NASA TLX. The NASA TLX is a standard instrument used in aviation research to assess pilot workload and was used in this study to assess analysts' subjective cognitive workload while completing each scenario. There are six factors in the standard NASA TLX:

- Mental demand: whether this searching task affects a user's attention, brain, and focus.
- Physical demand: whether this searching task affects a user's health, makes a user tired, etc.
- Temporal demand: whether this searching task takes a lot of time that a user can't afford.
- Performance: whether this searching task is heavy or light in terms of workload.
- Frustration: whether this searching task makes a user unhappy or frustrated.
- Effort: whether a user has spent a lot of effort for this searching task.

Each factor must be rated across the same scale but the choice of the scale range is arbitrary. We chose a 7-point scale for each of the factors. The '1' position was labeled 'little' and the '7' position was labeled 'much'.

Analysts were allotted ten minutes to complete this Questionnaire. This Questionnaire can be viewed in Appendix L.

2.2.3.6 Post-Scenario Questionnaire

Following the NASA TLX, analysts completed the six-item Scenario Questionnaire. The Scenario Questionnaire was used to assess particular dimensions of the scenario, such as its realism and difficulty. Analysts were allotted ten minutes to complete this Questionnaire. This Questionnaire can be viewed in Appendix M.

2.2.3.7 Post-Session Questionnaire

After completing the Post-Scenario Questionnaire, analysts completed the fifteen-item Post-Session Questionnaire. The Post-Session Questionnaire was used to assess analysts' experiences using a particular system to complete a scenario. This Questionnaire was designed so that submitting the responses echoed the completed questionnaire to the screen. Observers examined these responses and used them to construct follow-up questions for the subsequent Post-Session Debriefing Interview, which is described in more detail below. Analysts were allotted ten minutes to complete this Questionnaire. This Questionnaire can be viewed in Appendix N.

2.2.3.8 Post-Session Debriefing Interview

After completing the Post-Session Questionnaire, the observer used a Post-Session Debriefing Interview Schedule (Appendix O) to privately interview each analyst. The Interview Schedule

contained instructions to the observer for conducting the interview, and also provided a list of seven open-ended questions. One of these questions required the observer to use their notes from the Observation Worksheet, while two of these questions required the observer to use analysts' responses to the Post-Session Questionnaire items.

Observers prepared for the interviews by examining the Observation Worksheet and the results of the Post-Session Questionnaire, and recording appropriate follow-up items on the Post-System Debriefing Interview Schedule. Observers were allowed to select any item(s) from the Observation Worksheet for which they had questions. Observers were instructed to select items from the Post-Session Questionnaire for which analysts gave extremely high or low scores. In cases where there were none (i.e. where analysts responded to most items with "3"), observers were instructed to use their judgment to select items about which to ask questions. While the observer prepared for this interview, analysts were asked to wait outside the system room either in the conference room or the lobby of the building.

Once the observer finished preparing for both interviews (approximately 5-7 minutes) one analyst was asked to come back into the system room to participate in the interview, which was recorded using the Glass Box software. Once the interview with the first analyst was finished, the observer conducted the interview with the second analyst. Each interview was designed to last for 15 minutes. In all cases, analysts with code numbers 1-4 went first during the Post-Session Debriefing Interview at the end of the first session of the block, and analysts with code numbers 5-8 went first during the interview at the end of the second session of the block. The qualitative data produced by the Post-Session Debriefing Interview is not included in this report.

2.2.3.9 NASA TLX Weighting Instrument

At the end of the two-day block, analysts completed the NASA-TLX Weighting instrument. The NASA-TLX Weighting instrument was used to elicit a ranking from analysts about the factors that were probed with the NASA-TLX instrument. There are 15 pair-wise comparisons of 6 factors and the analyst is forced to choose one in each pair as more important. By weighting the ratings from the TLX instrument, it is possible to calculate meaningful workload component values when analysts might have rated all the components as having an exactly equivalent rating. For example, if an analyst rated all 6 factors '3' on the ratings, the total TLX score will be 3 but the relative contribution of each component will add based on the number of times that a component was deemed to be the more important factor. Analysts were allotted 10 minutes to complete this instrument. This Instrument can be viewed in Appendix P.

2.2.3.10 Post-System Questionnaire

After completing the NASA-TLX Weighting instrument, analysts completed a thirty-three item Post-System Questionnaire, which was used to assess their experiences using the system that they were assigned during the block. Analysts were allotted 20 minutes to complete this Questionnaire, but in most cases, finished sooner. As with the Post-Session Questionnaire, the Post-System Questionnaire was designed so that submitting the responses echoed the completed questionnaire to the screen. Observers were then able to examine these responses and ask follow-up questions about analysts' responses during the Post-System Debriefing Interview. This Questionnaire can be viewed in Appendix Q.

2.2.3.11 Post-System Debriefing Interview

After analysts completed the Post-System Questionnaire, they were asked to wait outside the system room while the observer prepared for the interviews (approximately 5-7 minutes). Observers used a Post-System Debriefing Interview Schedule to guide the interviews (Appendix R). The Interview Schedule contained instructions to the observer for conducting the interview, as well as six open-ended questions. Observers were instructed to construct content for two of these questions from analysts' responses to the Post-System Questionnaire. This was done in a manner identical to the one observers used to construct Post-Session Debriefing Interview items

from the Post-Session Questionnaire; this method is described in the preceding section 2.2.3.8. Each interview was scheduled to last for no more than 15 minutes. The qualitative data produced by the Post-System Debriefing Interview is not discussed in this report.

2.2.3.12 Product: Cross Evaluation

The fourth component of each block, following the Post-System Debriefing Interview, was the Cross Evaluation. Each analyst reviewed (using a paper copy) all seven reports prepared for each of the two scenarios in the block (14 total reports). Analysts used an online tool to rate each report according to seven criteria. After analysts completed the independent ratings of each report according to the seven criteria, they were asked to sort the stack of reports into rank order, placing the best report at the top of the pile. Analysts were then asked to use a pen to write the appropriate rank number at the top of each report, and to use an online tool to enter their report rankings. The criteria that the analysts used were:

- Covers the important ground
- Avoids the irrelevant materials
- Avoids redundant information
- Includes selective information
- Is well organized
- Reads clearly and easily
- Overall rating

Analysts completed all Cross-Evaluation activities for the first scenario of the block before moving on to the evaluation of the second scenario of the block. The Cross-Evaluation exercise took place in the system room to which analysts had been assigned for the block. Analysts were allotted two hours to complete the Cross Evaluation exercise. After the first block, it was discovered that analysts could complete this exercise within 1:45, and thus, the time allotment was shortened. The Cross-Evaluation tool can be viewed in Appendix S.

2.2.3.13 Product: Mini-focus Groups

After the Cross Evaluation, “mini-focus groups” of four analysts were formed to discuss the results of the Cross Evaluation. The membership of these groups was designed to avoid having the two analysts who used the same system together in a group, and to allow for a sufficient rotation of analysts amongst the two groups. It should be noted that the number of participants for the mini-focus groups was uneven for blocks 1, 2, and 3 because there were only seven analysts. The eighth analyst was present for the last block and participated in the mini-focus group. Correspondingly, during the final Cross-Evaluation for the last block, analysts evaluated eight, rather than seven, reports.

The two groups were led by two of the observers (the Lead Observer and the Co-PI), and occurred in two of the largest system rooms. These mini-focus groups had two purposes: to develop a consensus ranking of the seven reports for each scenario, and to elicit the aspects, or dimensions, which led each analyst to rank a report high or low in overall quality. These discussions were taped using the Glass Box software; an additional observer, who took notes, was present in each room.

The group discussions were free-form in that there were no pre-determined questions used to guide the discussion. Instead, group leaders started each discussion by asking each analyst to call out their report rankings, and recording this information on a white board for everyone to see. The group leaders had different methods for accomplishing this. Recall that reports had random numbers assigned to them for identification purposes, and thus, a report could not be linked back to an individual, unless that individual volunteered the information. In Method 1, the observer listed the report numbers in numerical order on the white board, and then asked each analyst to

call out his/her rankings for each of the reports before moving to the next report. Thus, the report numbers were listed in the first column, and analysts' rankings were listed in the three (or four) subsequent columns. When analysts read their rankings aloud, the observer randomly ordered the rankings across the three (or four) columns so that no single column corresponded to any one analyst's rankings. In Method 2, the observer asked each analyst to read from his or her stack of reports in sequence, so that the first report identified was the report that they ranked as first. This method resulted in three (or four) lists of report identification numbers ordered from position one to seven (or eight). With Method 2, the rankings provided by an individual analyst appeared in the same column of the display.

From this information, the group worked out a collaborative ranking of the reports. Analysts were asked to discuss which elements lead them to rank a particular report in the way that they did. Groups were allotted 20 minutes to discuss the rankings of the reports for each scenario, for a total of 40 minutes. The qualitative data produced during this Cross-Evaluation Group Discussions have not been analyzed for this report.

2.3 System logs and Glass Box Infrastructure

The Glass Box software was developed by Battelle (PNNL) under the auspices of ARDA. It supports capture of analyst workstation activities including keyboard/mouse data, window events, file open and save events, copy/paste events, instant messaging, and web browser activity. The Glass Box makes extensive use of a relational database to store time-stamped events and a hierarchical file store where files and the content of web pages are stored. The Glass Box "snatches" a copy of every file the analyst opens and saves so there is a complete record of the evolution of documents. The material on every web page the analyst visits is explicitly stored so that each web page can be later recreated (by researchers) as it existed at the time it was accessed by the analyst; screen images are also captured. In addition, screen capture and audio capture are available.

The data captured by the Glass Box provides details about a user's interaction with normal desktop components, such as MS Office and Internet Explorer. User interaction with applications that do not run in a browser or even Java applications that may run in a browser are not sufficiently logged by the Glass Box. Although limited information, e.g. Window Title, application name, information copied to the system Clipboard, is captured by default, the quantity and quality of the information is not sufficient to serve as a log of user interaction with a system. In this study, GNIST was the only tool whose Glass Box logging was adequate; the other three systems needed to develop system logs to capture information.

During initial meetings with the System participants, a minimal set of logging requirements was developed and included:

- Time stamp
- Set of documents the user copied text from
- Number of documents viewed
- Number of documents that the system said contained the answer
- Analyst's query/question

This set of logged event would allow measurement of

- Number of queries
- Number of documents viewed
- Query paths
- Growth of documents

- Number of productive questions/queries, i.e. user input that results in a useful document being found as indicated by cutting and pasting from it or typing while it is being viewed

2.4 End-of-Workshop Activities

On the final day of the workshop, several data collection techniques were used to gather information about analysts' entire experiences using all of the systems, and about the various instruments and protocols of the study.

2.4.1 Scenario Difficulty Assessment

On the final day of the experiment, analysts completed the Scenario Difficulty Assessment. In this activity analysts rated each scenario on twelve dimensions, and also rank-ordered the scenarios according to level of difficulty. Analysts were also given six new scenarios that had not been used in this study, and asked to make the same evaluations. This paper-based activity occurred in the conference room. Analysts were given two hours to complete this activity. This assessment worksheet can be viewed in Appendix T.

2.4.2 System Discussions

After the Scenario Difficulty Assessment, analysts self-selected into two groups of four (all eight analysts participated in this activity), and visited each of the three experimental system developers in turn, for a 40-minute freeform discussion to provide feedback about the systems. [At any moment one developer was idle]. The spontaneously formed groups appeared to interact easily and openly with each of the system developers. An observer accompanied each group and served as a guide and timekeeper. When requested, the observer took notes for the system developer. The goal of these discussions was to provide feedback to the developers; this data is not used in this report.

2.4.3 Focus Group on Systems

In the afternoon, a focus group was held with all eight analysts. A senior project member, who was experienced in leading focus groups and who had only been present for the first two days of the experiment, led the group using a Focus Group Guide, a copy of which can be found in Appendix U. The audio from this focus group was recorded with a digital recorder as well as with a VCR (audio only, no video). Other persons present at this focus group included the four observers, who took notes during the session, and one other project person (a manager). The discussion lasted for one hour and twenty minutes. The principal finding was a strong consensus in favor of one of the systems. This consensus agrees with the results of the Cross Evaluation. A ranking of a second system was extracted from the group, which also seemed to agree with the results from Cross Evaluation. Detailed analysis of this data is not included in this report.

2.4.4 Discussion of the Entire Workshop

After the Focus Group on Systems, a Focus Group of the entire workshop was conducted to elicit feedback about the method of study and about analysts' and observers' experiences. This focus group was led by the same leader of the Focus Group on Systems, and was also recorded with a digital recording device as well as a VCR (audio only, no video). This discussion lasted for 45 minutes. One of the principal findings of this discussion was that analysts did not feel that the process of being observed distorted or biased their responses and behaviors. Observers also indicated that they learned a lot from their observations and that, for the most part, did not feel awkward observing the analysts while they worked. There are plans to transcribe the recording, and to analyze the data the further. However, analysis of this data is not included in this report. See Appendix V for a copy of the focus group guide.

3 Sources of error in the analysis reported here.

3.1 *Sampling errors*

Statistical uncertainties arise when a universe of possibilities is studied using a random sample from that universe. Hence, all statistical results are conditioned by the statement "if the analysts and tasks used are a random sample from the universe of relevant analysts and tasks". It is apparent that the **scenarios** are not a random selection among possible scenarios. They were tailored to the corpus. Similarly, the **analysts** are not a random sample of analysts, since they were gathered by an iterative recruitment process informally known as "snowball" recruitment.

The collection or corpus. The corpus consisted of a specialized collection of about 40,000 documents from CNS, which deals primarily with CBW. It was augmented by WWW searches using Google with queries that were focused on the domains represented in the scenarios. Hence, a larger number of documents were probably relevant to each scenario than would be found in sources such as the WWW.

The tasks. The tasks were selected from a larger set developed by two consultants from the Rome AFB. They were screened against the collection using an informal method of search on single keywords. Tasks for which there were "too few" retrievals were dropped. Thus the tasks have been, to some degree, customized to work with the collection. In a real working environment, we suppose that tasks and collection will already be in agreement.

The analysts. The analysts in this experiment were naval reservists, selected by a "Snowball method" of recruitment which is virtually certain to not produce a random sample. In real applications the decision maker evaluating a system should expend substantial effort to recruit analysts typical of those who will be using the system. Demographic information about the analysts appears in Appendix A. There was some large variation in the experience of analysts in using computer systems.

3.2 *Statistical error.*

Statistical confidence levels are used to assess whether an observed difference is meaningful. This is expressed in terms of the confidence level. It is customary to require confidence levels of 95% or higher, although in certain types of research, such as epidemiology, lower confidence levels are accepted because it is so important to detect an outbreak of disease that one willingly risks a larger number of false alarms. As noted above, in this workshop several of the mathematical prerequisites for the use of statistical theory have not been met. Thus the computed confidence intervals or measures of significance are informative, but not rigorously interpretable.

Generally 95% confidence in a conclusion such as "system A is superior to system B" means "if there were no difference between the systems, and many studies of them were done, using random selections of tasks and analysts, only 5% of those many replicates would yield an effect as large as the one observed here". Informally, there is only a 5% chance that we would see such a large effect if it was not real.

In this study, many possible relations were explored. Relations found to be statistically significant when multiple relations are examined must either have a higher apparent level of confidence (the Bonferroni approach) or must have been planned in advance. The most important finding (influence of system on the leading factor in product evaluation) was planned in advance, and has been analyzed using the rigorous Scheffe test. For most other results no multiple hypothesis correction was applied, and the results are exploratory rather than conclusive.

3.3 Systematic Error

Systematic error occurs when the method of measurement has an inherent bias in some direction. Much systematic error is controlled by considering differences between cases, rather than interpreting absolute numbers. However, when two or more effects are "confounded" it is not possible to know whether an observed difference is due to one of the effects or the other. The only apparent confounding in this experiment is that the physical location (which ranged from a small windowless room to large windowed room) may have an effect on the performance of an analyst, and on the analyst's behavior when judging either system or product. The systems were fixed in the rooms, and so any effect of the physical surround is confounded with the identity of the system, and will be, in the analyses reported here, attributed to the system.

4 Principal Findings

Table 5 shows the 15 hypotheses investigated in this workshop. For each of these hypotheses, we show the type of data collection method that was used to determine if there was evidence to support the hypothesis. In this section, we'll discuss each method and the findings.

Table 5: Matrix of hypotheses and methods of investigation

	Question answering systems should	Questionnaires	NASA TLX	SmiFro Status	Cross-evaluation	System Logs	Glass Box	Query Trails
H1	Support information gathering with lower cognitive workload		X			X	X	X
H2	Assist in exploring more paths/hypotheses	X						X
H3	Enable production of higher quality reports	X			X			
H4	Provide useful suggestions to the analyst	X				X	X	
H5	Provide more good surprises than bad	X		X				
H6	Enable more focus on analysis than data collection	X						
H7	Enable analysts to collect more data in less time	X					X	
H8	Reduce the time spent reading	X					X	
H9	Identify gaps in the knowledge base	X				X	X	
H10	Help the analyst recognize gaps in their thinking	X						
H11	Provide context for information	X				X		
H12	Provide context, continuity and coherence of dialogue	X				X	X	X
H13	Let analysts relocate previously seen materials	X						
H14	Be easy to use	X	X					
H15	Increase an analyst's confidence in exploration and report	X		X				

4.1 Cross Evaluation

The results were analyzed in terms of the individual questions, and were also examined using factor analysis. The leading factor of the cross-evaluation of the reports is proposed as the "gold

standard" measure of the quality of the product. In the analysis reported below this is found to be a consistent and meaningful single measure of report quality.

As mentioned previously, the seven components are:

- Covers the important ground
- Avoids the irrelevant materials
- Avoids redundant information
- Includes selective information
- Is well organized
- Reads clearly and easily
- Overall rating

4.1.1 Factor Analysis

If the instrument has a balanced set of questions that accurately reflect the decision maker's concerns, then factor analysis is the best way to summarize the set of questions. Since there is no decision maker at hand, we have analyzed the data using factor analysis, as a reasonable default. The reader is cautioned, however, that if a specific aspect of system quality is over-represented in the array of questions, the "leading factor" will be biased toward that particular aspect of system quality.

The key findings that emerge from factor analysis are:

- (1) the seven elements of the cross-evaluation are primarily explained by a single leading factor which accounts for 79% of the total variance
- (2) the 33 system evaluation items require seven factors with eigenvalue (a measure of importance) greater than 1. Together they account for 88% of the observed variance. The first single factor accounts for 42% of the variance and the first four together account for 68% of the variance. It is noteworthy that each of these factors places a different one of the four systems in the first position. The system effect

4.1.2 The System Effect

To see the effect of systems on the overall quality of the product, or on the evaluations provided by the users, we must control for several effects: the effect of the task, the effect of the user, the effect of the observer, and the effect of self-judgment bias. This is done using a model called the General Linear Model. The General Linear Model represents the score assigned by a judge, to a report, as a linear combination of effects representing these factors.

$$\begin{aligned} \text{Score_on_Report (judge, author, system, task, self_bias, observer)} = \\ \text{Constant +} \\ \text{Judge_effect +} \\ \text{Author_effect +} \\ \text{System_Effect +} \\ \text{Task_Effect +} \\ \text{Self_Bias_Effect +} \\ \text{Observer_Effect} \end{aligned}$$

The relative magnitudes of these effects are shown in Table 6. The range is the difference between the highest and lowest values for the particular effect. The effects of system and scenario are smaller than all but the observer effect. The strength of conclusions about the

system is therefore limited. All differences refer to numbers computed on a 5-point scale whose maximum difference would be 4 units.

The experimental design minimizes the variance of the estimates of these several effects, which makes the design efficient. The experimental design precludes study of possible interaction effects, which would require a substantially larger factorial design.

Table 6: Relative magnitudes of various potential effects on Cross-Evaluation Results

Effects	Range
Judge	2.30
Self-judgment bias	2.01
Author	1.21
System	0.45
Scenario	0.43
Observer	0.21

It is apparent that extraction of the system effects depends on using a model that accounts for the large effects of judge, author and task, as well as the self-judgment bias.

Since the goal of this work is to develop metrics for systems, we concentrate on the System_effect. We find that it successfully separates the systems into two groups. One group contains a single system, and the other contains the other three systems. The more rigorous post-hoc Scheffe test shows that the difference between the top and the third (or the fourth system) is significant. The second system cannot be distinguished from any of the other three. The third and fourth are indistinguishable. These results are shown in Table 7.

Table 7: Effect of factors on the summary measure ("Factor_1") of the seven cross-evaluation scores

Parameter		Value	Sigma	t-Statistic	p-value	95% Confidence	
						Lower Bound	Upper Bound
Intercept		1.566	0.309	5.062	0	0.958	2.174
Scenario Effect							
SCENARIO=H	a	0					
SCENARIO=D		0	0.158	0.001	0.999	-0.311	0.311
SCENARIO=F		-0.016	0.158	-0.102	0.919	-0.327	0.295
SCENARIO=E		-0.228	0.158	-1.441	0.15	-0.539	0.083
SCENARIO=B		-0.359	0.158	-2.27	0.024	-0.67	-0.048
SCENARIO=C		-0.381	0.158	-2.407	0.017	-0.691	-0.07
SCENARIO=G		-0.39	0.156	-2.495	0.013	-0.697	-0.082
SCENARIO=A		-0.426	0.158	-2.695	0.007	-0.737	-0.115
Judge Effect							
JUDGE=ch4		0.141	0.418	0.336	0.737	-0.682	0.963
JUDGE=ch8	a	0					
JUDGE=ch1		-0.588	0.418	-1.405	0.161	-1.41	0.235
JUDGE=ch7		-0.933	0.418	-2.229	0.026	-1.755	-0.11
JUDGE=ch2		-1.475	0.418	-3.526	0	-2.297	-0.652
JUDGE=ch3		-1.9	0.418	-4.544	0	-2.723	-1.078
JUDGE=ch5		-2.16	0.418	-5.163	0	-2.982	-1.337
Author Effect							
ANALYST=ch7		0.275	0.16	1.721	0.086	-0.039	0.59
ANALYST=ch5		0.091	0.16	0.568	0.57	-0.224	0.406
ANALYST=ch8	a	0					
ANALYST=ch3		-0.068	0.16	-0.428	0.669	-0.383	0.246
ANALYST=ch2		-0.132	0.16	-0.825	0.41	-0.447	0.183
ANALYST=ch4		-0.385	0.16	-2.408	0.017	-0.7	-0.071
ANALYST=ch1		-0.937	0.16	-5.857	0	-1.252	-0.623
System Effect							
SYSTEM=		0.329	0.113	2.908	0.004	0.107	0.552
SYSTEM=	a	0					
SYSTEM=		-0.064	0.113	-0.566	0.572	-0.287	0.158
SYSTEM=Gnist		-0.12	0.113	-1.06	0.29	-0.342	0.103
Observer Effect							
OBSERVER=	a	0					
OBSERVER=		-0.01	0.113	-0.089	0.929	-0.233	0.212
OBSERVER=		-0.162	0.113	-1.431	0.153	-0.384	0.061
OBSERVER=		-0.21	0.113	-1.853	0.065	-0.432	0.013
Self-Judgment Effect							
[JUDGE=ch8]SelfBias		0.971	0.319	3.040	0.003	1.599	0.343
[JUDGE=ch1]SelfBias		0.791	0.319	2.476	0.014	1.419	0.163
[JUDGE=ch4]SelfBias		0.649	0.319	2.032	0.043	1.277	0.021
[JUDGE=ch2]SelfBias		0.608	0.319	1.904	0.058	1.237	-0.020
[JUDGE=ch7]SelfBias		-0.155	0.319	-0.486	0.627	0.473	-0.784
[JUDGE=ch3]SelfBias		-0.772	0.319	-2.418	0.016	-0.144	-1.401
[JUDGE=ch5]SelfBias		-1.037	0.319	-3.248	0.001	-0.409	-1.666

Note (a): One of the values must be arbitrarily set to zero, as only differences are meaningful

For each effect, the values of that factor have been arranged in decreasing order by their effect on the Factor_1 score. For example, in "Author effect" the Analyst called ch7 had the highest effect on the overall Factor_1 (roughly, wrote the best reports). The second column shows the standard error of this estimate (0.16). The third shows the t-statistic. The fourth shows the chance that this difference from 0 would occur by chance (8.6%). The fifth column shows the lower edge of the 95% confidence interval (-0.039), and the 6th shows the upper edge. Roughly this

means that the first value which could not be the same as that for ch7 is the value -0.068, for ch3. Such results can be shown graphically.

For clarity we have rerun the analysis using a varied name for the GNIST system so that it is presented at the arbitrary value of 0 (Table 8 and Figure 2). This seems appropriate since it is lower than the other systems on the Factor_1 score.

Table 8: Rearranged values for the effect of the system on Factor_1

Factor_1	System
0.449	Γ
0.120	A
0.056	B
0	GNIST

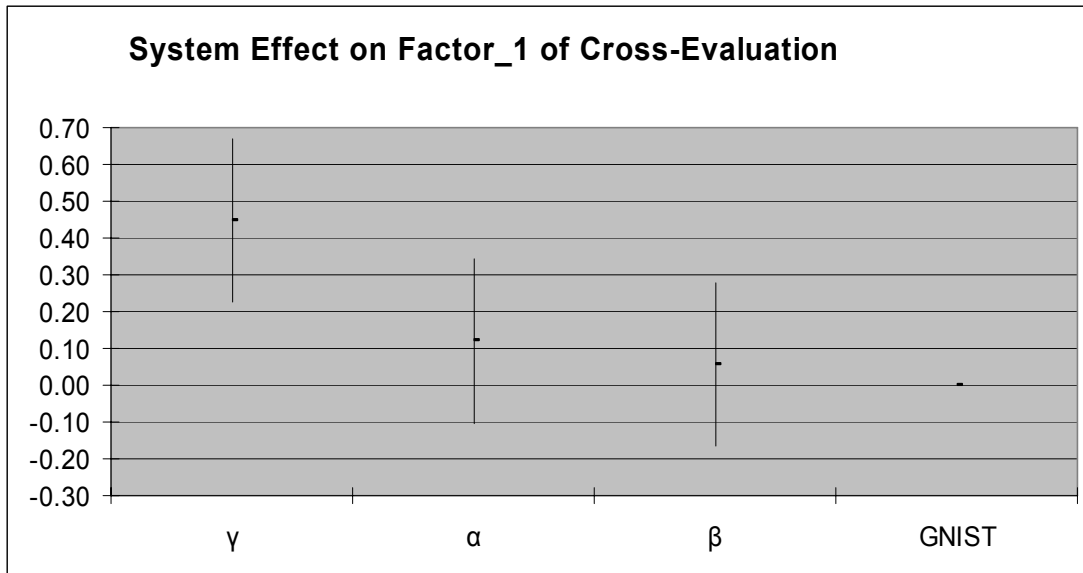


Figure 2: System effect on Factor_1

We see that all three systems score above the baseline GNIST system. The bars represent the 95% confidence intervals. We see that with 95% confidence the system labeled “gamma” is different from the system labeled GNIST. The other two systems are not different at that level of confidence from GNIST. They are significantly below the system labeled “gamma”.

Finding: The cross evaluation method with careful statistical analysis, can resolve distinctions among the prepared systems as tested here.

This is a weaker result than one might wish for, but it does validate the methodology. We return to the problem of scaling these results in more meaningful terms in Section 5.1.

4.1.3 Individual Aspects of the Product

The same decomposition of effects can be applied to scores on each of the seven individual factors of the cross-evaluation.

Finding: We find that five of the seven product criteria are able to separate at least one of the systems from at least one of the other experimental systems. All seven are able to separate at least one of the experimental systems from the baseline system, GNIST.

While one might have wished for a stronger separation of the experimental systems, this result validates the method of data collection and analysis.

4.1.4 Cross-evaluation Validation

In any experimental study of metrics, the crucial questions are: (1) are the metrics reliable (yielding the same value over and over again)? (2) are the metrics valid (do they assess the appropriate qualities of the thing being measured)? and (3) how accurate or sensitive are the metrics. In the sections below we discuss the reliability, validity, and sensitivity of the cross-evaluation metrics.

4.1.4.1 Reliability

In this study we have a single instance of measurement of a set of four systems. Thus the issue of reliability cannot be fully addressed. We can say, however, that the statistical analysis yields results that are fully consistent with the existence of measurable properties of the systems that are measured in this process.

4.1.4.2 Validity

The specific seven criteria used in the cross evaluation were developed in discussions with analysts, and were seen as reasonable by the analysts participating in the Challenge Workshop.

4.1.4.3 Accuracy or sensitivity

Using the hypothesis that a system should produce measurable improvements in the work product of the analyst when using it, how accurately can cross-evaluation distinguish between different systems? The data gathered here use the 7 scales used to measure the quality of work product. Each product was measured 7 times, on each scale, with each measurement being provided by a different analyst. When analyzed in a "non-parametric" or model-free way, these show some measurable effects.

4.1.4.3.1 Non-parametric measurement using scales

The seven specific scales used to assess documents can be thought of as the seven games of a "World Series" competition. The 7 documents prepared, and 7 judgments made on each, produce an average score both for the GNIST system, and for each other system. In some cases the average score for the other system is higher than GNIST, in others lower. Viewed as such a series competition, the results show that one system beats GNIST 7 times; another wins 5 and loses 1, with 1 tie, while the third wins 4 and loses 3. When compared with a "null hypothesis" which is the equivalent of saying that the outcome is determined by the toss of a coin, the chance of the first result is less than 1%. The chance of the second result is about 23%, if there were no difference, while the chance of the third is 50%. Thus the first result would be deemed significant at the "99%" confidence level, if the seven scales could be thought of as independent trials.

4.1.4.3.2 Parametric Measurement and Factor Analysis

A more sophisticated approach asks whether the scores on the several scales are related to some underlying factors. This "factor analysis" reveals strong correlations among the scores, with one underlying factor accounting for most of the variation. This factor can be interpreted as the "overall quality" of the report. Thus the compound procedure (a) conduct the experiment, and (b) apply factor analysis, seems to produce a valid measure of the proposed quality: the ability of the system to improve the work product of the analyst.

Findings: When systems are as different as the systems tested here, and are prepared as were the systems here, the cross-evaluation method and the questionnaire method are able to detect differences between systems at a statistical confidence level exceeding 95% confidence. The

methods of analysis are sophisticated, but standard, and easily replicated. Very large amounts of objective, scale-based, and qualitative data were collected, and form a resource for further investigation.

Conclusion: Since cross evaluation is an expensive method of measurement, it is of interest to ask whether less expensive methods such as questionnaires provide an adequate estimate of the more rigorous evaluation based on "product". Finally, since all of these methods depend on work by analysts, spending at least 5 hours with each system, it is of interest, to developers, to ask whether some of the information in the system logs can be used to estimate performance, without requiring the use of analysts at all. The actual cross-evaluation took less than an hour of analyst time, but takes time and expertise to analyze and forces the experiments to be conducted simultaneously.

4.2 Post-Session and Post-System Questionnaires

Using the same approach, we find that many of the questions on the Post-Session and Post-System Questionnaires are able to resolve the systems into two groups, showing that, with this design, they are able to provide useful metric information on the systems. In examining the following results, the reader is reminded that systems are referred to as how they ranked on a particular question, rather than their particular name or identification code. This is not necessarily fixed, as different systems might be ranked differently according to the question being asked.

In Tables 9 and 10, these questions are shown, grouped in relation to the original hypotheses, and ranked in terms of the statistical significance of the distinctions that they are able to reveal. The rankings in the tables reflect the highest score for each question. The best value for the question is dependent on the wording of the question. In most cases, the best value is the high score, but in some cases the best value is the low score (this is scale dependent). It should be noted that a column does not represent all scores for a single system. Instead, scores are presented according to rank, regardless of which system they were associated. The score for the baseline system (GNIST) is shaded. When significant, the F-value and degrees of freedom from the General Linear Model is provided along with the probability value (p-value). The values in the column labeled "Diff" are the results of the Scheffe's pair-wise comparison tests.

Table 9: Results of the Post-Session Questionnaire

H2: QA system should assist the user in exploring more paths/hypotheses.							
RANK:	1	2	3	4	F(3,58)	p-value	Diff.
Q12: If you had to perform a task like the one described in the scenario at work, do you think that having access to the [X] system would help you find information that you usually have trouble finding? [not at all ... a lot]	4.00 (.68)	3.07 (1.28)	2.79 (1.25)	2.47 (.64)	7.62	.000	1 > 2, 3, 4
H3: QA systems should enable analysts to produce higher quality reports.							
RANK:	1	2	3	4	F(3,58)	p-value	Diff.
Q4: If you had to perform a task like the one described in the scenario at work, do you think that having access to the [X] system would help answer specific questions that you are currently having trouble answering? [not at all ... a lot]	4.14 (.66)	3.13 (1.13)	2.87 (.64)	2.56 (1.28)	7.66	.000	1 > 2, 3, 4

Q8: In general, were the answers that the system provided helpful in meeting the goals set forth in the scenario? [unhelpful ... helpful]	4.14 (.53)	3.13 (1.06)	3.00 (1.24)	2.67 (1.18)	6.17	.001	1 > 2, 3, 4
H7: QA systems should allow the analyst to collect more data in less time.							
RANK:	1	2	3	4	F(3,58)	p-value	Diff.
Q2: In comparison to other systems that you normally use for work tasks, how would you assess the length of time that it took to perform this task using the [X] system? [less time ... about the same ... more time]	2.21 (.97)	3.43 (1.28)	3.73 (.70)	3.73 (1.22)	9.21	.000	1 < 2, 3, 4
Q13: If you had to perform a task like the one described in the scenario at work, do you think that having access to the [X] system would help increase the speed with which you find information? [not at all ... a lot]	4.07 (.92)	3.00 (1.20)	2.50 (1.29)	2.27 (.70)	10.827	.000	1 > 2, 3, 4
H9: QA systems should provide the analyst with identification of gaps in the knowledge base.							
RANK:	1	2	3	4	F(3,58)	p-value	Diff.
Q10: After using the [X] system, it is clear to me that there is more information that I need to know about the scenario that I cannot find in the collection.	2.93 (.92)	3.36 (1.01)	3.47 (.99)	3.60 (.99)	-	-	-
H10: QA systems should help the analyst recognize gaps in their thinking.							
RANK:	1	2	3	4	F(3,58)	p-value	Diff.
Q3: The [X] system helped me to better understand the scenario by the information that it provided.	3.79 (.80)	3.13 (1.06)	2.79 (.89)	2.53 (.83)	5.54	.002	1 > 2, 3, 4
Q11: The [X] system stimulated my thinking about the scenario.	3.79 (.70)	3.20 (.77)	3.00 (1.24)	2.47 (.83)	5.86	.002	1 > 3, 4
Q15: The [X] system expanded my understanding of the scenario.	3.71 (.83)	3.13 (.83)	3.07 (1.21)	2.60 (.51)	4.28	.009	1 > 4
Q7: The [X] system helped me to think about the scenario in new ways.	3.64 (.93)	3.13 (.99)	2.79 (.97)	2.53 (.83)	3.96	.013	1 > 3, 4
H12: QA systems should provide coherence and continuity for the user.							
RANK:	1	2	3	4	F(3,58)	p-value	Diff.
Q6: In general, how appropriate were the answers that the system provided for the questions that you asked? [inappropriate ... appropriate]	4.00 (.78)	3.33 (1.11)	2.86 (1.41)	2.73 (.96)	4.82	.005	1 > 3, 4
H13: QA systems should allow analyst to locate a specific document or piece of information previously seen.							
None							

H14: QA systems should be easy to learn and use.							
RANK:	1	2	3	4	F(3,58)	p-value	Diff.
Q5: In general, how did you find formulating questions that resulted in useful responses from the system? [easy ... difficult]	2.00 (.96)	2.73 (.88)	3.53 (.64)	3.71 (1.27)	13.90	.000	1 < 2, 3, 4
Q14: It was easy to gather relevant information about this scenario with the [X] system.	4.14 (.66)	3.13 (1.30)	2.79 (1.37)	2.53 (.92)	6.44	.001	1 > 2, 3, 4
H15: QA systems should make its user confident in the exploration of the available data, and in the report produced as a result.							
RANK:	1	2	3	4	F(3,58)	p-value	Diff.
Q9: I felt that the system showed me all of the available relevant information about this scenario.	3.57 (.85)	2.87 (1.13)	2.57 (1.02)	2.20 (.68)	7.15	.000	1 > 2, 3, 4
Q1: How confident were you of your ability to make the [X] system work to accomplish the assigned tasks? [unconfident ... confident]	4.43 (.65)	3.80 (1.08)	3.47 (1.19)	2.86 (1.35)	5.00	.004	1 > 3, 4

Table 10: Results of the Post-System Questionnaire

H2: QA system should assist the user in exploring more paths/hypotheses.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q10: The system helped me think of new ways to search for information.	3.86 (.69)	3.14 (.90)	3.14 (1.07)	2.29 (.49)	3.90	.030	1 > 4
Q1: The system allowed me to easily change my line of questioning.	4.00 (.58)	3.43 (.98)	2.71 (1.11)	2.29 (1.25)	3.51	.042	1 > 3, 4
Q22: The system often presented me with novel information.	3.49 (.79)	3.00 (.82)	2.86 (.69)	2.29 (.49)	3.11	.058	1 > 4
Q15: The system allowed me to easily change my search strategy.	3.57 (.96)	3.29 (1.38)	3.14 (1.21)	3.14 (1.21)	-	-	-
Q27: Having the system at work would help me find information that I cannot currently find.	3.29 (1.25)	2.86 (1.07)	2.86 (1.07)	2.43 (.53)	-	-	-
H3: QA systems should enable analysts to produce higher quality reports.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q9: The system helps me find important information.	4.29 (.49)	3.43 (.79)	2.86 (.69)	2.86 (1.25)	5.78	.008	1 > 3,4
Q32: The system often presented me with redundant information.	2.71 (1.11)	3.57 (1.40)	3.71 (1.11)	4.14 (.90)	3.91	.030	1 < 4 2 < 4
Q18: I could not find enough documents with relevant information.	2.00 (.58)	2.86 (1.07)	3.43 (.98)	3.71 (1.50)	3.57	.040	1 < 3,4
H4: QA systems should provide useful suggestions to analysts.							

RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q11: The system provided me with suggestions about which I had not thought.	3.57 (.98)	3.29 (1.50)	2.57 (.53)	1.86 (.69)	3.68	.036	1 > 4 2 > 4
H5: QA systems should provide more good surprises than bad.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q24: The system often surprised me by doing bad things.	1.86 (.69)	2.71 (.76)	3.00 (.58)	3.29 (1.50)	3.45	.044	1 < 3, 4
Q7: The system often surprised me by doing good things.	3.14 (.69)	3.00 (.58)	2.71 (1.25)	2.57 (.53)	-	-	-
H6: QA systems should allow the analyst to focus more on analysis, higher level, than to focus on data collections efforts.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q13: I spent more time reading documents than I did analyzing the information that the system provided.	3.14 (.69)	3.43 (.79)	3.43 (1.51)	4.14 (.69)	-	-	-
H7: QA systems should allow the analyst to collect more data in less time.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q23: Having the system at work would help me find information faster than I can currently find it.	4.00 (.82)	2.57 (.96)	2.57 (.96)	2.43 (1.13)	5.95	.007	1 > 2, 3, 4
Q6: The system slows down my process of finding information.	2.43 (1.27)	3.00 (1.63)	3.14 (.69)	3.42 (1.40)	-	-	-
H8: QA systems should allow the analyst to reduce reading time.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q3: I spent more time reading documents than I did searching for information.	3.43 (1.40)	3.86 (.38)	4.14 (.90)	4.71 (.49)	-	-	-
H11: QA systems should provide relevant context for information.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q5: Most of the time, I understood the relationship between the question that I asked and the answer that the system provided.	4.00 (1.15)	3.57 (.98)	3.00 (.82)	3.00 (1.00)	3.57	.040	1 > 3,4
H12: QA systems should provide coherence and continuity for the user.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q25: The system provided coherent answers.	4.00 (.82)	3.00 (.58)	3.00 (1.00)	2.86 (.69)	4.23	.024	1 > 2, 3,4
Q8: The system allowed me to navigate easily between searching activities.	4.57 (.53)	3.57 (.98)	3.00 (1.15)	3.00 (1.73)	4.08	.026	1 > 3,4
Q20: The system did a good job of helping me to integrate the materials that I found while searching.	3.86 (.69)	3.57 (1.13)	2.86 (1.21)	2.29 (.76)	-	-	-
H13: QA systems should allow analyst to locate a specific document or piece of information previously seen.							

RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q29: It was easy to re-examine my previous searching activities with the system.	4.43 (.79)	3.57 (1.13)	3.57 (1.13)	3.00 (.58)	-	-	-
H14: QA systems should be easy to learn and use.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q19: The system is hard to use.	1.29 (.49)	2.14 (.90)	2.43 (1.27)	3.00 (.82)	4.44	.020	1 > 3,4
Q28: The system was inflexible.	2.00 (1.00)	2.86 (1.25)	3.00 (.82)	3.57 (1.27)	4.31	.022	1 > 3,4
Q2: It was difficult to get the system to do what I wanted it to do.	2.14 (.69)	2.86 (.90)	3.14 (.90)	3.29 (1.11)	-	-	-
Q14: My skill at using the system improved over the course of the workshop.	4.29 (.49)	4.00 (.58)	4.00 (1.00)	3.57 (1.27)	-	-	-
Q21: I feel that I have become pretty proficient at using the system.	4.29 (.76)	3.71 (1.11)	3.43 (1.27)	3.00 (.82)	-	-	-
Q33: In general, I like using the system.	4.43 (.53)	3.14 (1.35)	3.14 (1.46)	3.00 (.82)	-	-	-
H15: QA systems should make its user confident in the exploration of the available data, and in the report produced as a result.							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q4: I often wanted to verify the system's answers.	2.43 (.96)	3.00 (.82)	3.43 (1.27)	3.43 (1.27)	-	-	-
System Readiness							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q12: The system is not ready yet to be used in my regular work environment.	2.14 (1.07)	2.71 (.95)	3.43 (1.13)	4.14 (.90)	6.15	.006	1 > 3,4
Q16: The system would be a useful addition to the tools that I already have at work.	4.43 (.53)	3.29 (.49)	3.00 (1.41)	2.86 (1.21)	4.40	.021	1 > 2, 3, 4
Q17: The system would let me stop using some of the tools that I currently use at work.	2.71 (1.12)	2.29 (.76)	2.14 (1.07)	1.86 (.90)	-	-	-
Training							
RANK:	1	2	3	4	F(3,27)	p-value	Diff.
Q30: The training materials were hard to understand.	1.57 (.79)	1.57 (.98)	2.43 (1.13)	3.00 (.58)	5.09	.013	1 > 4 2 > 4
Q26: The training on the first day gave me the skills needed to use the system successfully.	4.57 (.53)	4.29 (.95)	3.57 (1.40)	3.00 (1.15)	3.21	.053	1 > 4 2 > 4
Q31: The training materials contained most of the information I needed to learn to use the system.	4.14 (.69)	4.14 (.90)	3.71 (.95)	3.14 (1.06)	-	-	-

The results from the Post-Session Questionnaire are the strongest, as judged by statistical significance and p-values, and show support for a full six of the seven hypotheses for which they were designed to investigate. Indeed, statistically significant differences were found for fourteen of the fifteen items on the questionnaire, with p-values equally less than .01 in all cases but one, where it was .013. In nine of these fourteen cases, the post-hoc analysis identified the first-ranked system as receiving significantly higher scores than the other systems. In the remaining five cases, the post-hoc analysis demonstrated differences between the first-ranked system, and the third- and fourth-ranked systems. Overall, these results provide evidence that the first-ranked system assisted the user in exploring more paths/hypotheses (H2), enabled analysts to produce higher quality reports (H3), allowed analysts to collect more data in less time (H7), provided analysts with identification of gaps in the knowledge base (H9), helped analysts recognize gaps in their thinking (H10), provided coherence and continuity for analysts (H12), was easy to learn and use (H14), and made the user confident in the exploration of the available data and in the report produced as a result (H15).

The results from the Post-System Questionnaire were also somewhat strong, although weaker than those produced by the Post-Session Questionnaire. Results demonstrated statistical support for eight of the twelve hypotheses. Specifically, statistical differences between systems were detected for thirteen of the twenty-seven questions that assessed various aspects of the system. The results were weaker than those from the Post-Session Questionnaire, with p-values between .05 and .01 for eleven of the thirteen questions. Moreover, the post-hoc analysis usually only demonstrated significant differences between the first-ranked system and the third- or fourth-ranked system, rather than identifying the first-ranked system as significantly different from all other systems. Overall, these results provide some evidence that the first-ranked system assisted the user in exploring more paths/hypotheses (H2), enabled analysts to produce higher quality reports (H3), provided useful suggestions to analysts (H4), provided more good surprises than bad (H5), allowed analysts to collect more data in less time (H7), provided relevant context for information (H11), provided coherence and continuity for analysts (H12), was easy to learn and use (H14).

In addition to these questions, there were six Post-System questions that assessed system readiness and the training that the system developers provided. For these additional questions, statistical differences were detected between systems for two of the three questions about system readiness and for two of the three questions about training. System readiness assessed the extent to which analysts felt that the system was ready to be used in their real work environments. Results demonstrated statistical support ($p=.006$, $.021$) that the first-ranked system was closer to being ready to be used in this environment than the other systems. The questions about training asked analysts to evaluate their experiences in the training sessions. Statistical differences were found between the perceived quality in the training sessions, although these differences were not very strong ($p=.013$, $.053$). The post-hoc analysis demonstrated differences between the first-ranked system, and the fourth-ranked system, and the second-ranked system and the fourth-ranked system. It is worth noting that the first-ranked system in each of these cases was the baseline (i.e. Google), so this result is not that surprising.

Finding: The Post-session and Post-System Questionnaires were able to resolve the systems into two groups.

4.3 *SmiFro Console and Status Questionnaires*

The cross tabulation of the use of the SmiFro Console by analyst and by system is shown in Table 11. There was a wide variation in the number of times that the analysts used the tool. Analysts ch1 and ch7 made a total of 3 and 6 uses of the Console across the eight 2.5 analytic sessions. Although the other analysts contributed more indicators, there were insufficient numbers to allow any meaningful analysis. On average, there were nearly twice as many negative (frowny) as positive (smiley) submissions. Even this number must be viewed cautiously since analyst ch3 with system α and analyst ch2 with system β used the frowny indicator

repetitiously during a short period of time; therefore, roughly 28 of the 96 frowny records are related to 2 instances of system misbehavior.

During the final outbrief, we explored with the analysts their use or non-use of the SmiFro Console. Two primary reasons emerged: 1) the smile and frown icons were 'silly' causing the analysts not to take them seriously; 2) the analysts' military background doesn't match well with giving this type of feedback.

Table 11: Summary of SmiFro Console Activity

system	Counts	ch1	ch2	ch3	ch4	ch5	ch7	ch8	Total
α	Smiley		4	1	3	2		1	11
	Frowny	3	7	12	2		4	2	30
	Comments		1	1	2				4
β	Smiley		1	1	4	7		1	14
	Frowny		16	2	3	1	1	10	33
	Comments					1			1
γ	Smiley		1	1	2	8		4	16
	Frowny		5	3	1	4	1	6	20
	Comments			3	2				5
GNIST	Smiley		1			8		5	14
	Frowny		5	1	4			3	13
	Comments								0
Total – smiley		0	7	3	9	25	0	11	55
Total – frowny		3	33	18	10	5	6	21	96
Total – comments		0	1	4	4	1	0	0	10

The Status Questionnaires were administered every 30 minutes during each analytic exercise. A total of 229 questionnaires were received. This qualitative data from the SmiFro comments and from the status questionnaires were collated for each system. Each system team was provided with copies of the feedback for their system and for the GNIST baseline system. No further additional analysis was made of the data.

4.4 System Logs

During the Kick-Off meeting, the participating teams developed a set of required logging elements. These included:

- time stamp
- User's request – question or query
- List of documents the user copied text from
- Number of documents the user viewed
- Number of documents that the system returned in response to the user's request

Data for the GNIST system was derived from the Glass Box database with no additional system logging.

Table 12 shows the mean values for the data recorded by the systems. Analysts asked more questions when using the GNIST system than when using the QA systems. 'Good' questions were defined as questions that led to a document from which text was copied. GNIST had the lowest ratio of 'good' questions to total questions – 28%. The values for the α, β, and γ systems

were 39%, 58% and 57%, respectively. The data are shown graphically in Figure 3. Analysts viewed significantly more documents when they used GNIST than when they used QA systems. The final two columns of the table show that analysts copied from more documents using the γ system and they also copied more text fragments.

Table 12: Comparison of logged data for all systems (mean \pm SEM)

system	Questions	'Good' Questions	Docs returned	Docs Viewed	# copied from	# copy events
GNIST	16.71 \pm 3.09	4.64 \pm 0.70	577 \pm 130	75.64 \pm 9.68	11.07 \pm 1.83	22.50 \pm 4.24
α	8.64 \pm 1.5	3.36 \pm 0.86	71 \pm 11	-- ^a	5.36 \pm 0.90	12.79 \pm 2.87
β	8.71 \pm 1.3	5.07 \pm 0.56	75 \pm 10	26.71 \pm 2.93	11.71 \pm 1.20	15.64 \pm 1.60
γ	12.71 \pm 2.11	7.21 \pm 1.14	1191 \pm 214	37.07 \pm 4.84	18.21 \pm 2.30	33.50 \pm 4.09

^aData not logged.

Figure 3: Ratio of 'Good' Question to Total Questions

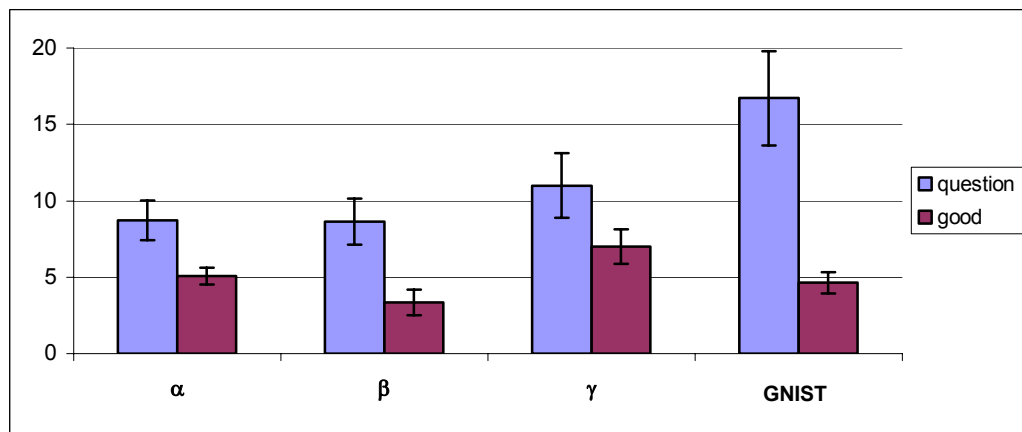


Table 13 shows a more detailed report for the GNIST system. This data is shown to give an idea of the variability in the data. Two additional columns are shown in this table related to 'unique' documents. Within a scenario, the queries that analysts submit are tightly constrained and the corpus contains a limited set of potentially documents. Therefore, we were interested in knowing how often analysts viewed the same document more than once. The data show that 87% of the time analysts viewed different documents and only 13% of the time did they view a document that they had seen before. We believe that this indicates that the corpus was sufficient for the analysts to accomplish their tasks without running out of material. We would have expected a much higher rate of repeat documents if the collection were too sparse in relevant documents.

Table 13: GNIST (baseline) data

analyst	Task	Questions	'Good' Questions	Docs returned	Docs viewed	Unique docs	% unique	#docs copied from	# copy events
ch1	C	17	3	260	27	20	74%	3	3
ch1	D	13	1	180	28	27	96%	1	1
ch2	A	32	10	580	72	66	92%	19	31
ch2	B	4	4	220	99	89	90%	26	58
ch3	G	43	7	1360	111	90	81%	8	13
ch3	H	32	8	1200	125	113	90%	13	28
ch4	E	22	8	380	143	129	90%	15	28
ch4	F	8	3	330	75	54	72%	11	28
ch5	G	9	5	290	85	82	96%	16	29
ch5	H	10	4	230	42	36	86%	7	16
ch7	C	11	4	600	46	39	85%	9	24
ch7	D	16	3	1700	52	49	94%	15	41
ch8	A	12	2	510	99	86	87%	3	3
ch8	B	5	3	240	55	47	85%	9	12
	<i>mean</i>	16.71	4.64	577	75.64	66.21	87.0%	11.07	22.50
	<i>sem</i>	3.09	0.70	130	9.68	8.77	2.0%	1.83	4.24

Analysis of system logs was tedious and time-consuming. Each system's logs were written in its own proprietary format, which meant that analysis could not be distributed easily across the teams. In addition, it should be noted that even though requirements were set for what needed to be logged, not all required data was logged. In some instances, time-stamps were not recorded and in another, the attempt to log when users viewed documents failed. The lack of time-stamps was addressed by correlating Glass Box capture data with the system's data after the fact to assign time stamps.

4.5 Glass Box Data

Access to Glass Box data allows us to make a variety of interesting and useful measurements. In this section, two such examples are shown – time allocation and activity trails. In addition, we have measure document growth over time by using the keystrokes targeted at the Word application and adding the number of characters from each copy event.

4.5.1 Allocation of time

The Glass Box detects which window on the computer display has the focus. The following figure shows a summary across scenarios and across analysts of the time (minutes) spent in the test system versus the time spent in MS Word. These were the main applications used by the analysts in this study. The overall time allotted for each exercise was 2.5 hours (150 minutes). The sum of the bar pairs for each system show that roughly 125 to 135 minutes was accounted for by these two applications. Other activities that occurred but are not represented in the Figure include: analysts paused the Glass Box when they took breaks or 5-10 minutes; they used File Manager to save files; and spent short periods of time in miscellaneous applications. The error bars indicate the standard error of the mean of 14 values.

The data show that GNIST required the analysts to spend more time using their browser at the expense of time to write into their Word reports. None of the other systems demonstrated this lop-sided effect; they showed roughly equal amounts of time spent in the application and Word.

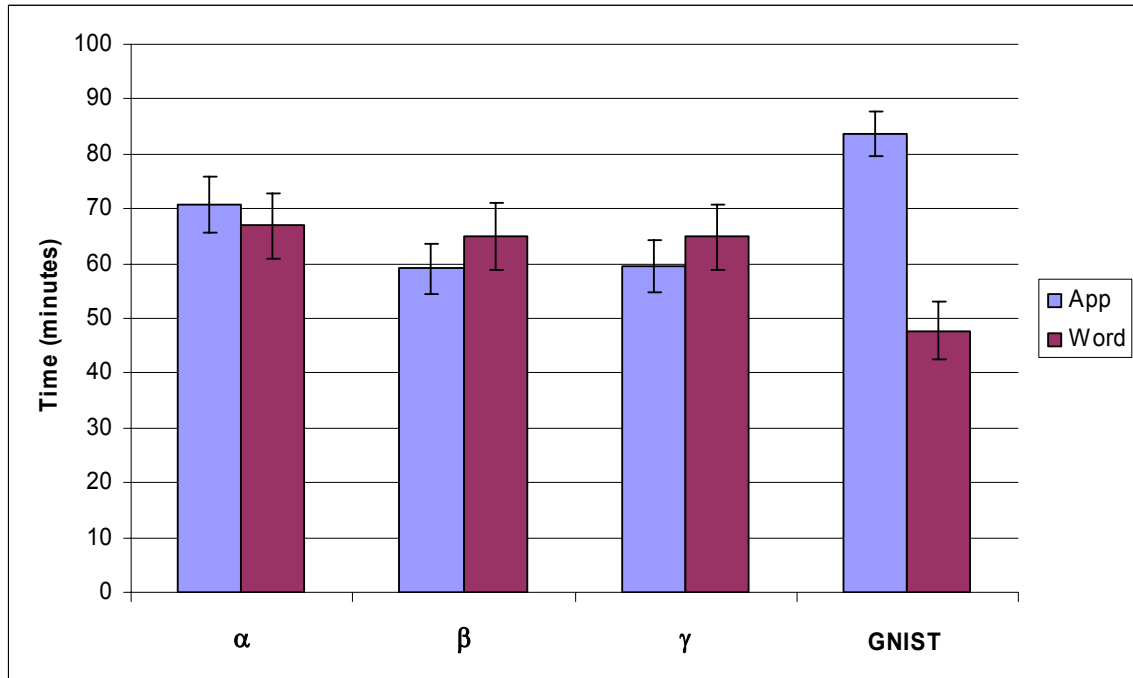


Figure 4: Comparison of Time Spent in Word vs. Test System Application

4.5.2 Activity trails

The Glass Box captures information that allows us to detect patterns of user activity. The following Figures shows a trace of key points for two sessions for one analyst. Cross-evaluation showed that analyst ch7 was the highest performing of the group of analysts, so his data has been chosen for this example. Figure 5 shows one of the GNIST sessions for analyst ch7. The Ψ system has been shown throughout the results to be a top performer, so the comparison in Figure 6 uses the data from one of analyst ch7's sessions on the Ψ system. Glass Box data was used for all the data in Figure 5. In Figure 6, the Glass Box provided the start and end time, as well as the copy events; the Documents viewed and the actual Question/Query markers were determined from the system logs.

The patterns of activity are quite different. In the GNIST session, the analyst started off by doing a query and then he spend over an hour reviewing and copying from documents returned by that single query. Inspection of the log shows that he viewed the first 12 pages of Google results (with 10 document snippets/page). After this the analyst did a few more query – read – copy sequences, followed by a set of non-productive queries.

In the trace from the Ψ system shown in Figure 6, the same analyst exhibited a markedly different pattern. The distribution of Question submitted by the analyst is fairly evenly distributed over time. The number of documents viewed is much smaller. Toward the end of the session, the analyst was still having productive queries, even though there appear to be several sets of non-productive questions within the middle of the session.

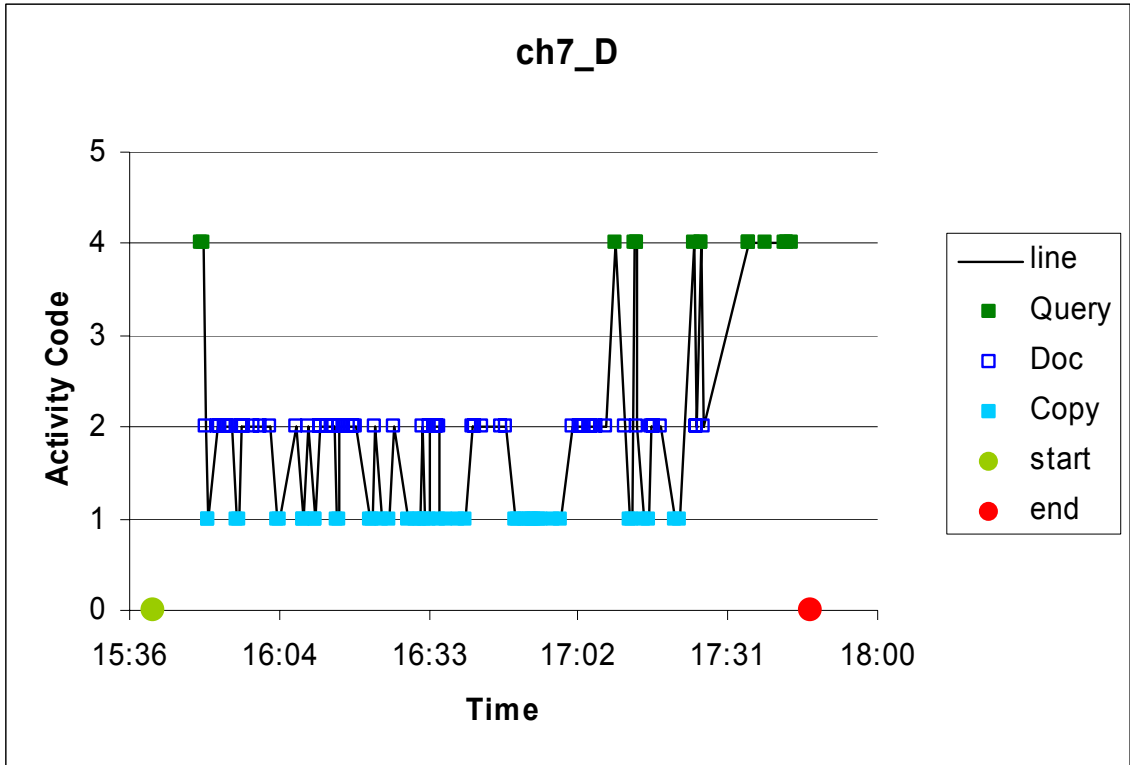


Figure 5: Best performer using GNIST system

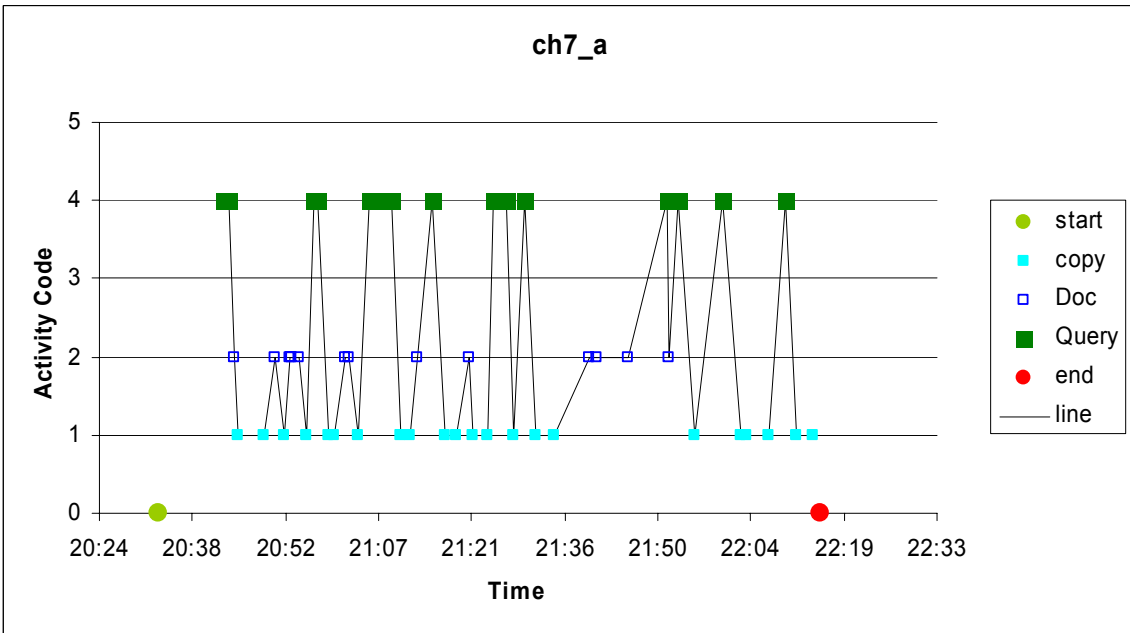


Figure 6: Best performer using Y system

4.6 Query Trails

System logs and Glass Box data allowed us to extract the set of queries done by each analyst for each scenario. These data have been shared with other AQUAINT researchers. We also continue

to use these traces to develop measures to analyze the strategies that analysts use in working on complex, yet tractable, problems. A full list of these query trails is available on the Metrics Challenge web site.

4.7 NASA TLX

The data in Table 14 show the mean values across all 8 scenarios for each of the 7 analysts who worked the full set of tasks (ch6 has been excluded). The overall TLX score shows that the **Y** system was associated with a significantly lower cognitive workload, while all the other systems were not different from one another. Inspection of the individual factors reveals many interesting things. For instance,

- GNIST system was associated with a much higher Temporal effect. Our working explanation for this is that Google returns answers so fast that the onus is always on the analyst to do something. On the other hand, it might be that the information that Google returns to the user is so poor that the analyst starts to worry about time passing.
- The mental component for the **Y** system is much higher than for the other systems even in the face of an overall reduction in workload.

Table 14: NASA TLX Cognitive Workload Measures

	α	β	γ	GNIST	Sig.
Mental	0.66	1.00	1.28	0.52	$\gamma > \beta > \{ \alpha, \text{GNIST} \}$
Physical	0.43	0.17	0.14	0.15	$\alpha > \{ \beta, \gamma, \text{GNIST} \}$
Temporal	0.89	1.17	0.62	1.70	$\text{GNIST} > \beta > \alpha > \gamma$
Performance	0.72	0.47	0.83	0.99	$\{ \text{GNIST}, g \} > \alpha > \beta$
Frustration	1.10	0.93	0.59	0.82	$\{ \alpha, \beta \} > \text{GNIST} > \gamma$
Effort	1.10	1.02	0.75	1.44	$\text{GNIST} > \{ \alpha, \beta \} > \gamma$
TLX Score	4.91	4.75	4.20	5.61	$\{ \text{GNIST}, \alpha, \beta \} > \gamma$

We have adopted the NASA TLX as a routine instrument when evaluating interactive systems. The only recommendation that we would make to others is that the TLX weighting is more effective when it is measured each time the base instrument is administered. In the case of this evaluation, the Weighting was done at the end of a block while the ratings were done at the completion of each scenario. The fact that scenarios are not all equally difficult (see Section 4.9), it is possible that analysts would provide different weights if they were given a chance.

Finding: The NASA TLX is sensitive to differences among systems. It allows not only overall workload estimates but seems to be able to offer pointers into the actual factors that comprise cognitive workload.

4.8 Scenario Assessment

We asked the analysts to rate the scenarios for difficulty as we wanted to see if there were effects of more difficult scenarios on the metrics for the Q&A systems.

4.8.1 Post-Scenario Questionnaires

The results of the Post-Scenario Questionnaires are displayed in Appendix Z. Overall, only one statistically significant difference was detected between scenarios. This difference was in

response to a question that asked analysts to assess the similarity of the scenario to tasks that they typically perform at work. For this question, which yielded a p-value of .041, the post-hoc analysis demonstrated that Scenario B was significantly less similar to their typical tasks, than Scenarios E, F, G and H, and that Scenario A was significantly less similar to their typical tasks than Scenarios G and H. These results should be viewed cautiously, as Scenarios A and B were the first scenarios encountered during the experiment. These ratings may be because of the newness of the experiment, or because of analysts' uncertainties about how to complete the task given the lack of contextual information in the early scenario descriptions.

4.8.2 Scenario Difficulty Assessment

The results of the Scenario Difficulty Assessment rankings for the scenarios used in the analytic sessions are shown in Table 15. Data on the ratings and all data for scenarios not used in this study are not included in this report.

Table 15: Scenario Difficulty Rankings

scenario	ch1	ch2	ch3	ch4	ch5	ch6	ch7	ch8	mean	mean(-ch6)
A	5	3	6	2	3	4	1	5	3.63	3.57
B	4	4	7	8	4	2	2	2	4.13	4.43
C	6	7	2	6	5	7	4	6	5.38	5.14
D	8	8	5	7	2	6	5	7	6.00	6.00
E	7	6	8	3	6	5	6	8	6.13	6.29
F	3	1	1	1	1	8	3	1	2.38	1.57
G	1	2	3	5	8	1	8	4	4.00	4.43
H	2	5	4	5	7	3	7	3	4.50	4.71

The data for analyst ch6 are shaded and an additional mean value is shown in the last column based on only the unshaded data points. Analyst ch6 rated all the scenarios for difficulty but since he arrived late, he actually only analyzed scenarios G and H. On average scenarios D and E were considered hard, while scenario F seems to have been the easiest.

The following Figure plots the two measures of scenario difficulty – post-scenario questionnaire and scenario complexity survey – against one another. There is a modest relationship of the two measures. More research will be needed in order to determine which of these methods is more sensitive to the type of difference that we are interested in measuring about scenarios.

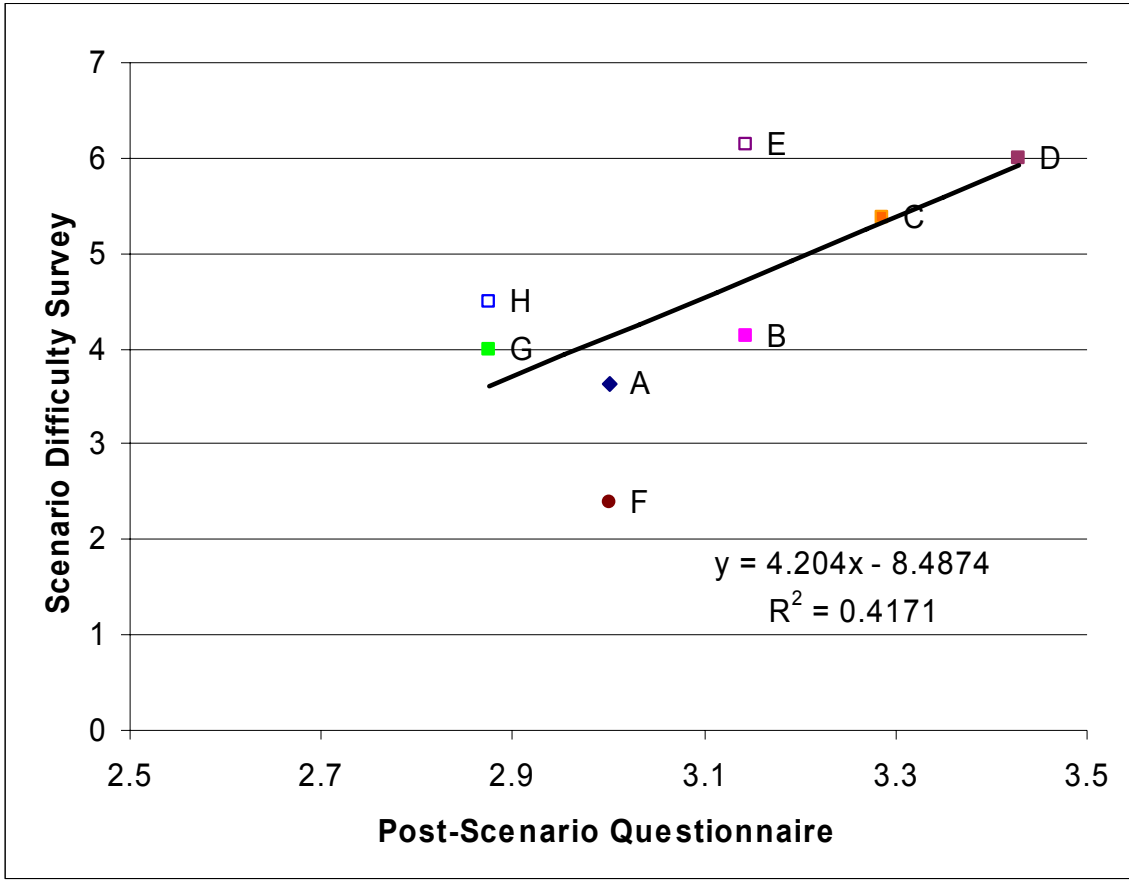


Figure 3: Comparison of two measures of scenario difficulty

5 Summary

The approach of this investigation was based on using hypotheses about question-answering systems to drive the development of appropriate methods and metrics for measuring system effectiveness. The marked-up matrix in Table 16 once again shows the hypotheses that we attempted to address. Previous sections of this report have described how we used each of the methods, shown as headers in the table, to look for differences and similarities across the software systems that were tested. The cells in the table are color-coded to show successes in green and failures in red. Yellow markers show situations where data has not yet been analyzed. It is important to note that 'success' is defined as identifying a method that was able to demonstrate a difference among the participating systems. Failure denotes methods that were unable to detect differences among the systems.

Table 16: Updated Hypothesis-Method Matrix

Question answering systems should	Questionnaires	NASA TLX	SmiFro & Status	Cross-evaluation	System Logs	Glass Box	Query Trails
SES SYS							
H1 Support information gathering with lower cognitive workload		X			X	X	X
H2 Assist in exploring more paths/hypotheses	1/1 3/5						X
H3 Enable production of higher quality reports	2/2 3/3			X			
H4 Provide useful suggestions to the analyst	1/1				X	X	
H5 Provide more good surprises than bad	1/2		X				
H6 Enable more focus on analysis than data collection	0/1						
H7 Enable analysts to collect more data in less time	2/2 1/2					X	
H8 Reduce the time spent reading	0/1					X	
H9 Identify gaps in the knowledge base	0/1				X	X	
H10 Help the analyst recognize gaps in their thinking	4/4						
H11 Provide context for information	1/1				X		
H12 Provide context, continuity and coherence of dialogue	1/1 2/3				X	X	X
H13 Let analysts relocate previously seen materials							
H14 Be easy to use	2/2 2/6	X					
H15 Increase an analyst's confidence in exploration and report	2/2 0/1		X				

While each of the methods has been discussed in detail with respect to its implementation and results, it is important to look at the hypotheses that guided this investigation. First of all, it is clear that some hypotheses were easier to operationalize with a variety of methods. For example, H1 (Support information gathering with lower cognitive workload) and H 12 (Provide context, continuity and coherence of dialogue) were studied with four different methods, all of which were able to discern differences among the systems. As mentioned previously, the project team decided to table H13 since there didn't appear to be any easy way to operationalize it. Although H10 was also tabled during our initial discussions, it turned out that questionnaire items could be constructed to address the relevant issues.

5.1 Key Findings

- Questionnaires are powerful discriminators across the range of hypotheses tested. They are also relatively economical to develop and analyze.
- The NASA TLX is sensitive within a limited domain, i.e., components of cognitive workload. It is cheap to administer and to analyze.
- The SmiFro Console was less than optimal in its implementation. Feedback from the analysts should allow a better re-design. Capturing analysts' 'in the moment' thoughts remains a challenge.
- Formative techniques, such as interviews and focus groups, provide the most useful and timely feedback to developers. Status reports as implemented in the SmiFro Console are another formative method. They were easy to obtain but defy analysis.
- Cross-evaluation of reports was shown to be a sensitive and reliable method. Except for questionnaires, it is the only method used that can measure the quality of the analysts' work products. The method is costly in terms of an analyst's time and analysis requires skill in statistical methods.
- System logs provided answers to several questions that were not addressable with other methods except the Glass Box. Logging is very expensive and rarely reusable. Development of a standard logging format for interactive QA systems would be advisable.
- The Glass Box provided data on user interaction across all systems at some level of granularity and at a fine level for the GNIST system. The cost of collection is low but the cost of analysis is probably prohibitive for most groups. NIST's experience through participation in the NIMD Program enabled the Glass Box to be used for this Workshop. Other commercial tools are available that capture some of the data and should be evaluated for future studies of this sort.
- Query trails were a by-product of this study rather than a full-blown method. This data has proven useful to several of the AQUAINT teams as well as to the participants.

5.2 Lessons learned

In addition to the Findings inserted in the relevant sections above, we have made some observations that we wish to bring to the attention of people who may want to perform a similar study in the future. This is by no means an exhaustive list. The logistics of setting up this experiment went quite smoothly considering the complexity of the undertaking. However, some lessons learned for future evaluations are:

- **Subjects:** Care should be taken to ensure that subjects have a working knowledge of basic tasks and systems, such as using browsers, Microsoft Word, and possibly Microsoft Excel.

- Local Arrangements: Badging at the local site is the first activity that all personnel need to deal with. It is an opportunity to set the tone for the event. Analysts are out of their normal surroundings and may be somewhat intimidated at all the activities going on around them.
- Preparation: Even with having the scenarios, questionnaires, and data sets ready beforehand, a full week is needed onsite to get the materials ready for the experiment.
- Schedule: The first block takes the longest and, once analysts get accustomed to the instruments and protocol, things take less and less time, so the schedule should allow for an acceleration.
- Coordinating the activities of two analysts is difficult especially when they work at very different paces. The best situation is one observer/one room/one analyst.

5.3 Available Resources

The NIST team maintains a password-protected website (<http://control.nist.gov/amc/>) for materials related to this project. To obtain a password, send email to emile.morse@nist.gov. A partial list of available materials includes:

- Document collection
- Various text-only versions of the corpus
- Scenarios
- Analyst reports
- Questionnaires

Many other resources are not posted but are available upon request. This includes such things as list of queries, lists of documents, and list of copied snippets; each is coded by analyst and scenario.