Natural Language Engineering **15** (1): 119–141. © 2008 Cambridge University Press
doi:10.1017/S1351324908004932 Printed in the United Kingdom

Questionnaires for eliciting evaluation data from users of interactive question answering systems

D. KELLY†
University of North Carolina, Chapel Hill, NC 27599-3360, USA
e-mail: dianek@email.unc.edu

P. B. KANTOR
Rutgers University, New Brunswick, NJ 08901, USA
e-mail: kantor@scils.rutgers.edu

E. L. MORSE
National Institute of Standards & Technology, Gaithersburg, MD 20899, USA
e-mail: emile.morse@nist.gov

J. SCHOLTZ‡
Pacific Northwest National Laboratory, Richland, WA 99352, USA
e-mail: jean.scholtz@pnl.gov

Y. SUN
University at Buffalo, The State University of New York, Buffalo, NY 14260, USA
e-mail: sun3@buffalo.edu

(*Received 6 March 2007*)

Abstract

Evaluating interactive question answering (QA) systems with real users can be challenging because traditional evaluation measures based on the relevance of items returned are difficult to employ since relevance judgments can be unstable in multi-user evaluations. The work reported in this paper evaluates, in distinguishing among a set of interactive QA systems, the effectiveness of three questionnaires: a Cognitive Workload Questionnaire (NASA TLX), and Task and System Questionnaires customized to a specific interactive QA application. These Questionnaires were evaluated with four systems, seven analysts, and eight scenarios during a 2-week workshop. Overall, results demonstrate that all three Questionnaires are effective at distinguishing among systems, with the Task Questionnaire being the most sensitive. Results also provide initial support for the validity and reliability of the Questionnaires.

1 Introduction

Question answering (QA) systems have been under development for many years and include systems designed to support retrieval from structured databases (c.f.,

† To whom all correspondences should be addressed.
‡ This work was completed while the author was at the National Institute of Standards & Technology.

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Woods, Kaplan and Nash-Webber, 1972) and those used to support retrieval from unstructured collections of documents (cf., Harabagiu *et al*., 2005). Current development is moving towards QA systems that support a range of question types and interactions with users (cf., Maybury, 2004; Strzalkowski and Harabagiu, 2006). In this paper, we focus on interactive QA systems that have been designed to support human users engaged in complex, analytical information tasks, such as those that are typically performed by intelligence analysts (Maiorano, 2006). By *interactive QA system* we mean a QA system that supports at least one exchange between the user and the system, and that allows the user some control over the content that is displayed and the actions that are taken.

With the design of interactive QA systems comes the challenge of determining how to evaluate such systems in their interactions with end-users. The TREC (Text Retrieval Conference) QA and Interactive Tracks (Voorhees and Harman, 2005) provide ways to evaluate QA systems for fact retrieval and systems that support interactive information retrieval (IR), respectively, but they do not show us how to evaluate interactive QA systems. Evaluation approaches created for dialogue systems such as those described in Walker *et al*. (1997) and den Os and Bloothooft (1998) provide techniques for investigating systems that are driven by dialogue interactions, but they do not show us how to evaluate systems that employ a variety of interactive techniques.

The TREC QA Track evaluation paradigm is system-oriented and evaluates techniques used for answer extraction (Voorhees 2005). No users are involved in these evaluations, and interaction and use are not focal points of the evaluation framework. Historically, this evaluation approach has focused on retrieval for factual questions rather than retrieval for other question types, such as analytical questions (Maybury 2004). Analytical QA differs from factual QA in that answers are often multi-dimensional and require combining information from multiple sources. Analytical QA includes factual QA and other activities, such as comparison and synthesis, and therefore demands much richer interactions among the system, the information and the user. Examples of analytical QA research include Harabagiu *et al*. (2005) and Small *et al*. (2004) (see Strzalkowski and Harabagiu, 2006 for an overview of advances in analytic QA and descriptions of other research efforts).

The TREC QA Track has expanded recently to other question types, such as relationship questions, and by adding an interactive QA task (Dang, Lin and Kelly, 2007). A relationship is said to exist between two entities when there is 'the ability of one entity to influence another,' and relationship questions address both the means of influence and the motivation for doing so. Such questions have been identified as a type of information need often experienced by intelligence analysts. The complex interactive QA task (ciQA) provided an evaluation framework for participants to investigate QA interaction for relationship questions within the confines of a single interaction. In particular, ciQA measured the impact of various *single-cycle* interactions on the performance of QA systems, using evaluation measures used in noninteractive QA research. The goal of the ciQA task was not to suggest holistic evaluation measures for interactive QA systems or measures for evaluating the quality of the interaction or the process of finding information. The focus was

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

instead on the outcome of the interaction; this is obviously important, but it does not tell the whole story. The ciQA evaluation method did not elicit evaluation data from system users and given that users are one of the major players in the interaction, we feel that this is an important limitation. Overall, the method used in ciQA is most suited to formative evaluation of individual aspects of QA interaction, rather than for summative system evaluations.

In the context of IR, the TREC Interactive Track led research and development of evaluation methods for interactive IR systems for 9 years (Hersh and Over, 2001; Dumais and Belkin, 2005). The overall design issues associated with conducting user-centered IR evaluations were presented by Tague-Sutcliffe (1992) (see also Borlund, 2003a). One basic change in viewpoint needed to study interactive systems with real users is to leave the TREC evaluation model which posits that specific items are *relevant*, leading to measures such as precision and recall. The TREC model assumes fixed relevance judgments made by a single (idealized) user (i.e., a TREC assessor). It is not clear that the same judgments would be made by other users, and it is well documented that relevance judgments vary according to user, task, and context (Borlund, 2003b). Another major finding of interactive IR research conducted so far is that results obtained in non-user evaluations do not always translate to user-centered evaluations (Hersh, 2006). For example, two systems might perform (measurably) differently within the context of a TREC QA evaluation, but this difference might not be observed when people use the systems in either laboratory or naturalistic settings. Thus, there is a need to create methods, and metrics supporting multi-user evaluations which are not dependent on objective relevance judgments are needed. In this paper we focus on questionnaires. Although individual researchers have developed a range of questionnaires to supplement their evaluations of interactive IR systems, these questionnaires have varied as much as the systems studied and no major effort has been made to develop a single questionnaire that is general enough to be used across a range of interactive IR systems.

Recent work in the QA community considers how to evaluate QA systems that support complex analytical tasks, with real users (Diekema *et al*., 2004; Liddy, Diekema and Yilmazel, 2004; Kelly *et al*., 2007; Wacholder *et al*., 2007). Diekema *et al.* (2004) evaluated the Knowledge Acquisition and Access System (KAAS), a QA system for undergraduates in aeronautical engineering. They worked in a natural environment, using data from system logs and users' responses to open-ended survey questions. Wacholder *et al.* (2007) describe a series of 3-day evaluations of HITIQA (Small *et al*., 2004), a complex QA system designed to support intelligence analysts' exploratory research for analytical questions. Analysts used HITIQA to prepare reports on various scenarios and participated in a number of evaluation exercises including questionnaires, interviews, and report ratings. Finally, Hersh (2006) describes a series of studies conducted within the medical domain that evaluated users' abilities to answer clinical questions using relatively traditional IR systems. The methods and metrics employed in these studies were similar to those used in interactive IR evaluations and the systems being studied were not designed specifically to support the resolution of complex, analytical questions. While these

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

research efforts offer measurement methods, they do not focus on the development and evaluation of the methods and metrics *per se*. Such a step is needed to validate them so that they can be used by other researchers.

Walker *et al.* (1997) and den Os and Bloothooft (1998) proposed methods for evaluating spoken dialogue systems. These methods were meant to be general, in that no specific task or use situation was described. The performance function proposed by Walker *et al.* (1997) was based on the success and costs of spoken dialogue interactions, as well as a single-user satisfaction measure. It is an important limitation of this method that only a single satisfaction measure was elicited from users. Satisfaction is a complex construct and it is difficult to imagine that any single question can reliably assess this construct in this context. For instance, den Os and Bloothooft (1998) proposed and evaluated a general questionnaire designed to assess the quality of spoken dialogue interactions. den Os and Bloothooft (1998) found that *five* factors contributed to user opinions of the systems: (1) general appreciation; (2) functional capabilities; (3) intelligibility of the output speech; (4) proficiency in the system's language; and (5) familiarity with spoken dialogue systems. While some interactive QA systems of the sort that are the focus of this paper have a dialogue module, not all of them do; few, if any, have a spoken dialogue component. Therefore, dialogue-centered evaluation methods are not entirely appropriate to the situation that is under study in this work. Finally, we note that the purpose of our work is to evaluate the *utility of systems*, rather than the *effectiveness of dialogue episodes*, so the unit of analysis also differs.

2 The problem

The work reported here seeks to fill the gap just mentioned by developing and validating methods and metrics for evaluating the utility of interactive, analytical QA systems. An overview of the study method is provided elsewhere (Kelly *et al.*, 2006). In this paper, we provide a detailed look at the questionnaires used in the evaluation. These questionnaires proved to be effective at distinguishing among systems, and they can be applied easily to other interactive QA evaluations. We focus on three types of questionnaires: a Cognitive Workload Questionnaire (NASA Task Load Index), a Task Questionnaire (completed after each specific task), and a System Questionnaire (completed after using a system). To our knowledge, this is the first research effort to develop and evaluate a set of questionnaires for use in assessing the utility of interactive QA systems designed to support the investigation of complex, analytical tasks.

Development of the instruments followed a set of working hypotheses about the general benefits that a 'good' QA system should provide (Table 1). The hypotheses were developed during planning meetings involving the authors, several system developers, and representatives of the intelligence community, and were informed by research described in Kelly *et al.* (2007). No claim is made about the exhaustiveness and comprehensiveness of these hypotheses. We pose them as starting points for evaluating systems with interactive QA capabilities. As advances occur, this list may grow and change, but we do not expect that this will happen rapidly.

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Table 1. *Dimensions and hypotheses used to guide the development and evaluation of Questionnaires*

Dimension	Hypotheses: 'A good interactive QA system should . . .'
Discovery	[1] Assist in exploring more paths/hypotheses. [2] Provide useful suggestions. [3] Provide more good surprises than bad. [4] Help users identify gaps in the knowledge base. [5] Help users recognize gaps in their thinking. [6] Provide relevant contextual information.
Effort	[7] Support users' information gathering with lower cognitive workload. [8] Be easy to use.
Efficiency	[9] Enable users to focus more on analysis than data collection. [10] Enable users to collect more data in less time. [11] Reduce the amount of time spent reading documents.
Interaction	[12] Provide coherence and continuity with respect to the interaction. [13] Allow users to relocate previously seen materials.
Outcome	[14] Enable users to produce higher quality reports. [15] Increase user's confidence in exploration and quality of report.

The hypotheses relate to five putative dimensions of either the task or the system: discovery, effort, efficiency, interaction, and outcome. After considerable discussion we converged on these hypotheses as general enough to address many of the functionalities of current interactive QA systems. We note that each hypothesis implies a comparison, which requires the testing of at least two systems. This might be a direct comparison of two systems at the same point in time or a comparison of the same experimental system at two points in time.

This study includes three experimental, interactive QA systems and one baseline system. We stress that the goal is *not* to evaluate the systems, but to evaluate the efficacy of our *instruments* in *distinguishing* among these four systems, with respect to the hypotheses. Our subjects were representative of the target users, intelligence analysts who use such systems to meet information needs related to their jobs. The systems, users, and tasks are described in more detail in the next section.

3 Methods and materials

The research was done at the Pacific Northwest National Laboratory in Richland, WA as a U.S. government Department of Defense Advanced Research and Development Activity (ARDA) Challenge Workshop. The work spanned 3 weeks: one week in on-site preparation of tools and systems, and two weeks conducting the evaluation.

3.1 Systems

The systems were three state-of-the-art interactive QA systems and a Google baseline. Each QA system was an example of an end-to-end system developed

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

under the ARDA Advanced Question Answering for Intelligence (AQUAINT) program. The QA system developers allowed us to use their systems to validate our *evaluation methodology*, which was the purpose of our research. Their participation was encouraged by the program manager and individual systems were chosen so that the set of systems exhibited widely divergent approaches to addressing the problem of QA. The Google appliance with its standard search interface was used to create the baseline system.

The three experimental QA systems are not named here as they are not the focus of the research reported in this paper. Although the features and interaction techniques differed across system, the underlying goal of all systems was to help analysts in complex, analytic tasks. All systems required analysts to enter natural language queries to initiate their interactions and all provided answer sets for analysts to review. Aspects that distinguished systems included whether or not they engaged analysts in dialogue, how they presented results to analysts, and how they elicited and used feedback from analysts.

3.2 Subjects

Seven intelligence analysts used each system. The analysts were volunteers serving a yearly 2-week service requirement as U.S. Naval Reservists. They were recruited by email solicitation to a large pool of potential volunteers. The first eight people to respond were accepted into the study, but only seven analysts were present for the entire experiment. The analysts' ages ranged from 30 to 54 years. Their education levels ranged from high school diploma to Ph.D.; over half had Master's degrees. The number of years they had served in the military ranged from 2.5 to 25, and the number of years spent doing intelligence work ranged from 0 to 23.

3.3 Scenarios and tasks

Fourteen scenarios were developed by analysts from the Air Force Rome Research Lab (U.S.) and were later vetted to ensure their appropriateness to the corpus. In total, eight scenarios in the domain of chemical/biological weapons of mass destruction were used. Each scenario included a title and description of an information need, as well as information about the target audience for the report. An example scenario is shown in Figure 1. The assigned 'task' for analysts was to gather information about the scenario using one of the QA systems and to write a report similar to the kind they would write while doing intelligence work for a specific 'customer.' The customer, the role, and a description of what was desired in the report was provided at the bottom of each scenario. We refer to analysts' work products as 'pseudo-reports,' since they were not as comprehensive as the ones they would produce in their normal jobs, given the time and corpus constraints.

The primary corpus was the AQUAINT Center for Non-Proliferation Studies (CNS) corpus. This consists of the January 2004 distribution of the *Eye on Proliferation* CD-ROM collection, which has been disaggregated by CNS into about 40,000 documents. When the initial fourteen scenarios were delivered to the National

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Table 2. *Size of corpus in bytes*

Source	All files	Documents	Images
CNS	40,192	39,932	945
Web	261,590	48,035	188,729

Scenario A: [country's] Chemical Weapons Program

Before an [agency] is reestablished in [country], a current, thorough study of [country] chemical weapons program must be developed. Your task is to produce a report for [person] regarding general information on [country] and the production of chemical weapons. Provide information regarding [country] access to chemical weapons research, their current capabilities to use and deploy chemical weapons, reported stockpiles, potential development for the next few years, any assistance they have received for their chemical weapons program, and the impact that this information will have on the [country]. Please add any other related information to your report.

Customer: [customer]
Role: Country desk – [country]
What they want: General report on [country] and CW production

Fig. 1. Example scenario (information redacted as part of the funder's internal review process).

Institute of Standards and Technology (NIST), they were compared with the CNS corpus. Several scenarios had less than hundred documents in the CNS corpus and were subsequently judged inadequate for the pseudo-report task. The corpus was further supplemented by mining the Web. Documents were collected from the Web by semi-automated querying of Google and manual selection of documents in the result sets. The approximate sizes of the different kinds of files, as determined by their file extensions, are summarized in Table 2.

3.4 Evaluation design

The evaluation workshop consisted of four, 2-day blocks. In each block, two analysts were assigned to one of the four rooms that contained two machines with the same QA system. Analysts worked independently and after each block rotated to different rooms. To control for potential interaction effects among analysts, systems, and scenarios, we used a modified Greco-Latin 4×4 design (Table 3). The goal in using

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Table 3. *Greco-Latin* 4 × 4 *evaluation design (A: analyst)*

Dates	Days 1, 2	Days 3, 4	Days 5, 6	Days 7, 8
Scenarios	A, B	C, D	E, F	G, H
System 1	A1	A2	A3	A4
	A5	A6	A7	A8
System 2	A4	A3	A2	A1
	A7	A8	A5	A6
System 3	A2	A1	A4	A3
	A8	A7	A6	A5
System 4	A3	A4	A1	A2
	A6	A5	A8	A7

this design is to ignore the effect of 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 might be better for certain scenarios); system and analyst (some analysts might adapt more quickly to a system); and analyst and scenario (some analysts might be more expert for certain scenarios). This design ensured that each analyst used each of the four systems and that each system was, for some analyst, the first, second, third, or last to be encountered. This also ensured that no analyst did any scenarios twice and that analyst pairings were unique.

It was noted earlier that analysts varied according to their level of experience in performing intelligence tasks and also along other individual attributes, such as age and education. While the design in Table 3 distributes the potential effects of these attributes across systems by requiring that each analyst uses each system an equal number of times, it does not let us investigate the potential impact of these attributes. This would require a larger number of subjects, the identification of specific attributes of interest, and a stratified sampling technique to ensure that an equal number of subjects representing the specific attributes were studied.

3.5 Data collection

A variety of data collection methods and instruments were used in this study. Five techniques were used to assess experiences and attitudes: Task Questionnaire, System Questionnaire, Cognitive Workload Questionnaire, interviews, and focus groups. System logs, Glass Box (Cowley, Nowell and Scholtz, 2005; Cowley *et al.*, 2006), and naturalistic observation were the core logging methods and provided data about analysts' interactions with systems and their search processes. Cross-evaluation (Sun and Kantor, 2006) was the primary quantitative method for evaluating reports that analysts produced. Cross-evaluation is both an activity and an instrument; it requires analysts to review each other's reports anonymously and rate each report according to seven quality criteria. The Task, System, and Cognitive Workload Questionnaires are the primary focus of this paper and are discussed in greater detail in Sections 3.6

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

and 3.7. Other instruments and methods will be described only briefly. Interested readers are referred to Kelly *et al.* (2006) for more detailed descriptions of these instruments and methods.

System logs were standard logs that recorded analysts' interactions with systems. These logs were developed by the system developers, although some requirements were specified by the research team. Glass Box is a client-side application which captures and archives all workstation activities, and also allows for subsequent retrieval and analysis of this data (Cowley *et al.*, 2006). This tool was developed as part of another project to study intelligence analysts' natural research activities (*Novel Intelligence from Massive Data*). This tool allowed us to obtain a very complete record of analysts' activities, including those related to report creation. In addition to these two types of logging, a trained researcher observed each pair of analysts while they worked.

Experimental blocks were divided into two sessions: each session corresponded to a scenario. Instruments were used throughout the experimental block (e.g., logging), at the end of the session, (e.g., the Task Questionnaire for each scenario), or at the end of the block (e.g., the System Questionnaire). When all of the work of writing and evaluating reports was completed, focus group interviews were conducted to understand more about analysts' overall experiences, and to learn more about the study method. The evaluation design gave analysts an opportunity to fully interact with four QA systems, and to complete time-limited tasks similar to their normal work tasks. From our perspective, this meant that we could evaluate a range of instruments across a panel of different systems that shared a common purpose. In this paper, we focus on the three questionnaires: the Cognitive Workload Questionnaire (NASA TLX), the Task Questionnaire, and the System Questionnaire.

3.6 Cognitive workload questionnaire: NASA TLX

The NASA TLX (Hart and Staveland, 1988) was used to assess analysts' cognitive workloads as they completed scenarios. The NASA TLX was expected to evaluate hypotheses related to effort. Six factors comprise the NASA TLX:

(1) *Mental demand*: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
(2) *Physical demand*: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
(3) *Temporal demand*: How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
(4) *Performance*: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing goals?
(5) *Frustration*: How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

(6) *Effort* : How hard did you have to work (mentally and physically) to accomplish your level of performance?

Each factor is rated across the same scale, but the choice of scale is arbitrary. We used a seven-point scale, where '1' corresponded to 'little' and '7' corresponded to 'much.' Analysts completed these ratings twice for each system, once after each scenario.

After analysts completed two-factor ratings (two for each scenario), they completed the NASA TLX weighting instrument. This elicited rankings of the *relative importance* of the six factors, for each system. Analysts completed all possible pairwise comparisons of the six factors. The number of times that a factor was rated more important by an analyst indicates the relative importance of the factor. Results of these comparisons are converted to a weight for each of the six factors and the individual factor score is multiplied by this value. It is standard practice to elicit weightings for each task, but in this study analysts completed the weighting instrument four times–once for each system, which generated different factor weightings for each system.

3.7 Task and System Questionnaires

The Task and System Questionnaires were created for this evaluation exercise and most of the questions were designed to evaluate interactive QA systems. Analysts completed the Task Questionnaire after each task and the System Questionnaire after they finished using a system. The key difference between these two Questionnaires is that the Task Questionnaire asked analysts about their experiences using a system for the particular scenario just completed, while the System Questionnaire asked analysts to consider their experiences using the system to complete both scenarios.

Questionnaires were developed by a team of researchers including the authors and the QA system researchers; the development of the questions was guided by the hypotheses and dimensions identified earlier, and by questionnaires used in Wacholder *et al.* (2007). Question development was iterative; we started with a much larger list of questions and narrowed and refined the questions over time. Standard usability questionnaires, such as QUIS (Chin, Diehl and Norman, 1988) were not used because such questionnaires are too general for our purposes and do not address QA functionality. In particular, standard usability questionnaires do not address all of the dimensions we were interested in measuring.

Table 4 shows how many questions assessed each dimension. Many questions assessed the discovery dimension as the Questionnaires were the primary tool for investigating discriminability among systems with regard to this aspect. Nine questions were used in conjunction with the NASA TLX to assess effort. Six questions assessed outcome and were used in conjunction with cross-evaluation. The dimensions of efficiency and interaction were not heavily addressed by the Questionnaires, but data about these dimensions were supplemented with log data. Furthermore, it is, in general, difficult to develop questions that assess these dimensions for comparisons among systems because the *types of interactions* supported by systems differed. However, individual system developers added questions that were customized to

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Table 4. *Number of questions from questionnaire associated with each dimension*

Dimension	Task Questionnaire	System Questionnaire	Total
Discovery	7	9	16
Effort	3	6	9
Efficiency	2	4	6
Interaction	1	4	5
Outcome	2	4	6
System readiness	–	3	3

Table 5. *Means (standard deviations) of unweighted scores for each factor and ANOVA results (BL: baseline)*

Factor	System				ANOVA results		
	A	B	C	BL	$F(3,57)$	p-value	Diff.
Mental	4.75 (1.48)	4.71 (1.20)	5.31 (1.01)	**5.75 (1.06)**	2.25	.096	–
Physical	4.19 (1.87)	**4.43 (1.56)**	2.75 (1.29)	3.92 (1.62)	3.36	.025	B >C
Temporal	4.69 (1.45)	4.71 (1.44)	3.56 (1.37)	**5.83 (0.72)**	7.00	.000	BL > C
Performance	4.56 (1.09)	4.50 (1.45)	3.94 (1.65)	**5.83 (0.84)**	4.85	.005	BL > C
Frustration	4.13 (2.19)	4.57 (1.56)	3.44 (1.50)	**4.83 (1.40)**	1.83	.153	–
Effort	4.63 (1.46)	4.93 (1.21)	3.88 (1.46)	**5.83 (0.72)**	5.52	.002	BL > C

Highest scores are in bold.

unique features of their systems (these questions are not shown here). Finally, three questions assessed an issue unrelated to the comparative hypotheses: the readiness of the system for an operational environment.

The Questionnaires are presented in Section 4 of this paper (Figures 2 and 3). All questions were assessed with five-point Likert-type scales, where '1' was labeled 'strongly agree' and '5' was labeled 'strongly disagree.' For a few questions in the Task Questionnaire, labels were changed as appropriate to the question. For those questions, the scale labels are shown in Figure 2 in the Results section. For most questions, higher scores are better, but for some, the direction is reversed and smaller numbers are better (e.g., Q2 and Q5 from the Task Questionnaire). These questions are marked by the word 'reverse' in the tables. A five-point scale was selected since this is the number of points used by Likert (1932) who initially proposed such scales.

4 Results

4.1 NASA TLX questionnaire

Table 5 displays the unweighted mean ratings analysts assigned to the different systems for each factor. The highest mean rating for each factor is in bold. This table also displays results of the Analysis of Variance (ANOVA) tests (including F-statistics, degrees of freedom, and p-values), and the results of Scheffe's

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Table 6. *Means (standard deviations) of weights assigned to each factor*

	System				
Factor	**A**	**B**	**C**	**BL**	**Total**
Mental	2.00 (1.47)	3.00 (1.37)	**3.63 (1.15)**	1.40 (0.84)	2.64 (1.48)
Physical	**1.14 (1.41)**	0.50 (0.52)	0.63 (1.15)	0.40 (0.84)	0.68 (1.05)
Temporal	2.86 (1.03)	3.75 (1.13)	2.63 (1.54)	**4.40 (0.84)**	3.32 (1.35)
Performance	2.43 (0.76)	1.63 (1.15)	**3.13 (1.41)**	2.60 (1.08)	2.43 (1.25)
Frustration	**3.14 (2.18)**	2.75 (1.61)	2.13 (1.75)	2.40 (1.43)	2.61 (1.78)
Effort	3.43 (1.34)	3.38 (0.89)	2.88 (1.50)	**3.80 (0.42)**	3.32 (1.18)

Highest scores are in bold.

pairwise-comparison tests ('Diff.' column). In cases where no statistically significant differences were found dashes appear in the last column.

The ANOVA and Sheffe's tests are used in subsequent analyses reported in this paper and will be explained briefly. ANOVA is an inferential statistical test that determines whether differences among means are statistically significant (a fixed effects model is used). The p-value is the probability that the observed difference between means would happen by chance if the scores for the two systems were in fact drawn from the same distribution. For instance, a p-value of .01 means that there is a 1/100 chance that a difference of the observed size occurred by chance. It is standard practice in the social sciences to accept a p-value of .05 or less; this value will be used in analyses reported in this paper to identify statistically significant differences.

The Scheffe's test is a post-hoc test that is used to control the effect of multiple hypotheses, and is used if significant differences are found by the ANOVA. ANOVA can demonstrate that at least one statistically significant difference exists between means, but does not indicate between which pairs of means this occurred. In this case, there are four means (one for each system) and the Scheffe's test indicates the pair(s) of means between which the difference was significant. For instance, for the physical factor, the Scheffe's test showed a statistically significant difference between Systems B and C, but not among any other pair of systems. The notation in the Diff. column indicates the direction of this relationship; for instance, the mean score for System B was significantly *greater* than the mean score for System C.

For all but one factor, the baseline system received the highest score (indicating more effort) and System C received the lowest scores (indicating less effort). Statistically significant differences between means were detected among systems for four factors: physical, temporal, performance, and effort. In each case, the Scheffe's tests identified only one significant pair and System C's scores were always significantly lower. Overall, it appears that at least for two of the systems the unweighted TLX scores are sufficient at discriminating the top-performing system from the bottom-performing system.

Results of the NASA TLX weighting instrument should provide some indication of the relative importance of each factor, overall, and for each system. Table 6

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Table 7. *Means for each NASA TLX factor and ANOVA results (BL: baseline)*

	System				ANOVA results		
Factor	**A**	**B**	**C**	**BL**	***F*(3,55)**	***p*-value**	**Diff.**
Mental	0.66 (0.56)	1.00 (0.60)	**1.28 (0.42)**	0.52 (0.32)	6.16	.001	C > A, BL
Physical	**0.43 (0.60)**	0.17 (0.18)	0.14 (0.26)	0.15 (0.31)	2.05	.118	–
Temporal	0.89 (0.39)	1.17 (0.50)	**0.62 (0.43)**	1.70 (0.39)	13.68	.000	BL > A, B, C; B > C
Performance	0.72 (0.30)	0.47 (0.30)	0.83 (0.50)	**0.99 (0.36)**	4.35	.008	BL > B
Frustration	**1.10 (0.84)**	0.93 (0.87)	0.59 (0.54)	0.82 (0.57)	1.31	.281	–
Effort	1.10 (0.46)	1.02 (0.37)	0.75 (0.53)	**1.44 (0.22)**	5.60	.002	BL > C
TLX score	4.91 (1.05)	4.75 (1.26)	4.20 (1.00)	**5.61 (0.64)**	3.77	.016	BL > C

Highest scores are in bold.

displays the mean weights that analysts assigned to each factor overall and for each system. Data presented in bold indicate the highest mean weights. Overall, the temporal and effort dimensions were associated with the highest total weight (3.32) and the physical dimension was associated with the lowest total weight (0.68), which is reasonable for reading, thinking, and typing. It appears that the amount of time and effort expended while using the system were the most important factors contributing to analysts' cognitive workloads. When factor weights for individual systems are examined, it becomes apparent that the factors which were selected the most often varied. It is important to note that these weights were generated with respect to each system and comparing weights across systems is not entirely appropriate since the values are relative to each system.

Table 7 displays the mean weighted scores for each factor for each system, which is a combination of the data from Tables 5 and 6. Statistically significant differences were found for four of the six factors: mental, temporal, performance, and effort. These results differ slightly from the results for the unweighted scores (Table 5) in that a statistically significant difference was found for the mental factor, but not for the physical factor. It is also the case that these results are stronger than those reported in Table 5, as evidenced by the larger F-statistics. The Scheffe's tests also detected significant differences among more than one pair of means. For the mental factor, System C's mean was significantly higher than System A's and BL's. For the temporal factor, BL's mean was significantly higher than System A's and System B's, and System B's was higher than System C's. Thus, the weighting had some impact on the resulting factor scores.

Overall the final TLX score (last row in Table 7) shows that System C was associated with a significantly lower cognitive workload, while the cognitive workload associated with other systems was similar. Although System C was associated with the least cognitive effort, it was also associated with the highest weighted mental score, even though it was not associated with the highest unweighted mental score (Table 5). Since the scores in Table 7 are a function of those presented in Tables 5 and 6, this discrepancy may be an artifact produced by assessing individual weights for each system rather than for the task as a whole. For instance, it may have been the case that the effort associated with System C for all factors was relatively low

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

compared to other systems, but the mental factor may have demanded the most effort from analysts even if this demand was lower than that for other systems. Alternatively, it may be the case that a high mental factor is desirable in this situation since the analyst's work is primarily mental in nature.

4.2 Task Questionnaire

Results for the Task Questionnaire are shown in Figure 2. Questions are grouped by associated hypotheses. Figure 2 displays the means and standard deviations of the ratings analysts assigned to the systems for each question, the results of the ANOVA tests (including F-statistics, degrees of freedom, and p-values), and the results of Scheffe's pairwise-comparison tests ('Diff.' column). In cases where no statistically significant differences were found dashes appear in the last three columns. Systems are referred to according how they ranked on a particular question, rather than according to a particular identification code. This is not necessarily fixed, as systems were rated differently for different questions. Thus, the system whose mean score is 4.00 for the first question (Q12) may not be the same system whose score is 2.93 for the second question (Q10). Scores for the baseline system are in bold.

Overall, these results show that most questions on the Task Questionnaires are able to distinguish among at least two groups of systems. Statistically significant differences were revealed by fourteen of the fifteen questions, and in all but two cases, the p-values were less than .01. In most cases, the questions divided systems into two groups. Sometimes one group was a singleton—that is, the score for the top-ranked system for any question was significantly better than the scores for systems ranked 2, 3, and 4. In four cases, the score for the top-ranked system was only significantly better than the scores for third and fourth ranked systems and in one case the score for the top-ranked system was only significantly better than the score for the fourth ranked system.

Another interesting finding is that there was much less variance in the scores that analysts assigned to whichever system ranked first, and to the baseline system than to the systems ranking in the remaining spots. The standard deviations for these systems usually did not exceed one point, while for the other systems the standard deviations often exceeded one. These results suggest that most analysts agreed on the scores that they assigned to whichever system was ranked highest on a question, and to the baseline system. The baseline system was ranked fourth by most questions, and in most cases its score was significantly different from whichever system ranked highest.

4.3 System Questionnaire

Results for the System Questionnaire are displayed in Figure 3. The information contained in this figure is organized similarly to that of Figure 2. For fourteen of the twenty-seven questions that were associated with hypotheses, statistically significant differences among one or more systems were detected. In addition, statistically significant differences among systems for two of the three questions associated with the readiness of the system for deployment in an operational environment were found.

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

DISCOVERY H1: QA system should assist analysts 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 (0.68)	3.07 (1.28)	2.79 (1.25)	**2.47 (0.64)**	7.62	.000	1 > 2, 3, 4
DISCOVERY H4: QA systems should help analysts identify 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. [reverse]	2.93 (0.92)	3.36 (1.01)	3.47 (0.99)	**3.60 (0.99)**	–	–	–
DISCOVERY H5: QA systems should help analysts 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 (0.80)	3.13 (1.06)	2.79 (0.89)	**2.53 (0.83)**	5.54	.002	1 > 2, 3, 4
Q7: The [X] system helped me to think about the scenario in new ways.	3.64 (0.93)	3.13 (0.99)	2.79 (0.97)	**2.53 (0.83)**	3.96	.013	1 > 3, 4
Q11: The [X] system stimulated my thinking about the scenario.	3.79 (0.70)	3.20 (0.77)	3.00 (1.24)	**2.47 (0.83)**	5.86	.002	1 > 3, 4
Q15: The [X] system expanded my understanding of the scenario.	3.71 (0.83)	3.13 (0.83)	3.07 (1.21)	**2.60 (0.51)**	4.28	.009	1 > 4
EFFORT H8: QA systems should be easy to 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] [reverse]	2.00 (0.96)	2.73 (0.88)	**3.53 (0.64)**	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 (0.66)	3.13 (1.30)	2.79 (1.37)	**2.53 (0.92)**	6.44	.001	1 > 2, 3, 4
EFFICIENCY H10: QA systems should enable analysts 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] [reverse]	2.21 (0.97)	3.43 (1.28)	**3.73 (0.70)**	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 (0.92)	3.00 (1.20)	2.50 (1.29)	**2.27 (0.70)**	10.827	.000	1 > 2, 3, 4
INTERACTION H12: QA systems should provide coherence and continuity with respect to the interaction.							
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 (0.78)	3.33 (1.11)	2.86 (1.41)	**2.73 (0.96)**	4.82	.005	1 > 3, 4
OUTCOME H14: 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 (0.66)	3.13 (1.13)	**2.87 (0.64)**	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 (0.53)	3.13 (1.06)	3.00 (1.24)	**2.67 (1.18)**	6.17	.001	1 > 2, 3, 4
OUTCOME H15: QA systems should increase analyst's confidence in exploration and quality of report.							
RANK	**1**	**2**	**3**	**4**	*F*(3,58)	*p*-value	Diff.
Q1: How confident were you of your ability to make the [X] system work to accomplish the assigned tasks? [unconfident … confident]	4.43 (0.65)	**3.80 (1.08)**	3.47 (1.19)	2.86 (1.35)	5.00	.004	1 > 3, 4
Q9: I felt that the system showed me all of the available relevant information about this scenario.	3.57 (0.85)	2.87 (1.13)	2.57 (1.02)	**2.20 (0.68)**	7.15	.000	1 > 2, 3, 4

Fig. 2. Means, standard deviations, results of ANOVA tests (including F-statistics, degrees of freedom, and p-values) and Scheffe's pairwise comparisons (Diff. column) for the Task Questionnaire. For each question, scores are ordered from best to worst, not according to system. Higher scores are better except where the word 'reverse' appears. Scores for the baseline system are in bold. All scales contain five points with the labels 'strongly agree' and 'strongly disagree' unless otherwise noted.

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

DISCOVERY H1: QA system should assist analysts in exploring more paths/hypotheses.							
RANK	**1**	**2**	**3**	**4**	$F(3,27)$	p-value	Diff.
Q1: The system allowed me to easily change my line of questioning.	4.00 (0.58)	**3.43 (0.98)**	2.71 (1.11)	2.29 (1.25)	3.51	.042	1 > 3, 4
Q10: The system helped me think of new ways to search for information.	3.86 (0.69)	3.14 (0.90)	3.14 (1.07)	**2.29 (0.49)**	3.90	.030	1 > 4
Q15: The system allowed me to easily change my search strategy.	3.57 (0.96)	3.29 (1.38)	3.14 (1.21)	**3.14 (1.21)**	–	–	–
Q22: The system often presented me with novel information.	3.49 (0.79)	3.00 (0.82)	2.86 (0.69)	**2.29 (0.49)**	3.11	.058	1 > 4
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 (0.53)	–	–	–
DISCOVERY H2: QA systems should provide analysts with useful suggestions.							
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 (0.98)	3.29 (1.50)	2.57 (0.53)	**1.86 (0.69)**	3.68	.036	1 > 4 2 > 4
DISCOVERY H3: QA systems should provide more good surprises than bad.							
RANK	**1**	**2**	**3**	**4**	$F(3,27)$	p-value	Diff.
Q7: The system often surprised me by doing good things.	3.14 (0.69)	3.00 (0.58)	2.71 (1.25)	**2.57 (0.53)**	–	–	–
Q24: The system often surprised me by doing bad things. [reverse]	1.86 (0.69)	**2.71 (0.76)**	3.00 (0.58)	3.29 (1.50)	3.45	.044	1 < 3, 4
DISCOVERY H6: QA systems should provide relevant contextual 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 (0.98)	**3.00 (0.82)**	3.00 (1.00)	3.57	.040	1 > 3,4
EFFORT H8: QA systems should be easy to use.							
RANK	**1**	**2**	**3**	**4**	$F(3,27)$	p-value	Diff.
Q2: It was difficult to get the system to do what I wanted it to do. [reverse]	2.14 (0.69)	2.86 (0.90)	**3.14 (0.90)**	3.29 (1.11)	–	–	–
Q14: My skill at using the system improved over the course of the workshop.	4.29 (0.49)	4.00 (0.58)	4.00 (1.00)	**3.57 (1.27)**	–	–	–
Q19: The system is hard to use. [reverse]	1.29 (0.49)	**2.14 (0.90)**	2.43 (1.27)	3.00 (0.82)	4.44	.020	1 > 3,4
Q21: I feel that I have become pretty proficient at using the system.	4.29 (0.76)	**3.71 (1.11)**	3.43 (1.27)	3.00 (0.82)	–	–	–
Q28: The system was inflexible. [reverse]	2.00 (1.00)	2.86 (1.25)	**3.00 (0.82)**	3.57 (1.27)	4.31	.022	1> 3,4
Q33: In general, I like using the system.	4.43 (0.53)	3.14 (1.35)	3.14 (1.46)	**3.00 (0.82)**	–	–	–
EFFICIENCY H9: QA systems should enable analysts to focus more on analysis than data collection.							
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. [reverse]	3.14 (0.69)	3.43 (0.79)	3.43 (1.51)	**4.14 (0.69)**	–	–	–
EFFICIENCY H10: QA systems should allow analysts to collect more data in less time.							
RANK	**1**	**2**	**3**	**4**	$F(3,27)$	p-value	Diff.
Q6: The system slows down my process of finding information. [reverse]	2.43 (1.27)	3.00 (1.63)	**3.14 (0.69)**	3.42 (1.40)	–	–	–
Q23: Having the system at work would help me find information faster than I can currently find it.	4.00 (0.82)	2.57 (0.96)	**2.57 (0.96)**	2.43 (1.13)	5.95	.007	1 > 2, 3, 4
EFFICIENCY H11: QA systems should reduce the amount of time analysts spend reading.							
RANK	**1**	**2**	**3**	**4**	$F(3,27)$	p-value	Diff.
Q3: I spent more time reading documents than I did searching for information. [reverse]	3.43 (1.40)	3.86 (0.38)	**4.14 (0.90)**	4.71 (0.49)	–	–	–
INTERACTION H12: QA systems should provide coherence and continuity with respect to the interaction.							
RANK	**1**	**2**	**3**	**4**	$F(3,27)$	p-value	Diff.
Q8: The system allowed me to navigate easily between searching activities.	4.57 (0.53)	**3.57 (0.98)**	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 (0.69)	3.57 (1.13)	2.86 (1.21)	**2.29 (0.76)**	–	–	–
Q25: The system provided coherent answers.	4.00 (0.82)	3.00 (0.58)	3.00 (1.00)	**2.86 (0.69)**	4.23	.024	1 > 2, 3,4
INTERACTION H13: QA systems should allow analysts to relocate previously seen information.							
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 (0.79)	3.57 (1.13)	3.57 (1.13)	**3.00 (0.58)**	–	–	–
OUTCOME H14: 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 (0.49)	3.43 (0.79)	**2.86 (0.69)**	2.86 (1.25)	5.78	.008	1 > 3,4
Q18: I could not find enough documents with relevant information. [reverse]	2.00 (0.58)	2.86 (1.07)	**3.43 (0.98)**	3.71 (1.50)	3.57	.040	1 < 3,4
Q32: The system often presented me with redundant information. [reverse]	2.71 (1.11)	3.57 (1.40)	3.71 (1.11)	**4.14 (0.90)**	3.91	.030	1 < 4 2 < 4
OUTCOME H15: QA systems should increase analyst's confidence in exploration and quality of report.							
RANK	**1**	**2**	**3**	**4**	$F(3,27)$	p-value	Diff.
Q4: I often wanted to verify the system's answers. [reverse]	2.43 (0.96)	**3.00 (0.82)**	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. [reverse]	2.14 (1.07)	**2.71 (0.95)**	3.43 (1.13)	4.14 (0.90)	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 (0.53)	**3.29 (0.49)**	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 (0.76)	2.14 (1.07)	**1.86 (0.90)**	–	–	–

Fig. 3. Means, standard deviations, results of ANOVA tests (including F-statistics, degrees of freedom, and p-values) and Scheffe's pairwise comparisons (Diff. column) for the System Questionnaire. For each question, scores are ordered from best to worst, not according to system. Higher scores are better except where the word 'reverse' appears. Scores for the baseline system are in bold. All scales contain five points with the labels 'strongly agree' and 'strongly disagree' unless otherwise noted.

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Fewer statistically significant differences were elicited by the System Questionnaire than by the Task Questionnaire. In general, the values of the F-statistics were lower than those for the Task Questionnaire, which indicates weaker relationships. The results of the Scheffe's tests were also less discriminating. For most statistically significant relationships, the Scheffe's tests identified two groups. However, in only three cases did the score associated with the top-ranking system significantly differ from the scores associated with the second, third, and fourth ranking systems. In nine cases, the top-ranking system differed from the third and fourth ranked systems and in four cases it only differed from the fourth ranked system. In other words, this instrument seemed to group systems at the high end of the scale, while the Task Questionnaire seemed to distinguish one system from the rest.

Responses to the System Questionnaire also differed from the Task Questionnaire with respect to the range of scores analysts assigned to systems. Scores analysts assigned to systems were more similar to one another in the System Questionnaire than in the Task Questionnaire. Moreover, the position of the baseline system was more variable. While it still ranked fourth most often, its rank varied slightly more on questions in the System Questionnaire, than it did in the Task Questionnaire. However, as with the Task Questionnaire, most questions for which the baseline system did not rank fourth were related to analysts' comfort using the system and may be explained by their familiarity with the baseline system.

4.4 Combining evidence from questionnaires

By combining results of the NASA TLX, the Task Questionnaire and the System Questionnaire, we see how much evidence each instrument provides in support of particular hypotheses. Table 8 presents the original hypotheses, along with specific instruments and questions that assessed each hypothesis. Items with an * indicate those for which a statistically significant difference was found between at least one pair of systems. The row total indicates the number of statistically significant differences for each hypotheses and the column total indicates the number of statistically significant differences revealed by each instrument.

Overall, the three questionnaires elicited the most evidence for hypotheses related to the discovery and outcome dimensions: 75% of the discovery questions and 83% of the outcome questions demonstrated significant differences among systems. Although fewer than half of the Task and System questions for effort resulted in significant differences among systems, the additional evidence provided by the NASA TLX illustrates the effectiveness of combining these three questionnaires to examine this dimension. The least evidence was provided for the questions related to efficiency.

5 Discussion

The NASA TLX was sensitive in evaluating analysts' overall workloads for each scenario. It was inexpensive to administer and analyze and has established validity and reliability as an instrument in a different context, where there are real-time pressures to control mechanical systems (Hart and Staveland, 1988). However, the

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Table 8. *Summary of the degree to which the questionnaires were able to measure differences postulated in the hypotheses*

	Hypothesis	NASA TLX	Task	System	Total
Discovery	[1] Assist analysts in exploring more paths/hypotheses.		4*, 8*	1*, 10*, 15, 22*, 27	5
	[2] Provide analysts with useful suggestions.			11*	1
	[3] Provide analysts with more good surprises than bad.			7, 24*	1
	[4] Help analysts identify gaps in the knowledge base.		10		0
	[5] Help analysts recognize gaps in their thinking.		3*, 7*, 11*, 15*		4
	[6] Provide analysts with relevant contextual information.			5*	1
				Total	12/16
Effort	[7] Support analysts' information gathering with lower cognitive workload.	√*	12*		2
	[8] Be easy to use.	√*	5*, 14*	2, 14, 19*, 21, 28*, 33	5
				Total	7/11
Efficiency	[9] Enable analysts to focus more on analysis than data collection.			13	0
	[10] Enable analysts to collect more data in less time.		2*, 13*	6, 23*	3
	[11] Reduce the amount of time analysts spend reading documents.			3	0
				Total	3/6
Interaction	[12] Provide coherence and continuity with respect to interaction.		6*	8*, 20, 25*	3
	[13] Allow analysts to relocate previously seen materials.			29	0
				Total	3/5
Outcome	[14] Enable analysts to produce higher quality reports.			9*, 18*, 32*	3
	[15] Increase an analyst's confidence in exploration and quality of report.		1*, 9*	4	2
				Total	5/6
	System readiness			12*, 16*, 17	2/3
	Total	2	14/15	16/30	32/47

Entries in each cell correspond to question numbers on the different questionnaires. Items with an * indicate that statistically significant relationships were found among systems by these items. Check marks appear in the NASA TLX column since this entire instrument measured effort.

way in which the NASA TLX weighting was done in this study may prevent comparisons of individual factor scores for each system. One interesting result was that certain system scores for individual factors appeared inflated. For instance, System

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

C required significantly less effort than the baseline according to the final composite TLX score, but actually had a significantly higher weighted score for the mental factor than the baseline and System A (Table 7). This does not necessarily mean that System C required more mental effort than the baseline system or System A, but that during the pairwise comparisons of factors for System C, the mental factor was selected more times than the other factors. This approach to weighting masks the fact that System C could have required less effort across all six factors than the other systems because weights were derived for each system, rather than for the task. The weights are relative to an individual system, not the task for which the system is being used. The weighting process in previous applications of the TLX involved eliciting weights for each different task; indeed the NASA TLX is a measure of the *task* load. However, in this study weights were elicited for each system, which makes cross-system comparison of the weighted factor scores tenuous. Although there was only a single task type in this study and eliciting weights once might have provided a more stable weighting function across systems, the various topics of the tasks created a situation where the cognitive demand may have differed slightly for each task instantiation, which is why the TLX weightings were taken at the system level.

Overall 93% of the questions on the Task Questionnaire distinguished among systems, but only about 50% of the questions on the System Questionnaire did. It appears that the Task Questionnaire is more sensitive than the System Questionnaire and there are several possible explanations for these results. The Task Questionnaire was completed after analysts finished a scenario, while the System Questionnaire was completed after two scenarios (interrupted by a lunch break, and several other activities). The time lapse may have caused the strength of analysts' feelings or memories to decay. Analysts may have regarded some questions on the System Questionnaire as redundant to those on the Task Questionnaire and therefore felt that they had already told us this information. It may also be the case that a courtesy effect occurred when analysts completed the System Questionnaire: analysts may have felt uncomfortable evaluating the systems critically because they did not want to offend the researchers. The Task Questionnaire focused on the system and the task, rather than just the system, and analysts may have been more critical (and more truthful) in their responses to this questionnaire.

After combining evidence from all three questionnaires, it was shown that the questionnaires elicited the most evidence for the *discovery* and *outcome* dimensions of the hypotheses and the least evidence for the efficiency dimension. Since the systems were novel to analysts, it may be that benefits related to efficiency cannot be realized without more use. It may also be the case that our questions were not appropriate or sensitive enough. These questions attempted to elicit analysts' evaluations of different components of the search process, but it may not be appropriate to subdivide analysts' activities in this way. It is possible that this subdivision made it more difficult to identify what it means to be efficient.

Although the same system did not always score highest on every item, we found that when two or more instruments assessed the same hypothesis, the same system scored highest. For instance, for Hypothesis 1 the same system was rated the highest by Task questions four and eight, and by System questions one, ten, and twenty-two. This result is important since it provides some evidence that the instruments

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

are reliable and that they have some construct validity. That is, the various ways we operationalized the different constructs (e.g., effort) seemed to converge. These measures also appear to be correlated with one another. Since the validity and reliability of the NASA TLX as an instrument for measuring effort has been established, this provides some evidence of the concurrent validity of the new measures of effort in the Task and System Questionnaires.

There are several limitations of this study. Although the validity and reliability of the NASA TLX has been established in numerous studies, this was only the first study where the Task and System Questionnaires were used. While our results provide preliminary evidence regarding the validity and reliability of the Questionnaires with respect to distinguishing among interactive QA systems, these aspects will be strengthened as the Questionnaires are used in more interactive QA evaluations. The study also had a relatively low number of participants. This was necessitated by the study design, which allowed us to study the same people's use of a number of systems to complete a number of scenarios. In this study, we chose to have more intense interactions with participants so that we could critically evaluate our instruments with a number of systems and scenarios. Although we found very strong effects for a number of measures, studies with larger numbers of participants and different interactive QA systems are needed to further validate these questionnaires.

6 Conclusions

There have been a number of advances in the evaluation of QA systems, interactive IR systems, and dialogue systems but little work has investigated evaluation methods and metrics for interactive QA systems. Previous work has focused on measuring usability and relevance, as opposed to the utility measures posed in this paper. One challenge to evaluating interactive QA systems using traditional evaluation measures is that relevance is unstable in multi-user evaluations. The problem with usability measures is that they are too general to assess specific aspects of interactive QA systems. The work reported in this paper evaluates alternative techniques for assessing the utility of interactive, analytical QA systems. Three questionnaires were evaluated with four systems, seven analysts, and eight scenarios. The major finding of this study is that the questionnaires were powerful discriminators across a range of hypotheses. While considerable time was spent developing the questionnaires, they were relatively economical to administer and analyze, and most analysts were comfortable completing them.

Results showed that the questionnaires used in this study – the NASA TLX, the Task Questionnaire and the System Questionnaire – were able to discriminate between systems with respect to several hypotheses related to discovery, effort, efficiency, interaction, and outcome. These Questionnaires were developed specifically for interactive QA systems that support a particular kind of task. These results show that it is feasible to use task-based questionnaires to find significant differences among interactive QA systems. However, the questionnaires might not extend well to a comparison of systems used for different types of tasks, for example answering procurement questions or planning recreational travel, as they were not developed

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

for these types of systems and tasks. These questionnaires are most appropriate for evaluations that involve interactive QA systems that support analytical tasks.

A natural question is whether the questionnaires could be reduced and streamlined. Standard techniques used in psychometrics might identify some independent factors underlying the set of responses, and one might experiment with keeping only those questions which load most heavily on these factors. This would reduce the number of questions on the questionnaires, the amount of time spent by participants completing them and the amount of time spent by researchers analyzing them. This might, in particular, be useful for a customer contemplating a major investment in systems of the type studied here.

Each evaluation situation is somewhat different and one hesitates to strive for a one-size-fits-all evaluation strategy for interactive systems. Our questionnaires addressed general characteristics that most interactive QA systems should have, but they did not necessarily allow one to evaluate system-specific features. For this, researchers added their own questions customized to these features. However, there are many opportunities for developing shared frameworks and infrastructures for evaluation. One common approach is to share scenarios and corpora, since they are generally expensive to construct from scratch. The creation of the questionnaires presented here, which are customizable to individual systems, contributes to an infrastructure for interactive QA evaluation. We hope that they will be reused, refined, and further validated. As interactive QA systems advance, these instruments may need to be updated or changed. In fact, changes in interactive QA functionality might lead to changes in the very hypotheses about utility that were used to generate these questionnaires. To facilitate further adoption and use, NIST maintains a password-protected website for materials related to this project. Those who are interested in obtaining the questionnaires in electronic format should email Emile Morse (emile.morse@nist.gov).

References

Borlund, P. 2003a. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research* **8**(3), Paper no. 152.

Borlund, P. 2003b. The concept of relevance in IR. *Journal of the American Society for Information Science* **54**(10), 913–925.

Chin, J. P., Diehl, V. A., and Norman, K. L. 1988. Development of an instrument measuring user satisfaction of the human–computer interface. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '88)*, Washington, DC, pp. 213–218.

Cowley, P., Haack, J., Littlefield, R., and Hampson, E. 2006. Glass box: capturing, archiving and retrieving workstation activities. In *Proceedings of the second ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '05)*, Santa Barbara, CA, pp. 13–18.

Cowley, P., Nowell, L., and Scholtz, J. 2005. Glass box: an instrumented infrastructure for supporting human interaction with information. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, Waikoloa, Hawaii.

Dang, H., Lin, J., and Kelly, D. 2007. Overview of the TREC 2006 question answering track. In E. Voorhees, and L. P. Buckland (eds.), *TREC2006, Proceedings of the Fifteenth Text Retrieval Conference*, Washington, DC. GPO.

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

den Os, E., and Bloothooft, G. 1998. Evaluating various spoken language dialogue systems with a single questionnaire: analysis of the ELSNET Olympics. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC '98)*, Granada, Spain, pp. 51–54.

Diekema, A. R., Yilmazel, O., Chen, J., Harwell, S., He, L., and Liddy, E. D. 2004. Finding answers to complex questions. In M. T. Maybury (ed.), *New Directions in Question Answering*, pp. 141–152. Cambridge, MA: MIT Press.

Dumais, S. T., and Belkin, N. J. 2005. The TREC Interactive Tracks: putting the user into search. In E. M. Voorhees, and D. K. Harman (eds.), *TREC: Experiment and Evaluation in Information Retrieval*, pp. 123–153. Cambridge, MA: MIT Press.

Harabagiu, S., Hickl, A., Lehmann, J., and Moldovan, D. 2005. Experiments with interactive question-answering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, pp. 205–214.

Hart, S. G., and Staveland, L. E. 1988. Development of a NASA-TLX (task load index): results of empirical and theoretical research. In P. Hancock, and N. Meshkati (eds.), *Human Mental Workload*, Amsterdam, North-Holland, pp. 139–183.

Hersh, W. 2006. Evaluating interactive question answering. In T. Strzalkowski, and S. Harabagiu (eds.), *Advances in Open Domain Question Answering*, pp. 431–455. Dordrecht, The Netherlands: Springer.

Hersh, W., and Over, P. 2001. Introduction to a special issue on interactivity at the Text Retrieval Conference (TREC). *Information Processing and Management* **37**(3): 365–367.

Kelly, D., Kantor, P., Morse, E. L., Scholtz, J., and Sun, Y. 2006. User-centered evaluation of interactive question answering systems. In *Proceedings of the Workshop on Interactive Question Answering at the Human Language Technology Conference (HLT-NAACL '06)*, New York, NY.

Kelly, D., Wacholder, N., Rittman, R., Sun, Y., Kantor, P., Small, S., and Strzalkowski, T. 2007. Using interview data to identify evaluation criteria for interactive, analytical question answering systems. *Journal of the American Society for Information Science and Technology* **58**(7): 1032–1043.

Liddy, E. D., Diekema, A. R., and Yilmazel, O. 2004. Context-based question-answering evaluation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, pp. 508–509.

Likert, R. 1932. A technique for the measurement of attitudes. *Archives of Psychology* **140**: 1–55.

Maiorano, S. J. 2006. Question answering: technology for intelligence analysis. In T. Strzalkowski, and S. Harabagiu (eds.), *Advances in Open Domain Question Answering*, pp. 477–504. Dordrecht, The Netherlands: Springer.

Maybury, M. T. 2004. Question answering: an introduction. In M. T. Maybury (ed.), *New Directions in Question Answering*, pp. 3–14. Cambridge, MA: MIT Press.

Small, S., Strzalkowski, T., Janack, T., Liu, T., Ryan, S., Salkin, R., Shimizu, N., Kantor, P., Kelly, D., Rittman, R., Wacholder, N., and Yamrom, B. 2004. HITIQA: scenario-based question answering. In *Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, Boston, MA, pp. 52–59.

Strzalkowski, T., and Harabagiu, S. 2006. *Advances in Open Domain Question Answering*. Dordrecht, The Netherlands: Springer.

Sun, Y., and Kantor, P. 2006. Cross-evaluation: a new model for information system evaluation. *Journal of American Society for Information Science and Technology* **56**(5): 614–628.

Tague-Sutcliffe, J. 1992. The pragmatics of information retrieval experimentation, revisted. *Information Processing and Management* **28**(4): 467–490.

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press

Voorhees, E. M. 2005. Question answering in TREC. In E. M. Voorhees, and D. K. Harman (eds.), *TREC: Experiment and Evaluation in Information Retrieval*, pp. 233–257. Cambridge, MA: MIT Press.

Voorhees, E. M., and Harman, D. K. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press.

Wacholder, N., Kelly, D., Rittman, R., Sun, Y., Kantor, P., Small, S., and Strzalkowski, T. 2007. A model for realistic evaluation of an end-to-end question answering system. *Journal of the American Society for Information Science and Technology* **58**(8): 1082–1099.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97)*, Madrid, Spain, pp. 271–280.

Woods, W. A., Kaplan, R. M., and Nash-Webber, B. L. 1972. *The Lunar Sciences Natural Language Information System: Final Report, BBN Report 2378*, Cambridge, MA.

https://doi.org/10.1017/S1351324908004932 Published online by Cambridge University Press