Technical Report 1036

Questionnaire Measuring the Utility of Knowledge-Based Systems

Leonard Adelman and James Gualtieri
Decision Systems Research

Sharon L. Riedel
U.S. Army Research Institute

Ann P. Trent
University of Kansas

January 1996

United States Army Research Institute
for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

Research accomplished under contract for the Department of the Army

Decision Systems Research

Technical review by

Theodore Shlechter
Joan Silver

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-STP, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1996, January	Final	November 1993-February 1995

4. TITLE AND SUBTITLE

Questionnaire Measuring the Utility of Knowledge-Based Systems

5a. CONTRACT OR GRANT NUMBER
DAAL03-91-C-0034

5b. PROGRAM ELEMENT NUMBER
0605803A DO 1082

5c. PROJECT NUMBER
D730

5d. TASK NUMBER

5e. WORK UNIT NUMBER

6. AUTHOR(S)

Leonard Adelman and James Gualtieri (DSR), Sharon L. Riedel (ARI), and Ann P. Trent (University of Kansas)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Decision Systems Research
3312 Tuckaway Court
Herndon, VA 22071

8. PERFORMING ORGANIZATION REPORT NUMBER

MIPR No. AECOM1DRS43075

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Artificial Intelligence Center
The Pentagon, Room 1D659
Washington, DC 20310-2000 (continued)

10. MONITOR ACRONYM
ARI

11. MONITOR REPORT NUMBER
Technical Report 1036

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Task was performed under Scientific Services Agreement issued by Battelle, Research Triangle Park Office, 200 Park Drive, P.O. Box 12297, Research Triangle Park, NC 27709. COR: Sharon L. Riedel

14. ABSTRACT *(Maximum 200 words)*:

This paper describes the development and validation of an off-the-shelf questionnaire designed to be tailored, as needed, to obtain the opinions of potential users of knowledge-based systems. Development began with a literature review to identify criteria used by different researchers to assess system utility and usability. The identified criteria then were organized into a multi-attributed hierarchy with the top three dimensions being Effect on Task Performance, System Usability, and System Fit. The bottom-level attributes were used to develop the questions for assessing system utility. In May 1994, the questionnaire was successfully tailored and used by the Army's Battle Command Battle Laboratory to evaluate 11 decision aiding prototypes. The questionnaire distinguished between those prototypes the soldiers liked and those that they did not. Psychometric analyses indicated the questionnaire passed required tests for reliability and validity.

15. SUBJECT TERMS

Questionnaire	Decision aids	Expert systems
Knowledge-based systems	Utility	Usability

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unclassified	77	

Unclassified
ARI Technical Report 1036

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) (Continued)

U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-RK
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Technical Report 1036

Questionnaire Measuring the Utility of Knowledge-Based Systems

Leonard Adelman and James Gualtieri
Decision Systems Research

Sharon L. Riedel
U.S. Army Research Institute

Ann P. Trent
University of Kansas

Fort Leavenworth Research Unit
Stanley M. Halpin, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

January 1996

Army Project Number
2O665803D730

Personnel and Training
Analysis Activities

ACKNOWLEDGMENT

This work was sponsored by the U.S. Army Artificial Intelligence Center and the U.S. Army Research Institute for the Behavioral and Social Sciences under the auspices of the U.S. Army Research Office Scientific Services Program administered by Battelle (Delivery Order 1082, Contract No. DAAL03-91-C-0034).

QUESTIONNAIRE MEASURING THE UTILITY OF KNOWLEDGE-BASED SYSTEMS

EXECUTIVE SUMMARY

Research Requirement:

To develop a questionnaire that could be easily tailored as needed to obtain the opinions of potential users of knowledge-based systems and other types of decision aids.

Procedure:

Development began with a literature review to identify criteria used by different researchers to assess system utility and usability. The identified criteria then were organized into a multi-attributed hierarchy with the top three dimensions being Effect on Task Performance, System Usability, and System Fit. The bottom-level attributes were used to develop the questions for assessing system utility. Multi-Attribute Utility Assessment (MAUA) concepts were used to combine the answers to these questions into utility measures for all nodes in the hierarchy. The validation effort focused on assessing if the questionnaire (a) could be tailored to different decision aiding systems, and (b) possessed good psychometric characteristics.

Findings:

The questionnaire was successfully tailored and used by the Army's Battle Command and Battle Laboratory (BCBL) to evaluate eleven decision aiding prototypes during the Prairie Warrior exercise in May 1994. BCBL personnel were able to identify a subset of utility criteria and attributes of critical concern to them, and the research team was able to develop a short version of the questionnaire that both measured these attributes and could be administered in 10 to 15 minutes. The resulting questionnaire was capable of distinguishing between those prototypes the soldiers liked and those that they did not. Moreover, psychometric analyses focusing on the questionnaire's split-half reliability and construct validity indicated that the questionnaire passed required tests for reliability and validity.

Utilization of Findings:

The questionnaire can be used as an off-the-shelf tool to obtain soldiers' opinions about the utility and usability of Army decision aiding systems, particularly early in the development life cycle.

QUESTIONNAIRE MEASURING THE UTILITY OF KNOWLEDGE-BASED SYSTEMS

CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

QUESTIONNAIRE MEASURING THE UTILITY OF KNOWLEDGE-BASED SYSTEMS

Introduction

This report describes the development and validation of a user questionnaire for assessing the utility and usability of decision aiding systems, including knowledge-based systems (KBSs). The goal was to develop a psychometrically valid questionnaire that could be easily tailored to the needs of different development efforts.

Berry and Hart (1990, p.200) argue that "The ultimate criterion of success for most ... systems is whether they will actually be used, and used to effect, by individuals other than the system developers." To help ensure that the final system will be used and useful, it is important to get users involved early in the development process, and keep them involved throughout it (See Gray, Roberts-Gray, & Gray, 1983; Shlechter, Bessemer, Rowatt, & Nesselroade, 1994; Shlechter, Brunside, & Thomas, 1987). If user assessments are considered early in development, changes to the system to reflect user needs will be relatively easy and inexpensive to make.

Several researchers (e.g., Adelman, Gualtieri, & Riedel, in press; Mitta, 1991; Nielsen, 1993; Sweeny, Maguire, & Shackel, 1993) discuss the process of matching different usability and utility assessment methods and measures to different stages in the system development life cycle. Of the many potential methods for obtaining user assessments, a questionnaire can be used early in system development and can obtain a standard set of information from a number of users while minimizing data collection time. However, if the evaluator constructs a new questionnaire for each system and each data collection occasion, the questionnaire will have unknown psychometric properties, will be time consuming to construct, and may not assess the most appropriate aspects of the system. Further, in the rapid prototyping development environment of knowledge-based systems, the window of opportunity to construct the questionnaire, obtain user feedback, and make recommendations for changes is often small and easily missed. One form of assistance to address these problems is a standard user questionnaire that can be quickly and easily tailored for different systems and for different stages of development.

This paper describes an "off-the-shelf" user questionnaire that contains a set of standard evaluation dimensions with ready made questions for each of the dimensions and a standard response format. The evaluator and sponsor select the dimensions that are of interest, are appropriate to test at the current maturity level of the prototype, and are appropriate for the user to answer. The questions for the desired dimensions are then tailored by the evaluator to make them appropriate for the system being evaluated and printed out using a standard format. With standard dimensions and questions, results can be compared between stages of development of the same system, between different systems, and to benchmark standards, when they are developed.

The questionnaire was developed in four steps. First, the authors performed a review to identify the different attributes of utility and usability defined in the literature. Second, a Multi-Attribute Utility Assessment (MAUA) hierarchy was created for combining individual usability attributes into broader utility concepts. The broader utility concepts include the system's effect on task performance, the usability of the human-computer interface, and the system's fit into the larger organization where it will be used.

Third, two or more questions were developed for measuring each bottom-level attribute in the hierarchy. The questions use a seven-point rating scale and were written in a general nature that permits developers (or evaluators) to modify them for different development efforts. Consistent with our goal, the questionnaire provides the advantages of (a) presenting a universe of dimensions for which developers might be interested in users' opinions, and (b) for each dimension, providing ready-made questions that can be tailored to the decision-aiding system under consideration. The fourth step in the development process was to pilot-test the questionnaire to ensure its content validity, and pre-test it (Adelman, Gualtieri, & Riedel, 1993) to demonstrate good, albeit preliminary, psychometric characteristics.

The goal of the validation effort was to ensure that (a) the questionnaire could be tailored to different KBSs and decision support systems, and (b) it possessed good psychometric characteristics. To achieve this goal, the questionnaire was used by five government and contractor employees to evaluate eleven different prototypes used during a military exercise. Prior to the exercise, senior Army personnel at the Battle Command Battle Laboratory (BCBL) identified the attributes in the hierarchy for which they wanted data, and the questionnaire was tailored to provide these data for each prototype. The study, which is described herein, showed that the questionnaire could be completed quickly, and that it could distinguish between those prototypes the soldiers liked and those they didn't. Moreover, psychometric analyses indicated that the questionnaire passed required tests for reliability and validity.

The term "decision aiding systems" is used throughout the paper to refer to different types of KBSs, such as expert systems (ES) and decision-analytic aids, and more general decision support systems (DSS).

The development and validation of the questionnaire are now considered, in turn.

Questionnaire Development

This section is divided into three parts. The first part describes the literature review; the second part describes the MAUA hierarchy of utility and usability attributes around which the questionnaire was developed; and the third part describes the questionnaire. The pretesting of the questionnaire's psychometric properties is presented in Adelman et al. (1993), and is not considered herein other than by comparing its procedures and results with those used in the validation effort described later in this paper.

Literature Review

The original purpose of the review was to identify the different definitions of system usability found in the literature. Based on previous research (e.g., Adelman, Rook, & Lehner, 1985), it was known that the definitions would be multi-faceted. These different facets represent the many different attributes that researchers have used to define usability. The goal was to have as broad a scope as possible, so that developers and evaluators would be able to tailor the questionnaire to measure those usability attributes of concern to them.

A systematic, structured approach was used to guide the literature review. Databases searched were ERIC, National Technical Information System (1990-1993) and Academy of Management (1988-1992). Keywords used for the search included usability, utility, man-machine interface, human-computer interface, ease of use, usefulness, decision support system, knowledge-based system, expert system, and decision aid.

As hypothesized, the review failed to find a readily available questionnaire that could be used as an off-the-shelf tool for obtaining users' opinions about decision-aiding systems. Consequently, the application need driving the questionnaire development effort was still appropriate.

A principal finding of the literature review was, as suspected, that usability is indeed a multi-faceted term. There is considerable disagreement on the definition given to usability in the literature. For example, we found that our use of "usability" did not match its use by other researchers. Our focus was on assessing the usability of the HCI. However, some researchers (e.g., Berry & Hart, 1990; Hammond, Morton, Barnard, Long, & Clark, 1987; Susskind, 1988; Hockey, Briner, Tattersall, & Wiethoff, 1989; Marshall, Nelson, & Gardiner, 1987)) took a much broader focus, basically equating system usability to the system's usefulness or utility in the user's actual environment. HCI usability was one aspect of this much broader focus.

Since the goal was to develop a tool that could collect user assessments of a wide variety of system aspects, we changed the focus from usability to the broader utility focus. The term "utility" will be used hereafter to convey this broader focus. "Usability" will be used only when referring to HCI usability. The reader should keep in mind, however, that different researchers use different terms.

Table 1 presents a simple pictorial representation of the different attributes (or characteristics) that different researchers used to define utility. The attribute names tended to be those that we have used previously (Adelman, 1992; Riedel, 1992) and, thus, tended to be the ones we used when constructing the questionnaire. The researchers, who are listed chronologically, were considering the wide array of systems and the broad utility focus when presenting their attributes. A shadowed cell entry indicates that the researcher used the attribute when defining system utility.

Examination of Table 1 shows that the researchers did not agree on what attributes should be used to measure system utility. There are only a few shadowed cells in any given column, and minimal agreement in the shadowed cells across columns. Moreover, this disagreement is even stronger than it appears if one examines specific definitions for the criteria.

Table 2 presents the concepts that different researchers used to define their attributes. The researchers' exact words were used, except in those cases where the concept's meaning was not intuitively obvious. Again, it can be seen how differently the researchers defined many of the attributes. For example, the quality attribute was defined in terms of "productivity," "effectiveness," "error reduction," "sensitivity to parameter changes," "accuracy," "performance," "usefulness of results," "verification," "validation," "functionality" and "quality of process and products." Some of these concepts probably mean the same thing, such as "performance" and "effectiveness." Other concepts, however, such as distinguishing between the quality of the process versus the products, are tapping different attributes of quality.

Nevertheless, the attributes can be combined at a more general level to form three utility dimensions: Effect on Task Performance, System Usability (in terms of HCI), and System Fit into the larger organization. The attributes that comprise these three utility dimensions are separated by bold horizontal lines in Tables 1 and 2.

Table 1

Pictorial Representation of the Utility Attribute Defined by Different Researchers

	Bennet (1984)	Shakel (1986)	Hammond *et al.* (1987)	Marshal *et al.* (1987)	Morris (1987)	Cleal & Heaton (1988)	Clegg *et al.* (1988)
Quality	■	■	■	■		■	■
Confidence							
Acceptability		■	■		■		
Reliability							
Timeliness						■	■
Ease of Use			■	■	■		
Use of Data			■				■
Learnability	■	■				■	■
Explanation				■			
Workload				■			■
Flexibility		■	■	■	■		■
Fit							
Interoperability			■				
Other's Attitudes					■		

Table 1

Pictorial Representation of the Utility Attribute Defined by Different Researchers (continued)

	Susskind (1988)	Hockey (1989)	Ravden & Johnson (1989)	Berry & Hart (1990)	Berry & Hart (1991)	Holcomb & Tharp (1991)	Adelman/ Riedel (1992)
Quality	■	■	■	■	■	■	■
Confidence					■		■
Acceptability							■
Reliability	■			■			■
Timeliness				■	■		■
Ease of Use	■		■	■	■	■	■
Use of Data	■		■	■	■		■
Learnability			■			■	■
Explanation			■		■	■	■
Workload		■				■	■
Flexibility		■	■			■	
Fit					■		■
Interoperability							■
Other's Attitudes					■		

Table 2

The Actual Concepts that Different Researchers Used to Define Their Utility Attributes

	Bennet (1984)	Shakel (1986)	Hammond *et al.* (1987)	Marshal *et al.* (1987)	Morris (1987)	Cleal & Heaton (1988)	Clegg *et al.* (1988)
Quality	Productivity	Effectiveness	Sensitive to Small Changes in Params.	Error Reduction		Error Rate; Failure Time	Error Prevention & Correction
Confidence							
Acceptability		Attitude Accept.	Face Validity		Level of Enjoy.; Perceived Use.		
Reliability							
Timeliness						System Speed	System Speed
Ease of Use			Simple; Systematic	Familiar; Position Known; Consistent	Ease of Use		
Use of Data			Knowledge Base System Uses All Available Data				Info. I/O (i.e., Compatibility)
Learnability	Training Time	Learnability				Training Time	Ease of Learning
Explanation				Feedback			
Workload				Match Cogn. Abilities			Degree of Effort
Flexibility		Flexibility	Flexibility	Adaptive; Locus of Control	Level of Direction		Being in Control
Fit							
Interoperability			Similar Modules				
Other's Attitudes					Other's Attitudes		

Table 2

The Actual Concepts that Different Researchers Used to Define Their Utility Attributes (continued)

	Susskind (1988)	Hockey (1989)	Ravden & Johnson (1989)	Berry & Hart (1990)	Berry & Hart (1991)	Holcomb & Tharp (1991)	Adelman/ Riedel (1992)
Quality	Accuracy	Performance	Error Prevention & Correction	Usefulness of Results, V&V	Accuracy; Overall Effect	Functional/Task Accomplish.	Quality of Process and Products
Confidence					Confidence		Confidence
Acceptability							Acceptability of Process, Results & Represent. Scheme
Reliability	Completeness			Reliability			Reliability
Timeliness				Speed	Time		Timeliness
Ease of Use	Logical		Visual Clarity; Explicitness; Functionality	Intelligibility	Match with User	Consistency; Intuitive	Ease of Use
Use of Data	Consistent Info.		Compatability & Consistency	Usefulness of Data	Form of Data		Use of Data
Learnability			Guidance & Support			User Help	Ease of Training; Documentation
Explanation			Information Feedback		User Support	Feedback	Explanation Capability
Workload		Workload				Minimal Memory	Workload
Flexibility		Discretion	Flexibility & Control			User Control	
Fit					Org. Match		Organization Fit with Doctrine
Interoperability							Interoperability of Different Systems
Other's Attitudes					Effect on Others		

Table 3 shows that there is considerable agreement among researchers if one looks only at these three general dimensions; that is, the cell is filled if the researcher identifies any of the attributes comprising the dimension. The agreement is striking for Effect on Task Performance and System Usability. Although the specific attributes may be different, every researcher identified at least one attribute for these two dimensions. In contrast, less than half the researchers identified a single attribute for System Fit.

The results of the literature review support the application of Multi-Attribute Utility Assessment (MAUA) as a conceptual framework for defining system utility. First, each researcher used multiple attributes to define the concept of system utility (i.e., usefulness or value). Second, different researchers use different attributes in their definitions. Sometimes the different sounding attributes meant the same thing, but in many cases they were defining different aspects of system utility. Third, even given these differences, three broad utility dimensions could be identified for categorizing the attributes. This is a necessary prerequisite for using a MAUA framework.

With these findings in mind, a MAUA hierarchy of attributes was developed to define system utility. This hierarchy, and the broader approach to using MAUA as a basis for questionnaire development, are described in the following section of the paper.

<u>MAUA Hierarchy</u>

Figure 1 presents the MAUA hierarchy of utility and usability attributes used to develop the questions for the questionnaire. As can be seen, **Overall System Utility** is decomposed into three broad categories or groupings of attributes: **Effect on Task Performance, System Usability, and System Fit**. Each of these three broad category represents an upper-level branch of the hierarchy, and is referred to hereafter as a dimension. Each dimension is, in turn, decomposed into different sub-groups of attributes, called criteria. Each criterion may be further decomposed into the specific attributes identified in the literature.

The questions in the questionnaire assess a system against the lowest-level attributes and criteria, if a criterion is not further decomposed into attributes in the hierarchy. A system's score on each dimension is a weighted average of the system's scores on the lower-level attributes and criteria that comprise it. This section of the paper briefly overviews the hierarchy of utility dimensions, criteria, and attributes moving down the MAUA hierarchy. Dimensions and the overall utility node in the hierarchy are presented in bold, capital letters; criteria are underlined; and attributes are presented in regular type. The next section overviews the questionnaire's characteristics,

Table 3

Agreement Among Researchers at the Level of General Utility Dimensions

	Bennet (1984)	Shakel (1986)	Hammond *et al.* (1987)	Marshal *et al.* (1987)	Morris (1987)	Cleal & Heaton (1988)	Clegg *et al.* (1988)
Task Performance	■	■	■	■	■	■	■
System Usability	■	■	■	■	■	■	■
System Fit			■		■		

	Susskind (1988)	Hockey (1989)	Ravden & Johnson (1989)	Berry & Hart (1990)	Berry & Hart (1991)	Holcomb & Tharp (1991)	Adelman/ Ridel (1992)
Task Performance	■	■	■	■	■	■	■
System Usability	■	■	■	■	■	■	■
System Fit					■		■

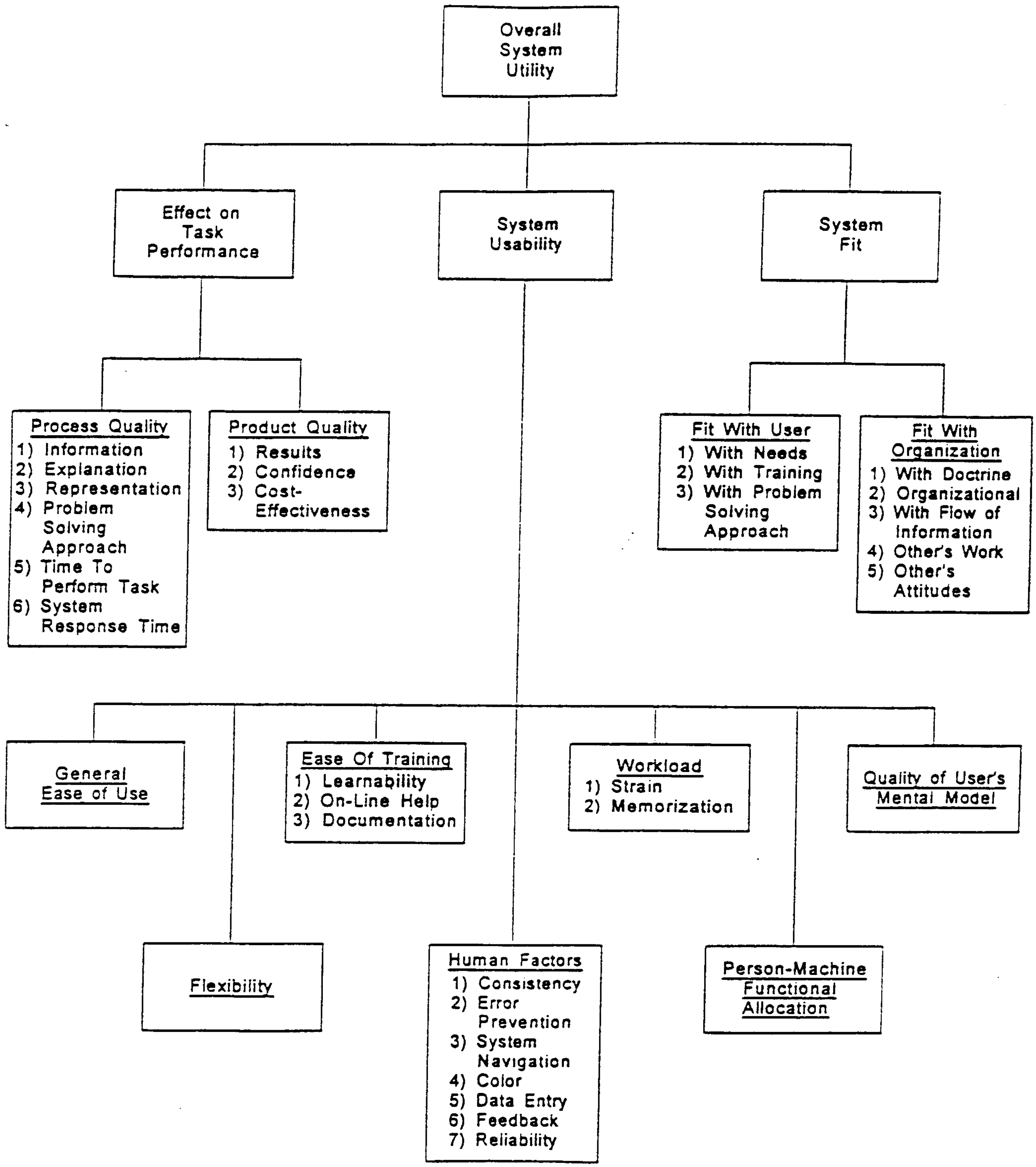

Figure 1. A MAUA evaluation hierarchy for assessing users' opinions about system utility.

including how one obtains scores on attributes, criteria, and dimensions moving up the hierarchy.

The first dimension, **Effect on Task Performance**, is composed of two criteria: process quality and product quality. The distinction between process and product quality has been made by a number of researchers, including Adelman (1992), Berry and Hart (1990), and Riedel (1992).

Process quality is composed of six attributes: (1) the quality of the system's information (i.e., data), (2) its explanation capability, (3) its knowledge representation scheme, (4) its problem solving approach, (5) the time to perform task, and (6) the system's response time. Each attribute is defined in turn.

Quality of information is the degree to which the system used the correct information in making its recommendations. Quality of explanation is the ability of the system to convey to the users how it arrived at its results. Quality of representation is the ease with which the user can understand and modify the judgments which the system uses to make its recommendations. Quality of the problem solving approach is how well the system represents the logic structure necessary for solving the problem, and how acceptable this representation is to the user. Performance time is the time it takes users to complete the task using the system. And response time is the amount of time it takes the system to respond to users' inputs and produce outputs.

The product quality criterion is decomposed into: (1) the users' assessment of the quality of results generated with the system, (2) their confidence in the overall products produced by using the system, and (3) its overall cost-effectiveness. Quality of results is the quality of the system's recommendations and accompanying explanations, analysis and reports. Berry Hart (1990) have a similar construct called usefulness of results; Holcomb and Tharp (1991) also have a related concept known as functionality.

Overall confidence is defined as a measure of how sure users are that the system is working effectively and giving them the correct answer. Cost-effectiveness is a measure of the efficiency of the system. Benefits and costs can be at the individual user level or at the organizational level. For example, benefits could be increased speed or higher quality output; while costs could include level of expertise and effort required to run the system.

The second dimension for assessing system utility is **System Usability**. **System Usability** is decomposed into seven criteria: (1) general ease of use, (2) flexibility of operations, (3) ease of training, (4) human factors attributes, (5) workload issues, (6) adequacy of the allocation of functions between the person and machine, and (7) the quality of user's mental model of how the system operates. Where necessary these criteria were further decomposed into relevant attributes.

The first criterion, general ease of use, is simply how easy the system is operate. Similar criteria have been proposed by Barry and Hart (1990), Clegg et al. (1988), Holcomb and Tharp (1991), and Ravden and Johnson (1989).

Flexibility, the second criterion, is defined as the degree of user discretion and user control. This criterion is a measure of the degree to which the operator of the system determines the type and rate of work rather than the system setting the agenda and pace. Hockey, Briner, Tattersall, and Wiethoff (1989) had a similar criterion labeled level of discretion.

The third criterion is ease of training. This criterion has three attributes that, together, characterize how easy the system is to learn. The three attributes are: (1) learnability, (2) on-line help functions, and (3) documentation. Holcomb and Tharp (1991) also identify this last dimension (documentation) as an important criteria in evaluating systems.

The fourth system usability criterion, human factors guidelines, is defined as the degree to which the system follows prescribed HCI tenets. Seven attributes contribute to the rating of this criterion. They are: (1) consistency, (2) error prevention, (3) system navigation, (4) use of color, (5) ease of data entry, (6) system feedback, and (7) reliability. A number of these attributes also were identified by other theorists. For example, Holcomb and Tharp (1991) proposed feedback as a criterion for system evaluation.

Workload is the fifth system usability criterion. Workload has two attributes: strain and memorization. Strain is defined as the amount of physical or mental pressure imposed by the system on the user. Memorization is the amount of information that the user is required to remember in order to operate the system. Clegg et al. (1988), Hockey et al. (1989), and Holcomb and Tharp (1991) also included measures of workload in their set of usability criteria.

The sixth usability criterion is functional allocation. This criterion measures the degree to which activities allocated to the machine are appropriate for the system to do. That is, the system performs those functions that the user wants it to perform.

The final usability criterion is the quality of the user's mental model of the system. This is defined in terms of how easy it is for a user to understand the system's layout and features. It is the match between the user's mental model of the system and the actual features of the system. The greater the match between the system and the user's model of the system, the easier it will be for the user to work with it.

The last dimension of the utility hierarchy is **System Fit.** This dimension has two criteria: fit with users and fit with organization. This dimension measures the degree to which the characteristics of the system match those of the users and their organization. The greater the degree of fit, the more likely the system will be utilized. Adelman et al. (1985) found this dimension to be highly correlated with domain experts' judgments of the potential utility of decision support and expert system prototypes.

Fit with user has three attributes: (1) match with needs, (2) match with training, and (3) match with problem-solving approach. Each of these attributes affects the degree to which the user can easily understand and work with the system. The extent to which there is a match between the user and the system determines how quickly the user will be able to utilize the system and how much training will be required.

Fit with organization has five attributes: (1) match with organizational doctrine, (2) organizational fit, (3) effect on information flow, (4) effect on other people's workload, and (5) attitude of others toward system. These attributes contribute to the probability that an organization will use a system. If the new system is not viewed positively by upper-level management, then it will not be implemented by the organization. Similar arguments can be made for each of the other dimensions.

This section has described a hierarchy of utility dimensions, criteria and attributes derived from the literature. The next section will show how this heirarchy is used to construct a questionnaire measuring utility.

The Questionnaire

Elsewhere we have discussed how MAUA can be used to evaluate decision aiding systems in general, and KBSs, ESs, and DSSs in particular (Adelman, 1992; Adelman et al., in press; Riedel & Pitz, 1986). Here, we only consider how we used the general MAUA approach to develop a questionnaire for obtaining users' subjective assessment of the perceived utility of such systems.

MAUA is typically used to evaluate the relative utility of competing alternatives. This is done by implementing six general steps: (1) specifying the evaluation dimensions, criteria, and attributes; (2) differentially weighting the dimensions, criteria, and attributes in terms of their relative importance; (3) scoring each of the alternatives against each of the bottom-level attributes and criteria; (4) creating utility functions so that the scores on the different attributes (and bottom-level criteria in the hierarchy) can be placed on the same utility scale; (5) summing the weighted utility scores for the alternatives; and (6) selecting the alternative with the highest overall utility score. Sensitivity analysis is used to assess how sensitive the highest rated alternative is to changes in the scores, utility functions, and relative importance weights.

Conceptually, the same steps were used to develop a questionnaire for obtaining subjective assessments of system utility. First, we used the results of the literature review to develop the MAUA hierarchy of utility dimensions, criteria, and attributes described above.

Second, we gave equal weights to the attributes comprising each criterion, such that the weights summed to 1.0. The exact weight given any attribute depended on the number of attributes comprising the criterion. The larger the number of attributes comprising a criterion, the smaller the weight on each attribute, so that the weights would sum to 1.0. Equal weights were used because the literature review did not provide any empirical basis for saying that one attribute was more important than another in defining any of the criteria. By using equal weights, we were simply averaging the system's scores on the attributes comprising any given criterion.

Similarly, we also gave equal weights to the criteria for each dimension. Again, there was no empirical basis for saying that one criterion was more important than another in determining users' judgments of **Effect on Task Performance**, **System Usability**, or **System Fit**.

The three dimensions also were given equal weights. Although the literature review suggested that **Effect on Task Performance** and **System Usability** are the two most frequently defined utility dimensions, we could think of instances where **System Fit** would be more important in determining a system's overall value to its potential users. More importantly from the perspective of developing the questionnaire, a MAUA approach lets users and evaluators specify the relative importance of different dimensions and criteria, as appropriate for tailoring the questionnaire to their particular context.

The third step in a MAUA application is scoring the alternatives against the bottom-level attributes and criteria. The questionnaire represents this scoring mechanism in this case. Specifically, there were two or more questions for obtaining users' opinions about the system for each of the bottom-level attributes and criteria in the hierarchy. Users' answers to these questions indicate, at a particular point in the life cycle, how well the system is doing on each bottom-level attribute and criterion. These scores represent feedback developers can used to improve system utility and usability.

Each question in the questionnaire is in the form of a statement. Participants answer the questions by using a 7-point scale going from "strongly disagree" (1) to "strongly agree" (7), with "neither agree nor disagree" (4) as the mid-point. Space is provided after each statement to provide room for the participants to write "comments" explaining their responses, if they choose to do so.

For the fourth step in the MAUA, that of creating a utility scales for each of the bottom-level attributes and criteria, we assumed that the seven-point scale used for the questionnaire represented a utility scale. The questions were written so that higher scores always meant that the system was performing better on the attribute (or criterion) being assessed by the question. We also assumed that the utility scale was a linear function, a reasonable assumption according to Edwards (1977) and Huber (1980) in most situations.

The scores and weights are combined by simple arithmetic operations to implement the fifth and sixth steps of the MAUA. By doing so, one obtains a user's assessment of the overall utility of the system. Specifically, one obtains a criterion score, indicating how well each participant thought the system performed on a criterion, by averaging each participant's answers to the questions measuring that criterion. (In those cases where the criterion was decomposed into separate attributes, one first averages the answers for the questions measuring the attributes, and then averages the scores for the attributes to obtain the criterion scores.) Then, moving up the hierarchy, one obtains a dimension score by multiplying each criterion's score by its weight, and then summing up the products for the criteria that comprise a dimension. As we noted earlier, each criterion that comprises a dimension was considered equally important; consequently, a dimension score is equal to the average of the criteria scores.

Finally, one obtains an overall utility score for the system, by summing the products of the dimension scores and their corresponding (equal) weights. One can obtain an average score for the participants who completed the questionnaire, at each level of the hierarchy, by averaging their scores at the appropriate levels. Sensitivity analysis can be performed by determining how sensitive the overall utility score is to changes in the relative weights on the criteria and dimensions, or to the system's scores on the criteria and attributes.

The next section of this paper describes the questionnaire validation effort. Before turning to it, we make seven points. First, Appendix A lists the 96 questions in the complete questionnaire. These questions are organized within the context of the complete MAUA hierarchy described above.

Second, ninety (90) of the 96 questions in the questionnaire assess the bottom-level criteria and attributes in the hierarchy. The other six questions assess the participants' global judgment as to the overall utility of the system. As will be illustrated in the next section of the paper, the global utility judgments can be correlated with the overall utility score calculated by the MAUA hierarchy (node 0.0) to assess their agreement and, in turn, the construct validity of the questionnaire.

Third, there are at least two questions for each bottom-level criterion and attribute in the hierarchy. At least one question is in each half of the questionnaire; and, on the average, each half of the questionnaire has half the questions for each bottom-level criterion and attribute. This permits one to calculate a split-half reliability score for the questionnaire. This is a psychometric measure indicating the extent to which questions that are supposedly measuring the same attribute (or criterion) are, in fact, doing so. Said differently, if the questionnaire is a reliable measuring instrument, then there should be a high correlation between the two halves of the questionnaire, for the questions were presumed to be measuring the same attributes (and criteria). In the future, the two halves of the questionnaire can constitute two separate versions of the questionnaire, each taking 15-20 minutes to administer.

Fourth, the reason that there are more than two questions for each bottom-level criterion and attribute in the current version of the complete questionnaire is so that, through repeated application of the questionnaire, one can determine which questions (supposedly) measuring the same criterion (or attributes) correlate the highest. These would be the only questions retained in later versions of the questionnaire.

Fifth, the initial version of the questionnaire had some questions phrased so that the respondent had to disagree with the statement, that is, give a low score in order to evaluate the system highly. This was done to ensure that respondents carefully read each question before responding. However, pilot-testing of the questionnaire indicated that these (reversed) statements were unnecessarily complex, and slowed down the respondents' speed in answering the questionnaire. Consequently, all questions were constructed so that higher scores on the scale meant a system was doing better.

Sixth, considerable efforts were made to ensure that the questions had content (or face) validity. The first and third authors have considerable experience working with Army personnel and evaluating military decision aids. In addition, the questionnaire was developed by surveying other researchers and evaluators to determine what dimensions and items they included in their questionnaires. This helped ensure that the set of items in the questionnaire represented the range of factors that users consider when evaluating system utility. And, finally, one active-duty Army officer and one retired Army officer participated in the pilot test, as did other personnel at ARI who have experience working with Army personnel. This helped to ensure that the questions used the right phrases and jargon and, basically, sounded right to the respondents.

Seventh, the complete questionnaire was completed by three participants in a small, limited evaluation of a KBS prototype for supporting Army tactical planners. The psychometric analysis for that application was quite encouraging. Neither, the pilot test nor the prior administration of the complete questionnaire is discussed herein; the interested reader is referred to Adelman et al. (1993).

We now consider the questionnaire validation effort.

Validation Effort

As noted in the Introduction, the goal of the validation effort was to ensure that (a) the questionnaire could be tailored to different decision aiding prototypes, and that (b) it possessed good psychometric characteristics. Each concern is addressed in turn.

The Application

The questionnaire was tailored to evaluate the utility of eleven decision aiding prototypes demonstrated in May 1994 during a week-long exercise, called Prairie Warrior, at Fort Leavenworth, Kansas. The tailoring process was implemented in two general steps.

First, senior personnel in the Battle Command and Battle Laboratory (BCBL) decided which dimensions, criteria, and attributes should be included in the questionnaire. They decided that the all eleven prototypes should be evaluated on all three dimensions: **Task Performance**, **Usability**, and **System Fit**. However, because (1) all the prototypes were still early in the development life, and (2) were only being demonstrated with limited time available for their use, BCBL personnel decided that the questionnaire should measure only seven (of the eleven possible) criteria. The criteria were process quality, product quality, quality of user's mental model, human factors guidelines, flexibility, fit with user, and fit with organization. Most of the attributes for these criteria were, in turn, selected for inclusion in the questionnaire.

The second step was to select specific questions for the questionnaire. Thirty-three of the 96 questions in the complete questionnaire were used in the questionnaire tailored for the Prairie Warrior exercise. For example, to measure the attribute "quality of the results," the evaluation team tailored two of the four standardized questions in the complete questionnaire for measuring this attribute. One of these two questions was, for example, "The system provided the user with useful results." In order to save time in completing the questionnaire, only ten (10) nodes (seven attributes, two criteria and the overall utility score) had two questions, one each in the first and second half of the instrument, for the split-half reliability calculations.

The questionnaires were completed by five Army data collectors. On average, each decision aiding prototype was evaluated by three of the five data collectors. It took the data collectors 10-15 minutes to answer the 33 questions, including written comments if they choose to elaborate on their answers. In addition, it took them another 10-15 minutes to complete an additional set of 10 to 12 questions, depending on the decision aid. Some of the additional questions were added to the questionnaire in order to perform the construct validity analysis described in the next section of the paper. Other questions were relevant to other data collection efforts for the Prairie Warrior exercise. None of the participants had any problems using the questions considered herein. Appendix B presents the questionnaire for one of the prototypes.

Table 4 presents the mean scores for all nodes in the hierarchy. The mean score for each attribute (i.e. the lowest level) in Table 4 was calculated by averaging all the responses for all the prototypes and all the data collectors answering the questions measuring the prototypes. The mean score for each criterion was calculated by averaging the mean scores for all the attributes comprising it. The mean score for a dimension was calculated by averaging the mean scores for the criteria comprising it. The mean Overall Utility score was calculated by averaging the mean scores for the three dimensions.

Although the sample size is too small to reach a conclusive position, the means provide an idea of what an acceptable level for the system utility score would be using the questionnaire. Because the questionnaire was used to evaluate a variety of decision aiding prototypes, the mean values could serve as benchmarks for future researchers. Prototypes that scored considerable lower than the means in Table 4 would be thought of

Table 4

Mean, Maximum, and Minimum Values for the Utility Dimensions, Criteria, and Attributes

Node	Mean	Max.	Min.	N
0.0 Overall System Utility [Based on Hierarchy]	4.85	7.00	2.85	18
[Based on Questions #1 & #23]	5.68	7.00	2.00	33
1.0 Effect on Task Performance	5.07	7.00	3.33	21
1.1 Process Quality	4.79	7.00	3.17	21
1.1.1 Quality of the Information [Questions #2]	4.91	6.00	2.00	22
1.1.2 Quality of the Explanation Capability/Reasons	----	----	----	
1.1.3 Quality of the Representation, Examination, and Modification of Knowledge Stored in System [Questions #3]	4.33	7.00	2.00	21
1.1.4 Quality of the Problem Solving Approach [Question #4]	5.15	7.00	2.00	26
1.1.5 Time to Perform Task(s) [Questions #21,29]	4.81	7.00	1.00	32
1.1.6 System Response Time [Questions #5]	4.69	7.00	2.00	26
1.2 Product Quality	5.35	7.00	3.50	21
1.2.1 Quality of the Results [Questions #6,24]	5.49	7.00	2.00	23
1.2.2 Overall Confidence [Questions #7,25]	5.21	7.00	2.00	21
1.2.3 Cost-Effectiveness	----	----	----	
2.0 System Usability	4.59	7.00	2.33	18
2.1 General Ease of Use Questions	----	----	----	
2.2 Quality of the User's Mental Model of the System [Questions #20,22,30]	4.73	7.00	2.00	21
2.3 Ease of Training	----	----	----	
2.3.1 Learnability	----	----	----	
2.3.2 On-Line Help Function	----	----	----	
2.3.3 Documentation	----	----	----	

Table 4

Mean, Maximum, and Minimum Values for the Utility Dimensions, Criteria, and Attributes (continued)

Node	Mean	Max.	Min.	N
2.4 Human Factors Guidelines for Person-Machine Interaction	4.41	7.00	2.00	20
2.4.1 Consistency	----	----	----	
2.4.2 Error Prevention and Handling [Questions #16]	3.55	5.00	2.00	20
2.4.3 System Navigation [Questions #17,28]	4.21	6.00	2.00	20
2.4.4 Use of Color [Questions #18]	5.09	7.00	2.00	22
2.4.5 Ease of Data Entry [Questions #33]	4.69	7.00	2.00	26
2.4.6 Feedback [Questions #19]	4.52	7.00	2.00	23
2.4.7 Reliability	----	----	----	
2.5 Workload	----	----	----	
2.5.1 Strain	----	----	----	
2.5.2 Memorization	----	----	----	
2.6 Flexibility [Questions #15,27]	4.63	7.00	1.00	18
2.7 Functional Allocation Between Person and Machine	----	----	----	
3.0 System Fit	4.90	7.00	2.88	18
3.1 Fit (i.e., Match) With User	4.61	7.00	2.00	18
3.1.1 Match With Users' Needs [Question #8]	4.32	7.00	2.00	22
3.1.2 Match With Users' Training	----	----	----	
3.1.3 Match With Users' Problem-Solving Approach [Questions #9]	4.29	6.00	2.00	18
3.2 Fit (i.e., Match) With Organization	5.18	7.00	3.13	21
3.2.1 Match With Doctrine [Questions #10,26]	4.52	7.00	2.00	21
3.2.2 Organizational Fit [Question #11]	5.46	7.00	3.00	33
3.2.3 Effect on Information Flow [Questions #12,31]	5.37	7.00	2.00	33
3.2.4 Effect on Other's Workload [Questions #13]	5.15	7.00	2.00	33
3.2.5 Attitude of Others [Questions #14,32]	5.39	7.00	2.00	33

as needing additional work, while those that scored above this level would require less, if any, improvement on the relevant attributes, criteria, and dimensions.

Examination of Table 4 shows that the mean **Overall Utility** score for the eleven prototypes was 4.85 on the 7-point scale, with 4.0 as the scale midpoint. This mean score indicates that, on the average, the data collectors thought the prototypes had more positive than negative attributes. Although this mean score is not high, we think it is an acceptable mean value for eleven initial prototypes.

It should be noted that the mean response to the two questions directly assessing the prototypes' Overall Utility was 5.68. This mean was considerably higher than the Overall Utility mean of 4.85 calculated from the attribute scores. There are several reasons for a discrepancy between a global judgement of overall utility and an overall utility score based on the integration of lower level attribute scores. First, the limitations of short term memory make it difficult for people to integrate a large amount of information (Hogarth, 1987). To reduce the mental effort involved in integrating information for a global judgement, people tend to use simplifying heuristics such as basing their judgement on only a few factors. This means that each data collector's judgement of overall utility was based on a small number of factors rather than the 21 factors that make up the calculated overall utility score. Secondly, in the present study, equal weights were used to combine lower level nodes into higher level nodes because we had no basis for doing otherwise. However, the data collectors may have been mentally weighting the relative importance of the different attributes (and/or criteria and dimensions) making up their global overall utility judgements. If they were, then their global utility scores would be different from the calculated hierarchical utility scores.

The difference between the global utility score and the calculated hierarchical utility score lends credibility to the use of MAUA to determine overall utility of decision aiding systems. The MAUA approach is based on the principle of decomposition. In MAUA, the overall judgement is decomposed into its elements so that judgements about the elements can be made separately. Pitz and McKillip describe the decomposition principle (1984, p. 76). "The decomposition principle asserts that judgements are indeed more reliable, more consistent with each other and less subject to bias and error when the event being judged is characterized by fewer features. The principle implies that there will be less systematic and random error when the judgements are concerned with simple, unidimensional components of the problem." A number of researchers have found evidence for the decomposition principle (Gardiner, 1977; Gardiner & Edwards, 1975; Pitz, 1980). If a global assessment of utility would yield the same, or even better values, than utility based on a MAUA analysis and data collection, there would be no point in spending the time and effort to do a MAUA analysis. In this study the two measures of overall utility did not yield the same scores and the decomposition principle argues that the calculated MAUA score is a better measure.

Table 4 also presents the maximum and minimum values for each of utility dimensions, criteria, and attributes. The maximum and minimum values for an attribute were the highest and lowest ratings, respectively, given to the question(s) measuring that attribute for any prototype by any data collector. When multiple questions were used to measure an attribute, we calculated the mean value for those questions separately for each data collector for each prototype.

The maximum and minimum values for the criteria and dimensions were the highest and lowest cumulative scores, respectively, given by any data collector for any prototype as one moved up the hierarchy. Consequently, the maximum and minimum values given in Table 4 for a criterion are not necessarily the average of the maximum and minimum values, respectively, given to the attributes comprising it. For example, the minimum value for <u>product quality</u> (criterion 1.2 in Table 4) is 3.5; yet the minimum values for quality of results (attribute 1.2.1) and overall confidence (attribute 1.2.2) are both 2.0. This occurred because one data collector rated one prototype lowest on quality of results, and another data collector rated another prototype lowest on overall confidence. The minimum rating of 3.5 for the <u>product quality</u> criterion represents the lowest mean score given by any data collector, for any prototype they evaluated, for both attributes. It is not the mean of the minimum values shown in Table 4 for these two attributes.

In three cases, the maximum value shown in Table 4 for a criterion is higher than the average of the maximum values for the attributes. For example, the maximum value shown in Table 4 for <u>process quality</u> (criterion 1.1) is 7.0; yet, the maximum value for "quality of information" (attribute 1.1.1) was only 6.0. This occurred because one data collector did not answer the question measuring quality of the information for one prototype. We calculated his score for the <u>process quality</u> criterion for that prototype by averaging his scores for the other four attributes comprising that criterion. His mean score was 7.0 for the dimension because he gave a 7 to all the questions measuring the other process quality attributes for that prototype. In contrast, 6.0 was the highest rating for the question measuring quality of information given by any of the data collectors who answered it.

The same logic holds for the maximum and minimum values shown in Table 4 for the dimensions and Overall Utility. The maximum and minimum values were the highest and lowest values, respectively, given to the dimensions for any prototype by any data collector.

It is important to emphasize that the wide range between the minimum and maximum values shown in Table 4 indicates that the data collectors were able to use the entire response scale to discriminate between good and poor prototypes. For example, the **Overall Utility** score had a range of 4.15, going from a minimum of 2.85 to a maximum of 7.0. The scores for the three utility dimensions ranged from 3.33 to 7.00 for **Effect on Task Performance**, 2.33 to 7.00 for **System Usability**, and 2.88 to 7.00 for **System Fit**. The highest score for many bottom-level nodes (15 of the 19 utility attributes) was 7.00, while the lowest score for many of the same nodes (18 of the 19 attributes) was 2.00 or less. [Note: The lowest rating score was 1.00 for time to perform task (attribute 1.1.5 in Table 4) and flexibility (criterion 2.6).]

Table 5 shows the means, standard deviations, and sample sizes for the overall utility scores of all twelve decision aids. The mean overall utility scores range from a low of 3.76 to a high of 6.17. This wide range for the different aids suggests the questionnaire is a sensitive instrument that can distinguish between aids that the participants liked and disliked. This large range of means demonstrates how flexible the utility hierarchy is in differentiating between good and poor decision aids and that it may be possible to use the questionnaire to discriminate between prototypes that need considerable improvement, and perhaps even for which development should stop, and prototypes that should be developed immediately.

We have described the application of the questionnaire to eleven decision aiding systems. The results of this application show that the questionnaire can be easily tailored to multiple and diverse systems, that respondents use the whole range of the questionnaire scale, and that the questionnaire can distinguish between aids liked and those they disliked. We now turn to the psychometric analysis of the questionnaire data.

Psychometric Analysis

This section presents the psychometric analysis of the five data collectors' responses. We note at the outset that five participants is too small a sample size for accurately assessing the psychometric characteristics of a questionnaire. For example, previous research by Adelman et al. (1985) used 29 participants to assess the psychometric characteristics for a questionnaire, and even this was considered to be a small sample size. However, it should be remembered that each of the eleven prototypes was evaluated, on the average, by three data

Table 5

Overall Utility Score Mean, Standard Deviation (SD), and Sample Size (N) for Eleven Decision Support Systems (DSS)

DSS	Mean	SD	N
1	5.30	.36	3
2	3.76	.78	3
3	5.25	1.02	5
4	4.26	.75	2
5	4.99	--	1
6	5.72	1.07	5
7	4.92	--	1
8	4.30	.33	3
9	3.85	.27	2
10	5.15	.95	3
11	6.27	.76	5

collectors. This increased the sample size to 30 evaluations. This number provided sufficient power for assessing the statistical significance of the obtained results. In fact, the statistically significant results presented below are particularly encouraging given the small sample size. Nevertheless, future applications of the questionnaire should continue to assess its psychometric properties and ways to improve it.

We present the results of two analyses, one in each of two subsections. The first subsection presents the results for the split-half reliability analysis. The second subsection assesses the construct validity of the questionnaire; that is, the extent to which the results of the questionnaire correlate with the results of other questions or approaches to supposedly measuring the same utility attributes.

Split-Half Reliability. As was noted above, each half of the questionnaire had at least one question measuring each of the ten nodes selected for measuring the split-half reliability. If the questionnaire is a reliable measuring instrument, then there should be a high correlation between the two halves of the questionnaire measuring these ten nodes. The split-half reliability correlation coefficient would indicate the extent to which this is, in fact, the case.

Two procedures were used to obtain the split-half reliability. The first used the traditional four-step procedure for calculating split-half reliability. First, we identified the items for the ten (10) nodes in each half of the questionnaire. One of the nodes, criterion #2.2, had three questions; all other nodes had two questions. For criterion 2.2, we correlated each of the first questions measuring that criterion with the third question measuring it. This gave us eleven pairs of questions for calculating the split-half reliability coefficient.

(Note: An alternative approach would have been to pair the average score to the first two questions with the score for the third question. We rejected the latter approach because averaging would have reduced the amount of variation between the questions and, given the relatively small number of comparisons, possibly inflated the split-half reliability correlation.)

The second step in the first procedure was listing the ratings for each data collector for each of the prototypes. These ratings were then standardized for each of the data collectors prior to combining the data. This step was done to control for systematic differences in how the data collectors used the rating scale. For example, some raters may never have rated any of the prototypes more than a five, while others' top score may have been seven.

Third, we calculated Pearson product-moment correlations. We correlated the eleven (11) pairs of standardized scores, for the two halves of the questionnaire, for each of the eleven (11) prototypes for all of the data collectors. This led to a total of 263 comparisons used in the split-half reliability calculation. The reason the number of comparisons is smaller than would be expected with five data collectors (11 aids x 11 comparisons x 5 data collectors = 605) is that not all data collectors evaluated all aids, and not all items were completed for those aids that were evaluated. The Pearson product-moment correlation was $r_a = 0.61$. This correlation is significantly different than zero at the $p < 0.01$ level ($\underline{t}$ (262) = 12.55).

Fourth, we used the formula below (from Cascio, 1991) to calculate the more traditional split-half reliability correlation coefficient.

$$r_e = 2r_a/(1 + r_a) \qquad [1]$$

The resulting split-half reliability was $r_e = 0.76$. This measure is above the traditional minimum level of 0.70.

Because the initial test of split-half reliability may have been biased by the large sample size (263 comparisons), a second more conservative test of split-half reliability was conducted. It must be remembered that each data collector evaluated more than one prototype. In the more conservative test, we first calculated each data collector's standardized mean score, over the prototypes they evaluated, for each question used in the split-half reliability analysis. This produced eleven standardized mean scores for each half of the questionnaire (i.e., one for each of the eleven comparisons) for each of the five data collectors. Overall, we correlated 53 data points (11 mean scores for each half of the questionnaire multiplied by 5 data collectors - 2 for missing data). The Pearson product-moment correlation was $r_a = 0.64$. This correlation is significantly different than zero at the $p < 0.01$ level ($\underline{t}$ (52) = 4.89). To obtain the split-half reliability coefficient, formula 1 was applied to this correlation. This calculation resulted in a split-half reliability of $r_e = 0.78$, also greater than the traditional minimum level of .70.

The results of these two calculations are encouraging in that they yield very similar results despite large differences in sample sizes (263 versus 53 comparisons). This study used a subset of the questionnaire because not all the items were applicable to the prototypes in their current state of development. However, the split-half reliability results are also similar to those found with the entire 96 item questionnaire in the pre-test (Adelman et al., 1993). In the pretest analysis, the split-half reliability for the entire 96 item questionnaire was $r_e = 0.65$. In the current study, the r_e values were 0.76 and 0.78, for the two methods of calculating the split half-reliability.

Construct Validity. By "validity" we mean that the instrument is measuring what it is supposed to measure. An instrument can be reliable (i.e., it produces the same results upon replication), but invalid (i.e., it reliably measures the wrong thing). Since the previous section suggests that the questionnaire is reliable, we now turn to consider its validity.

There are three different types of validity: content (or face) validity, predictive validity, and construct validity. As was noted in the section describing the questionnaire, we tried to ensure that the questionnaire had content validity by having both Army officers and ARI psychologists, all of whom had experience developing questionnaires for use by Army personnel, critique the questionnaire's content. As we will consider in the discussion section, future applications of the questionnaire need to assess its predictive validity; that is, its ability to predict respondents' actual performance behavior in a test setting. In this section, we consider the construct validity of the questionnaire; that is, the extent to which the results of the questionnaire correlate with the results of other questions or approaches supposedly measuring the same utility attributes.

In order to assess construct validity of the questionnaire, respondents answered six sets of questions in addition to the questionnaire items. Appendix C shows the construct validity questions used in these analyses. Six analyses were conducted. First, we correlated (a) the data collectors' mean responses to the two questions directly asking about the decision aids' overall utility with (b) their **Overall Utility** score based on the MAUA utility hierarchy (i.e., node 0.0). This procedure resulted in 33 comparisons (e.g., eleven decision aids with an average of three raters per aid). The correlation was $r = 0.60$. This correlation is significantly different from zero at the $p < 0.01$ level ($\underline{t}$ (32) = 4.16).

This statistically significant correlation means that higher responses to the global questions were related to higher scores on the **Overall Utility** score. Similarly, lower scores on the questions were related to low **Overall Utility** scores. Although the questions had a higher mean value than the **Overall Utility** score, as was discussed in the previous section, there was still a significant relationship between the two measures. This suggests that, overall, the questionnaire has construct validity.

For the second comparison, we correlated (a) the data collectors' **Overall Utility** scores with (b) their ratings for a question asking about the "extent to which the decision aid met their needs." The latter question was added to the questionnaire for construct validity purposes. The "needs" rating was on a five-point scale going from "Not At All" to "Very Much." The correlation was 0.35, which was significantly different from zero at the $p < 0.01$ level ($\underline{t}$ (32) = 5.76). This provides more evidence that the **Overall Utility** node is measuring a global utility construct.

For the third comparison, we correlated (a) the data collectors' **Overall Utility** scores with (b) their mean ratings for a question asking about the extent to which the aid would improve their performance on each of twenty-nine (29) separate tasks. The latter question was not developed for construct validation purposes, but we realized that it could be used for that purpose when doing the analysis. An example item from this second measure would be: "This system will help to improve my performance on course of action (COA) analysis." A five point scale was used to assess the extent to which the decision aid improved performance on the tasks. The correlation was 0.61. Again, this correlation was significantly different from zero at $p < 0.01$ ($\underline{t}$ (32) = 4.26).

This result is particularly encouraging because the questions in the two questionnaires were at very different levels of specificity. Moreover, we did not consider the relative importance of the 29 separate tasks when we calculated the mean ratings. Since different decision aids support different tasks to various degrees, it can be argued that each data collector should have differentially weighted the 29 tasks for each prototype to reflect the relative importance of the tasks being supported by it. The weights were not requested because the question was not initially developed for construct validity purposes. Consequently, the correlation of 0.61 probably reflects a conservative estimate of the degree of relation between the two measures.

The fourth comparison examined the relationship between (a) the five data collectors' ratings for the quality of results criterion (node 1.2.1) with (b) their rating of the extent to which the decision aid would "improve the quality of their work." The latter question was added for construct validation purposes. These two values were found to be correlated with each other. The correlation was 0.39, which was significantly different from zero at the $p < 0.01$ level (t (32) = 4.24). This correlation suggests that both measures were assessing the quality of the results produced by the prototypes, and again provides support for the construct validity of the questionnaire.

The fifth comparison correlated (a) the data collectors' organizational fit score (attribute 3.2.2) with (b) their answer to the question, "To include this decision aid in staff operations would be." The response was again on a 5-point scale going from Very Difficult to Very Easy. The correlation was 0.43. Again, this was significantly different from zero at the $p < 0.01$ level (t (32) = 9.67). This correlation suggests that both measures were assessing the organizational fit of the prototypes, and again provides support for the questionnaire' construct validity.

The final construct validation comparison was between a) the data collectors' rating for the time to perform task (attribute 1.1.5) and (b) their rating of the extent to which the decision aid would be "flexible in meeting varying task and time demands." The latter question was rated on a five point scale from Very Unsatisfactory to Very Satisfactory. Unlike the other comparisons this correlation was not significant. There are two possible reasons for the poor correlation, which was only 0.11. First, the comparison question occurred at the end of the questionnaire. The data collectors may have been tired and not accurate in their judgments. Second, although both sets of questions addressed aspects of time, the attribute in the utility hierarchy focuses on how a decision aid improves the speed with which a task is completed, while the comparison item focuses on the aids' ability for the user to flexibly make use of their time. In retrospect, these seem like two different constructs. Nevertheless, the low correlation indicates a construct validation failure of the two questions measuring the time to perform task attribute.

Two types of psychometric analyses, split half-reliability and construct validation, were described in this section. The results of these analyses suggest that the questionnaire is valid and reliable.

Summary and Conclusions

This paper described the development and validation of a questionnaire for obtaining users' opinions about the utility of decision aiding systems, including knowledge-based systems. The questionnaire was designed to be an off-the-shelf tool that evaluators could use to obtain users' opinions of an aid's utility throughout the development process. It was designed to be quickly and easily tailored to different decision aids at different stages of development.

Development of the questionnaire began with a literature review to identify the criteria used by different researchers to assess system utility and usability. The identified criteria then were organized into a multi-attributed hierarchy with the top three dimensions being **Effect on Task Performance**, **System Usability**, and **System Fit**. These three dimensions were decomposed into lower-level utility criteria and attributes. Questions were developed to assess each of the bottom and top level attributes. Multi-Attribute Utility Assessment (MAUA) concepts were used to combine the answers to these questions into utility measures for each criterion, dimension, and the system overall. The construct validity of the questionnaire is supported by the method by which the questionnaire was constructed, i.e. identification of utility dimensions used by other researchers and integtration of these dimensions into a hierarchy using MAUA concepts.

Data for a psychometric analysis of the questionnaire were collected at the Army's Battle Command and Battle Laboratory (BCBL). Five data collectors used the questionnaire to evaluate eleven different decision aid prototypes. First, BCBL personnel identified a subset of utility criteria and attributes of critical concern to them, and then the validation team developed a short version of the questionnaire that both measured these attributes and could be administered in 10 to 15 minutes.

The results of the validation effort are encouraging. First, the resulting questionnaire was capable of distinguishing between those prototypes the soldiers liked and those that they didn't. The mean overall utility scores for the eleven aids ranged from a low of 3.76 to a high of 6.27, on a scale of 1 to 7. In addition, the results showed that data collectors were able to use the entire range of the questionnaire scale when evaluating the aids. Specifically, when examining the minimum and maximum values given by individual data collectors for individual aids, all levels of the hierarchy had a large range of responses. In particular, this range went the entire length of the response scale (i.e., from 1 to 7) for one attribute and for one criterion, and most of the scale (i.e., from 2 to 7) for all the other attributes and criteria. The scores for the three utility dimensions ranged from 3.33 to 7.00 for **Effect on Task Performance**, 2.33 to 7.00 for **System Usability**, and 2.88 to 7.00 for **System Fit**. And the **Overall Utility** score went from a minimum of 2.85 to a maximum of 7.0.

Second, the high split-half reliability measures were high. Two different procedures were used to assess the split-half reliability of the questionnaire, one being more conservative than the other. The resulting reliability measures were r_e = 0.76 and r_e = 0.78. Both measures are above the traditional minimum level of 0.70. Moreover, the r_a correlations upon which they were based were significantly greater than zero at the $p < 0.01$ level with 262 and 32 degrees of freedom, respectively. These split-half reliability coefficients are comparable to, and higher than, the split-half reliability coefficient of r_e = 0.65 obtained for the entire 96-item questionnaire in the pretest.

Third, the construct validity results were encouraging. Five of the six construct validity correlations ranged from 0.35 to 0.61, and were statistically significant at the $p < 0.01$ level. The correlations were comparable to the construct validity correlations obtained with the 96-item questionnaire in the pretest. Of particular importance, all three comparisons assessing the questionnaire's **Overall Utility** score were statistically significant. The only construct validity correlation that was not significant was for time to perform task, and that was due to either a poor match in the question wording or the fact that the construct validity question did not come until late in the evaluation.

It is important to re-emphasize that considerable efforts were made to ensure that the questions had content validity. The questionnaire was developed by first surveying other researchers and evaluators to determine what dimensions and items they included in their questionnaires. This helped ensure that the set of items in the questionnaire represented the range of factors that users consider when evaluating system utility. In addition, the first and third authors have considerable experience working with Army personnel and evaluating military decision aids. And, finally, one active-duty Army officer and one retired Army officer participated in the pilot test, as did other personnel at ARI who have experience working with Army personnel. This helped to ensure that the questions used the right phrases and jargon and, basically, sounded right to the respondents.

Although further data collection and analysis is required, the above results suggest that the questionnaire is a reliable and valid measurement instrument. The current validation effort did not, however, assess the questionnaire's predictive validity. That is, there was no attempt to correlate the questionnaire's results with the data collectors' actual performance behavior with the prototypes. This decision was made because the prototypes were primarily for demonstration purposes within the context of a military exercise, with limited time available for their use. Nevertheless, the lack of predictive validity data represents a limitation of the current validation effort. It needs to be rectified by future research.

It is hoped that others will utilize the questionnaire to expand the sample size for the psychometric analyses. To achieve this goal, data collection, preferably using the entire 96 item questionnaire, will have to be expanded to additional aids and participants. Although it is anticipated that the questionnaire will be used to evaluate other Army prototypes, this does not prohibit the questionnaire from also being used in other domains. Indeed, by building the questions around general utility dimensions, criteria and attributes, we have tried to develop a questionnaire that can be easily modified to measure a wide array of decision aiding systems. It is our contention that this set of items could be used by developers, both inside and outside of the military, to control development costs and to help evaluate and select decision aids and, perhaps, other types of computer systems.

Additional research is also needed to determine whether the many different attributes do, in fact, collapse into the three general utility dimensions subjectively defined in this study. This question could be explored, for example using factor analysis, as part of future data collection efforts with the questionnaire. Factor analysis would require a much larger sample size (e.g., 100 to 200 participants) than we were able to obtain. However, empirically determining the major dimensional constructs of system utility would make an important contribution to understanding how users evaluate systems, particularly early in the development life cycle.

Because a relatively large number of decision aiding prototypes were evaluated in this study (i.e., eleven), the mean values presented in Table 4 might be used as benchmarks for the evaluation of future systems. These benchmarks, along with the range data, could be utilized to help select which prototypes should continue to be developed, and which ones should have development halted or rethought. By supporting these types of decisions, the questionnaire provides a tool to help control development costs, allowing the sponsoring organization to make a decision on a system's implementation potential prior to actually having to build the full system. And even for prototypes that score reasonably well overall, the questionnaire identifies areas (i.e., attributes and criteria) where additional work is required in the opinion of potential users. This is critical feedback for the development team.

The mean values for each of the attributes and criteria do, however, have to be used with caution for two reasons. First, additional applications of the questionnaire would provide a larger sample size upon which to base any benchmarks. Second, the mean values were obtained by using equal weights throughout the hierarchy. However, respondents using the questionnaire may consider certain attributes, criteria, and dimensions to be differentially important. As discussed earlier in the paper, this may account for why the mean value for the global overall utility judgments was higher than the mean score for the **Overall Utility** node in the hierarchy. The MAUA approach upon which the questionnaire is based permits users to differentially weight the attributes, criteria, and dimensions in the hierarchy. Indeed, this is one of the strengths of MAUA. However, use of differential weights will affect the benchmark values for higher level nodes in the hierarchy. Consequently, the benchmark means presented in Table 4 need to be used with caution.

The goal of this study was to develop an "off-the-shelf" questionnaire, measuring decision aid utility, which can be quickly and easily tailored for different systems at different stages of development. Strong psychometric support for the questionnaire was provided by the high reliability and validity measures for data collected over a range of systems. The method of constructing the questionnaire lends support to its content validity. In addition, the application of the questionnaire to eleven decision aids demonstrated the ease with which the questionnaire can be adapted for use with different systems. The availability of this ready made, easily adaptable, psychometrically valid, user questionnaire makes it possible for evaluators to routinely obtain user feed-back throughout decision aid development.

References

Adelman, L. (1992). Evaluating decision support and expert systems. New York: Wiley-Interscience.

Adelman, L., Gualtieri, J., & Riedel, S.L. (in press). A multi-faceted approach to evaluating expert systems. Artificial Intelligence in Engineering Design, Analysis and Manufacturing.

Adelman, L., Gualtieri, J., & Riedel, S.L. (1993). Questionnaire measuring the usability of knowledge-based systems (Working Paper). Fort Leavenworth, KS: U.S. Army Research Institute for the Behavioral and Social Sciences.

Adelman, L., Rook, F.W., & Lehner, P.E. (1985). User and R&D specialist evaluation of decision support systems: Development of a questionnaire and empirical results. IEEE Transactions on Systems, Man, and Cybernetics, SMC-15, 334-342.

Bennett, J.L. (1984). Managing to meet usability requirements: Establishing and meeting software development goals. In Bennett et al. (Eds.), Usability issues and health concerns. Englewood Cliffs, NJ: Prentice Hall.

Berry, D.C., & Hart, A.E. (1990). Evaluating expert systems. Expert Systems, 7, 199-207.

Berry, D.C., & Hart, A.E. (1991). User interface standards for expert systems: Are they appropriate. Expert Systems With Applications, 2, 245-250.

Cascio, W.F. (1991). Applied psychology in personnel management. Englewood Cliffs, NJ: Prentice Hall.

Cleal, D., & Heaton, N. (1988). Knowledge-based systems: Implications for human computer interfaces. Chichester, UK: Ellis Horwood.

Clegg, C.W., Warr, P., Green, T., Monk, A., Kemp, N., Allison, G., & Lansdale, M. (1988). People and computers: How to evaluate your company's new technology. Chichester, UK: Ellis Horwood.

Edwards, W. (1977). How to use multiattribute utility measurement for social decision making. IEEE Transactions on Systems, Man, and Cybernetics, SMC-7, 326-340.

Gardiner, P.C. (1977). Multiattribute utility measurement: Public values for public decision making. Department of Psychology, University of Southern California.

Gardiner, P.C., & Edwards, W. (1975). Public values: Multiattribute utility measurement of social decision making. In M.F. Kaplan & S. Schwartz (Eds.), Human judgement and decision making. New York: Academic Press.

Gray, T., Roberts-Gray, C., & Gray, W.D. (1983). A guide to implementation of training products (ARI Research Report 1350). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 143 669)

Hammond, N.V., Morton, J., Barnard, P.J., Long, J.B., & Clark, I.A. (1987). Characterizing user performance in command-driven dialogue. Behavior & Information Technology, 6, 237-254.

Hockey, R., Briner, R., Tattersall, A., & Wiethoff, M. (1989). Assessing the impact of computer workload on stress: The role of system controllability. Ergonomics, 32, 1401-1418.

Hogarth, R. (1987). Judgement and choice. New York: Wiley.

Holcomb, R., & Tharp, A.L. (1991). What users say about usability. International Journal of Human-Computer Interaction, 3, 49-78.

Huber, G.P. (1980). Managerial decision making. Glenview, IL: Scott, Foresman.

A

Marshall, C., Nelson, C., & Gardiner, M. (1987). Design guidelines. In M. Gardiner & B. Christie (Eds.), Applying cognitive psychology to user interface design. Chichester, UK: Wiley.

Mitta, D.A. (1991). A methodology for quantifying expert system usability. Human Factors, 33, 233-245.

Morris, C. (1987). The effect of user characteristics on system development. In W. Rouse & K. Boff (Eds.), System design: Behavioral perspectives on designers, tools, and organizations. New York: North-Holland.

Nielsen, J. (1993). Usability engineering. London, UK: Academic Press.

Pitz, G.F., & McKillip, J. (1984). Decision analysis for program evaluators. Beverly Hills, CA: Sage.

Ravden, S.L., & Johnson, G.I. (1989). Evaluating usability of human computer interfaces: A practical method. Chichester, UK: Ellis Horwood.

Riedel, S.L. (1992). ALBM-ATTD master plan for life cycle operational evaluations (Planning document). Fort Leavenworth, KS: U.S. Army Research Institute for the Behavioral and Social Sciences.

Riedel, S.L., & Pitz, G.F. (1986). Utilization-oriented evaluation of decision support systems, IEEE Transactions on Systems, Man, and Cybernetics, SMC-16, 980-996.

Shakel, B. (1986). Ergonomics in design for usability. In M. Harrison & A. Monk (Eds.), People & computers: Designing for usability. Proceedings of the second conference of the BCS HCI specialist group. Cambridge, UK: Cambridge University Press.

Shlechter, T.M., Bessemer, D.W., Rowatt, W.C., & Nesselroade, K.P. (1994). Evaluating the unit performance assessment system's after action review displays (Technical Report 997). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 281 712)

Shlechter, T.M., Burnside, B.L., & Thomas, D.A. (1987). Issues in developing and implementing computer-based instruction for military training (Research Report 1451). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA 189 479)

Susskind, K. (1988). Cognitive factors in human-computer interaction. In D. Cleal & O. Heaton (Eds.), Knowledge-based systems: Implications for human computer interfaces. Chichester, UK: Ellis Horwood.

Sweeney, M., Maguire, M., & Shackel, B. (1993). Evaluating user-computer interaction. International Journal of Man-Machine Studies, 38, 689-711.

APPENDIX A

UTILITY HIERARCHY AND QUESTIONS

Note: The question numbers noted in this section refer back to the complete ninety-six (96) item questionnaire used in the pretest to evaluate the AA Comparator KBS. The question numbers for the items used in the current study can be found in Table 4 and a sample instrument in Appendix B.

0.0 Overall System Utility

[Questions #10,24,33,51,61,86]

1.0 Effect on Task Performance

1.1 Process Quality

1.1.1 Quality of the Information (Data & Knowledge)

[Questions #14,36,44,65,83]

1.1.2 Quality of the Explanation Capability/Reasons

[Questions #19, 49, 90]

1.1.3 Quality of the Representation, Examination, and Modification of Knowledge Stored in System

[Questions #32,82]

1.1.4 Quality of the Problem Solving Approach

[Questions #11,42,58,85]

1.1.5 Time to Perform Task(s)

[Questions #18,75, 81]

1.1.6 System Response Time

[Questions #12,73]

1.2 Product Quality

1.2.1 Quality of the Results (i.e., products)

[Questions # 17,34,59,76]

1.2.2 Overall Confidence

[Questions #37,46,52]

1.2.3 Cost-Effectiveness

[Questions #5, 80]

2.0 System Usability

2.1 General Ease of Use Questions

[Questions #7,15,35,55,68,71]

2.2 Quality of the User's Mental Model of the System

[Questions #9,20,38,56,62,67,95]

2.3 Ease of Training

2.3.1 Learnability

[Questions #26,53,77]

2.3.2 On-Line Help Function

[Questions #3,70]

2.3.3 Documentation - Not used for AA Comparator, but the standard questions are presented in appendix

2.4 Human Factors Guidelines for Person-Machine Interaction

2.4.1 Consistency

[Questions #25,48,63]

2.4.2 Error Prevention and Handling

[Questions #16,40,64]

2.4.3 System Navigation

[Questions #41,54,72]

2.4.4 Use of Color

[Questions #29,78]

2.4.5 Ease of Data Entry

[Questions #30,50]

2.4.6 Feedback

[Questions #23,60]

2.4.7 Reliability - Not used for AA Comparator, but the standard questions are presented in appendix

2.5 Workload

2.5.1 Strain (i.e., effort)

[Questions #6,45,74]

2.5.2 Memorization

[Questions #28,92]

2.6 Flexibility (Discretion and User Control)

[Questions #21,69, 96]

2.7 Functional Allocation Between Person and Machine

[Questions #22,88,94]

3.0 System Fit (i.e., How Well the System Fits In)

3.1 Fit (i.e., Match) With User

3.1.1 Match With Users' Needs

[Questions #43,47,79]

3.1.2 Match With Users' Training

[Questions #8,84]

3.1.3 Match With Users' Problem-Solving Approach

[Questions #31,39,91]

3.2 Fit (i.e., Match) With Organization

3.2.1 Match With Doctrine

[Questions #13,57]

3.2.2 Organizational Fit

[Questions #1,87]

3.2.3 Effect on Information Flow

[Questions #2,89]

3.2.4 Effect on Other People's Workload

[Questions #27,93]

3.2.5 Attitude of Others (Political Acceptability)

[Questions #4,66]

Standardized Questions

Note: The information within brackets [] indicates how the evaluator needs to tailor the question.

0.0 Overall System Utility

10. I think an operational version of the [System Name] is good enough to use in a major training exercise.

24. The [System Name] is a valuable tool for [purpose of system].

33. Overall, the [System Name] is a useful approach for [purpose of system].

51. Use of the [System Name] will improve [purpose of system] performance.

61. I recommend continued development of the [System Name] for operational use.

86. Overall, the [System Name] improves [purpose of system].

1.0 Effect on Task Performance

1.1 Process Quality

1.1.1 Quality of the Information (Data & Knowledge)

14. The [System Name] is using the right data for [purpose of system].

36. I agree with the [use the word "knowledge" or identify a type of knowledge stored in the knowledge base] stored in the [System Name] for [purpose of system].

44. I agree with the [identify a second type of knowledge stored in the knowledge base] stored in the [System Name] for [purpose of system].

Note: If the above version of question #44 is inappropriate because one can not easily distinguish between the different types of knowledge stored in the knowledge base, use the following version of question #44.

44. The [System Name] contains the right knowledge for [purpose of system].

65. I agree with the [identify a third type of knowledge stored in the knowledge base] stored in the [System Name] for [purpose of system].

Note: If the above version of question #65 is inappropriate because one can not easily distinguish between the different types of knowledge stored in the knowledge base, use the following version of question #65.

65. The [System Name] contains an adequate level of expertise to support users performing [purpose of system].

83. The [System Name] uses the correct information in producing its results.

1.1.2 Quality of the Explanation Capability/Reasons

19. Overall, the reasoning underlying the results is acceptable.

49. The [System Name] provided good reasons for its results.

90. It is easy to interpret the results of the [System Name].

1.1.3 Quality of the Representation, Examination, and Modification of Knowledge Stored in System

32. In general, it is easy to modify the knowledge stored in the [System Name].

82. The [System Name] allows users to examine the expert judgments on which the system's recommendation is based.

1.1.4 Quality of the Problem Solving Approach

11. The [System Name's] approach to representing expert knowledge for [purpose of system] is acceptable.

42. The [System Name] uses a logically sound approach for [purpose of system].

58. The [System Name's] approach to [purpose of system] is acceptable.

85. The calculations [or analysis] performed by the [System Name] were helpful.

1.1.5 Time to Perform Task(s)

18. Using the [System Name] to [purpose of system] was fast enough for my needs.

75. I would feel comfortable using the [System Name] under time pressure.

81. Completing the task with the [System Name] is faster than current procedures.

1.1.6 System Response Time

12. The [System Name] responds quickly to the user's commands.

73. The [System Name's] response time is acceptable.

1.2 Product Quality

1.2.1 Quality of the Results (i.e., products)

17. Overall, the [System Name] provided me with useful results.

34. I found the [System Name's] results acceptable.

59. The [System Name] supports the preparation of high quality products.

76. The [System Name] would improve the quality of my work.

1.2.2 Overall Confidence

37. I have alot of confidence in the results obtained working with the [System Name].

46. I am confident that the [System Name] is well-built technically.

52. I have alot of confidence in the [System Name's] approach to [purpose of system].

1.2.3 Cost-Effectiveness

5. The [System Name] is cost-effective because the benefit of using it is worth the effort.

80. The [System Name] provides users alot of value for their efforts.

2.0 System Usability

2.1 General Ease of Use Questions

7. The displays are easy to read.

15. The displays are easy to understand.

35. It was easy to tell the [System Name] what to do.

55. The [System Name] is easy to use.

68. The [System Name's] input screens are easy to use.

71. The mouse and keyboard are easy to use.
[Note: The wording of this question depends on the type of input devices that the system uses.]

2.2 Quality of the User's Mental Model of the System

9. It was easy to form a mental picture of how the [System Name] works.

20. It was easy to understand why the results came out the way they did.

38. The organization of menu items is easy to understand.

56. It is clear what to do to get the [System Name] to perform the actions one wants.

62. The labels on the menu choices correctly describe the choice.

67. I understand how to use the [System Name] to do [purpose of system].

95. The system contains familiar terms.

2.3 Ease of Training

2.3.1 Learnability

26. The [System Name] requires no retraining for infrequent users.

53. One can learn to use the [System Name] in one two-hour training session.

77. The [System Name] was easy to learn.

2.3.2 On-Line Help Function

3. The [System Name] has sufficient help features.

70. The [System Name's] help features are easy to use.

2.3.3 Documentation - Not used for AA Comparator, but the standard questions are presented below.

- How to use [System Name] is well documented.

- The [System Name's] User's Manual is easy to understand.

2.4 Human Factors Guidelines for Person-Machine Interaction

2.4.1 Consistency

25. The [System Name] uses the same layout for all screens.

48. The [System Name] presents similar information at the same place on the screen.

63. The same commands produce the same actions throughout the [System Name].

2.4.2 Error Prevention and Handling

16. The [System Name] helps to prevent errors the user might make when using it.

40. The [System Name] provides immediate error notification.

64. The [System Name] is designed so that it is easy to recover from errors, if they should occur when using it.

2.4.3 System Navigation

41. It is always clear where the user is in the [System Name].

54. The user can easily move form one menu item to another without errors in the [System Name].

72. The user can easily move to different parts of the [System Name] as required to do the tasks.

2.4.4 Use of Color

29. The [System Name] uses color in an intuitive way.

78. I understand the meaning of the different colors used in the displays.

2.4.5 Ease of Data Entry

30. I can easily supply the information the [System Name] asks me for.

50. It is easy to enter data into the [Sys. Name].

2.4.6 Feedback

23. The [System Name] provides feedback when it's processing user commands.

60. The [System Name] provides the user with effective directions so that one always knows what to do next.

2.4.7 Reliability - Not used for AA Comparator, but the standard form of the questions is presented below.

- The number of system failures is acceptable.
- The level of down time is acceptable.
- The same inputs produce the same results.

2.5 Workload

2.5.1 Strain (i.e., effort)

6. The user does not have to exert much mental effort to use the [System Name] to compare avenues of approach.

45. The [System Name] reduces the amount of work required to compare avenues of approach.

74. The amount of effort required to use the [System Name] is acceptable.

2.5.2 Memorization

28. The user does not have to memorize commands to use the [System Name].

92. All necessary information is available on each screen.

2.6 Flexibility (Discretion and User Control)

21. I felt in control of the [System Name] when it was operated.

69. The system allows for adaptation to different scenarios.

96. The [System Name] permits the user to control the order in which different activities are done.

2.7 Functional Allocation Between Person and Machine

22. The [System Name] supports those tasks requiring support when [purpose of system].

88. The [System Name] is designed so that the right activities are allocated to the person and machine.

94. The [System Name] provides me with the right kind of support for [purpose of system].

3.0 How Well the System Fits In

3.1 Fit (i.e., Match) With User

3.1.1 Match With Users' Needs

43. The [System Name] meets my needs for [purpose of system].

47. The [System Name's] products meet my needs.

79. It is easier to [purpose of system] using the [System Name] than with my current procedures.

3.1.2 Match With Users' Training

8. The [System Name] is designed to match the computer skills of Army personnel who would use it.

84. The system's approach to comparing avenues of approach matches how I was trained to perform this task.

3.1.3 Match With Users' Problem-Solving Approach

31. The [System Name] performs [purpose of system] the way I do.

39. The [System Name's] approach to [purpose of system] matches my idea of how this task should be done.

91. In general, the [System Name] uses the same information that I use.

3.2 Fit (i.e., Match) With Organization

3.2.1 Match With Doctrine

13. The procedures used in the [System Name] are consistent with Army doctrine.

57. The procedures used in [System Name] follow Army doctrine.

3.2.2 Organizational Fit

1. The [System Name] fits well in the [organizational place for the system].

87. From a [organizational place for the system] perspective, the [System Name] is a good fit.

3.2.3 Effect on Information Flow

2. The [System Name] will facilitate the flow of information in the [organizational place for the system].

89. The [System Name] will not interfere with the flow of information in the [organizational place for the system].

3.2.4 Effect on Other People's Workload

27. The [System Name] would not increase the amount of work for other people involved in [purpose of system].

93. The [System Name] will decrease the workload of other people in the [organizational place for the system].

3.2.5 Attitude of Others (Political Acceptability)

4. Other people in the [organizational place for the system] will support the [System Name's] implementation.

66. My superiors would strongly favor the using the [System Name].

APPENDIX B

QUESTIONNAIRE FOR SAMPLE DECISION AID

MapInfo
(MapInfo Desktop Mapping Decision Aid)

MapInfo is a desktop mapping tool and decision aid which supports the display of maps and overlay graphics. MapInfo is a geographic information system (GIS) which can display Battle Command data by location. MapInfo takes user data and displays it in the correct location with appropriate colors and symbols. Data may be viewed in three ways: text, map and graphic. Data behind any object on a map may be queried and the distance between points may be calculated.

The purpose of this questionnaire is to obtain your opinion of MapInfo. You will be given a number of statements about MapInfo. Each statement will be followed by a scale and a section for comments. On the opinion scale, please indicate the extent to which you agree or disagree with each statement by circling the appropriate number on the 7 point scale.

Below each question is a space for comments. Please use this space to explain your agreement or to make comments about the system. If an item is not applicable, or if you cannot answer the question, circle DK (Don't Know) at the end of the ratings.

The following information is requested in order to better interpret and analyze responses. All individual information will be treated as confidential and will not be released to third parties. If you have already completed one survey for another system just enter your name. Please complete the information as appropriate.

NAME ______________________ RANK __________ BRANCH __________
ORGANIZATION (for data collectors only)______________________
DUTY POSITIONS WHILE USING MapInfo ______________________
DUTY POSITION LOCATION (circle): FORWARD INTEL CELL REAR OTHER SUBORDINATE CELL (LIST)

Approximate number of hours experience with this aid ______________________

Did you actually operate the aid, just observe its use, or direct its operation by an operator (circle): OPERATE OBSERVE DIRECT

PLEASE RETURN BY 27 MAY 94. RETURN VIA DISTRIBUTION ENVELOPES TO: ARI (DR. RIEDEL).

STATEMENT	OPINION
1. Overall, use of MapInfo will improve planning and execution performance.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
2. The system uses the right information in producing its results.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
3. The user can easily modify information already in MapInfo, if necessary.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
4. The calculations performed by the system are helpful.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
5. MapInfo responds quickly enough to user's commands.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	

STATEMENT	OPINION
6. The system would improve the quality of user's work.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
7. Users would have a lot of confidence in the results obtained working with MapInfo.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
8. It was easy to form a mental picture of how the system works.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
9. Users can easily supply the information the system asks for.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
10. The procedures used by the system are consistent with Army doctrine.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	

STATEMENT	OPINION
11. MapInfo would fit well into staff operations.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
12. The system will facilitate the flow of information during planning and execution.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
13. MapInfo will decrease the workload for staff officers.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
14. Staff officers will support the use of the system.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
15. Users would feel in control of MapInfo while using it.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	

STATEMENT	OPINION							
16. It is easy to recover from errors made while using the system.	*Strongly Disagree* 1	2	3	*Neither Agree Nor Disagree* 4	5	6	*Strongly Agree* 7	*Don't Know* DK
Comments:								
17. It is always clear to the user where he is in MapInfo.	*Strongly Disagree* 1	2	3	*Neither Agree Nor Disagree* 4	5	6	*Strongly Agree* 7	*Don't Know* DK
Comments:								
18. In general, MapInfo uses the same information user's would use in performing planning and execution tasks.	*Strongly Disagree* 1	2	3	*Neither Agree Nor Disagree* 4	5	6	*Strongly Agree* 7	*Don't Know* DK
Comments:								
19. MapInfo provides the user with effective directions so that he knows what to do next.	*Strongly Disagree* 1	2	3	*Neither Agree Nor Disagree* 4	5	6	*Strongly Agree* 7	*Don't Know* DK
Comments:								
20. It is clear how to get the system to perform the actions one wants.	*Strongly Disagree* 1	2	3	*Neither Agree Nor Disagree* 4	5	6	*Strongly Agree* 7	*Don't Know* DK
Comments:								

STATEMENT	OPINION
21. Planning and execution tasks can be completed faster using the system than not using MapInfo.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
22. The system's products meet user's needs.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
23. Overall, MapInfo is a valuable tool for tactical planning and execution.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
24. The system provided users with useful results.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
25. Users would have a lot of confidence in MapInfo as an aid in staff operations.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	

STATEMENT	OPINION
26. The terminology used by MapInfo is consistent with Army doctrine.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
27. The system permits the user to control the order in which different tasks are done.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
28. The user can easily move to different parts of MapInfo as required to perform the tasks.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
29. Users would feel comfortable using the system under time pressure.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
30. The organization of the menus is easy to understand.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	

STATEMENT	OPINION
31. MapInfo will not interfere with the flow of information in staff operations.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
32. The commander would strongly support using MapInfo during planning and operations.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	
33. It is easy to enter data into MapInfo.	*Strongly Disagree* 1 2 3 *Neither Agree Nor Disagree* 4 5 6 *Strongly Agree* 7 *Don't Know* DK
Comments:	

Please circle the appropriate response.

34. How well would MapInfo meet users' needs during an actual exercise?

Not At All		Some What		Very Much
1	2	3	4	5

35. To include MapInfo in staff operations would be:

Very Difficult	Moderately Difficult	No Problem	Moderately Easy	Very Easy

36. How much would use of MapInfo improve the quality of users' work?

Not At All		Some What		Very Much
1	2	3	4	5

37. To what extent would use of MapInfo <u>improve</u> performance on the following tasks? Listed are general tasks and selected specific tasks.

	Not At All		Some What		Very Much	Don't Know
Mission Analysis	☐	☐	☐	☐	☐	☐
Increase situational awareness	☐	☐	☐	☐	☐	☐
Track flow of key events	☐	☐	☐	☐	☐	☐
Develop shared understanding of battlefield	☐	☐	☐	☐	☐	☐
IPB	☐	☐	☐	☐	☐	☐
Increase situational awareness	☐	☐	☐	☐	☐	☐
Track flow of key events	☐	☐	☐	☐	☐	☐
Develop shared understanding of battlefield	☐	☐	☐	☐	☐	☐
COA Development	☐	☐	☐	☐	☐	☐
Track flow of key events	☐	☐	☐	☐	☐	☐
Develop shared understanding of battlefield	☐	☐	☐	☐	☐	☐
COA Analysis and Comparison	☐	☐	☐	☐	☐	☐
Track flow of key events	☐	☐	☐	☐	☐	☐
Develop shared understanding of battlefield	☐	☐	☐	☐	☐	☐
Rehearse Plans	☐	☐	☐	☐	☐	☐
Develop Synchmatrix	☐	☐	☐	☐	☐	☐
Commander's Decision	☐	☐	☐	☐	☐	☐
Synchronize tactical operations	☐	☐	☐	☐	☐	☐
Develop shared understanding of battlefield	☐	☐	☐	☐	☐	☐

37. (Continued)

To what extent would use of MapInfo improve performance on the following tasks? Listed are general tasks and selected specific tasks.

	Not At All		Some What		Very Much	Don't Know
Execute Plan.	☐	☐	☐	☐	☐	☐
Increase situational awareness.	☐	☐	☐	☐	☐	☐
Track flow of key events.	☐	☐	☐	☐	☐	☐
Develop shared understanding of battlefield.	☐	☐	☐	☐	☐	☐
Synchronize tactical operations.	☐	☐	☐	☐	☐	☐
Direct and lead subordinate troops.	☐	☐	☐	☐	☐	☐
Control battle tempo.	☐	☐	☐	☐	☐	☐
Lateral coordination.	☐	☐	☐	☐	☐	☐
Command on the move.	☐	☐	☐	☐	☐	☐
Track flow of key events.	☐	☐	☐	☐	☐	☐

38. This section examines the ability of specific MapInfo capabilities to:
  (1) increase situational awareness,
  (2) facilitate a shared understanding of the battlefield,
  (3) assist in the synchronization of tactical operations, including lateral coordination, and
  (4) facilitate plan rehearsal.

For each capability below please:
  (1) rate the potential of this type of capability to improve performance on each of the 4 tasks,
  (2) rate how much current MapInfo improves performance compared to unaided performance.
  (3) describe changes that should be made to MapInfo's capabilities to improve it.
Put a number in each blank using the following scale.

No Improvement	Slight Improvement	Improvement	Much Improvement	Very Much Improvement	Not Applicable
1	2	3	4	5	NA

MapInfo CAPABILITY	DEVELOP SIT. AWARENESS	SYNCHRONIZE TACTICAL OPS	DEV. SHARED UNDERSTANDING OF BATTLEFIELD	REHEARSE PLAN
Attaching data to a map object (point & click to obtain info)				
Potential	___	___	___	___
MapInfo	___	___	___	___
Changes to MapInfo? ______________________________				
Layering (displaying and moving up to 100 different overlays)				
Potential	___	___	___	___
MapInfo	___	___	___	___
Changes to MapInfo? ______________________________				
Thematic Representation of Data (visually shade objects based on attribute or as a result of a query)				
Potential	___	___	___	___
MapInfo	___	___	___	___
Changes to MapInfo? ______________________________				
Geographic Analysis (find objects and perform geographic searches/radius and polygon)				
Potential	___	___	___	___
MapInfo	___	___	___	___
Changes to MapInfo? ______________________________				
Multiple Views of Data (simultaneous viewing of maps, graphs and charts)				
Potential	___	___	___	___
MapInfo	___	___	___	___
Changes to MapInfo? ______________________________				

38. (Continued)

For each capability below please:

(1) rate the potential of this type of capability to improve performance on each of the 4 tasks,

(2) rate how much current MapInfo improves performance compared to unaided performance.

(3) describe changes that should be made to MapInfo's capabilities to improve it.

Put a number in each blank using the following scale.

No Improvement	Slight Improvement	Improvement	Much Improvement	Very Much Improvement	Not Applicable
1	2	3	4	5	NA

MapInfo CAPABILITY	DEVELOP SIT. AWARENESS	SYNCHRONIZE TACTICAL OPS	DEV. SHARED UNDERSTANDING OF BATTLEFIELD	REHEARSE PLAN
Map Display Capability				
Potential	___	___	___	___
MapInfo	___	___	___	___

Changes to MapInfo? ____

____

39. Should additional capabilities be added? Yes _____ No _____
If yes, please explain. ______________________________

____

40. Should any capabilities be eliminated or changed? Yes _____ No _____
If yes, please explain. ______________________________

____

41. What are the advantages, disadvantages, and limitations of MapInfo?

____

____

42. How satisfactory was MapInfo in being flexible to meet varying task and time demands?

Very Satisfactory	Satisfactory	Borderline	Unsatisfactory	Very Unsatisfactory

43. Overall comments and suggestions: ______________________________

____

____

____

____

APPENDIX C

CONSTRUCT VALIDATION QUESTIONS

CONSTRUCT VALIDATION QUESTIONS FOR SIX COMPARISONS

All question numbers refer to items in Appendix B.

COMPARISON 1: Overall utility score, calculated from attribute scores, correlated with mean of questionnaire items 1 and 23 (node 0.0) below.

1. Overall, use of [Decision Aid name] will improve planning and execution performance.

23. Overall, [Decision Aid name] is a valuable tool for tactical planning and execution.

COMPARISON 2: Overall utility score, calculated from attribute scores, correlated with added validation question 34.

34. How well would [Decision Aid name] meet users' needs during an actual exercise?

Not At All		Some What		Very Much
1	2	3	4	5

COMPARISON 3: Overall utility scores, calculated from attribute scores, correlated with mean of responses to added validation question 37 in Appendix B.

COMPARISON 4. Mean of quality of results criterion items (questionnaire items 6 and 24, node 1.2.2)) correlated with added validation question 36.

6. The system would improve the quality of users' work.

24. The system provides users with useful results.

36. How much would use of [Decision Aid name] improve the quality of users' work?

Not At All		Some What		Very Much
1	2	3	4	5

COMPARISON 5: Organizational fit criterion item (questionnaire item 11, node 3.2.2) correlated with added validation question 35.

11. [Decision Aid name] would fit well into staff operations.

35. To include [Decision Aid name] in staff operations would be:

Very Difficult	Moderately Difficult	No Problem	Moderately Easy	Very Easy

COMPARISON 6: Time to perform task criterion score (questionnaire items 21 and 29, node 1.1.5) correlated with added validation question 42.

21. Planning and execution tasks can be completed faster using the system than not using [Decision Aid name].

29. Users would feel comfortable using the system under time pressure.

42. How satisfactory was [Decision Aid name] in being flexible to meet varying task and time demands?

Very Satisfactory	Satisfactory	Borderline	Unsatisfactory	Very Unsatisfactory