AN ASSESSMENT THE USE OF A STRUCTURED, SUBJECTIVE METHOD TO EVALUATE THE QUALITY OF DECISIONS IN COMPLEX, ILL-STRUCTURED PROBLEMS

by

Walter A. Powell

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Information Technology

Committee:

Dr. Kathryn Blackmond Laskey, Dissertation Director

Dr. Leonard Adelman, Committee Member

Dr. Daniel Barbara, Committee Member

Dr. Karla Hoffman, Committee Member

Dr. Andrew Loerch, Committee Member

Dr. Stephen Nash, Senior Associate Dean

Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering

Date: 25 APRIL 2014

Spring Semester 2014
George Mason University
Fairfax, VA

An Assessment of the Use of a Structured, Subjective Method to Evaluate the Quality of Decisions in Complex, Ill-structured Problems

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Walter A. Powell
Master of Engineering
Cornell University, 1991
Bachelor of Science
Unites States Naval Academy, 1983

Chairman: Kathryn Blackmond Laskey
Volgenau School of Engineering

Spring Semester 2014
George Mason University
Fairfax, VA

ACKNOWLEDGEMENTS

I would never have been able to complete this dissertation without the guidance from my committee, support from my family, and input from the Friday afternoon KRYPTON seminar.

I would like to thank my committee for their patience and support. I would especially like to thank my advisor, Professor Katheryn Blackmond Laskey, for her limitless (ok, not so limitless) patience, steadfast support, and encouragement through my lengthy journey to get here. I would also give a special thanks to Professor Leonard Adelman for his guidance and providing the foundation for the knowledge of experimental design, conducting effective experiments, the collection of data, and the analysis of that data that made this research possible.

I would also like to thank my family for teasing and guilting me in to finally finishing writing this dissertation. I would also like to give my heartfelt thanks to Mom and Dad for their tireless efforts while editing several versions of this dissertation while on vacation in a trailer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

AN ASSESSMENT OF THE USE OF A STRUCTURED, SUBJECTIVE METHOD TO EVALUATE THE QUALITY OF DECISIONS IN COMPLEX, ILL-STRUCTURED PROBLEMS

Walter A. Powell, Ph. D.

George Mason University, 2014

Dissertation Director: Dr. Kathryn Blackmond Laskey

Improving the quality of decisions made by decision-makers is the ultimate goal of research into decision-making, and the ability to assess the quality of decisions is central to decision research. The ability to assess the quality of decisions is crucial to determining whether actual decision-making conforms to theories of decision making, whether decision-making tools affect real-world decision-making, and determining which series of complex decisions (plans) would be best to implement. Much research has been conducted into decision-making and decision quality, but most of this research concerns problems that are well-structured (the best answer is knowable) and that are of limited complexity. Decision research into this class of problems typically uses the desirability of outcomes or the rationality of the decision process as the basis for evaluating the quality of decisions. While these methods may be appropriate for well-structured problem, they do not seem appropriate for complex, ill-structured problems in which

rational decision processes do not necessarily lead to a single best solution, the solution generated may not be implemented or may not lead to a single best outcome. Therefore another method, ideally a direct assessment of decision quality, is needed in order to evaluate decision-making in complex, ill-structured problems.

Since complex, ill-structured problems are likely to have few objective measures and therefore require subjective judgments on the part of decision-makers, the evaluation of the quality of the decisions made to address these problems likewise needs to be primarily subjective. However, little research has been conducted into either subjectively assessing decision quality or directly assessing decision quality (without the use of a proxy) for complex, ill-structured problems. The research documented here evaluates the effectiveness of using a structured, subjective method for directly evaluating decision quality. The use of a structured subjective method was investigated in two cases studies in which real world military problems of different complexities were addressed using different decision-making processes. Together, the case studies demonstrated that a structured, subjective approach was effective in directly evaluating the quality of decisions and a that a structured, subject evaluation is robust in that it can be used to evaluate the decision quality of decisions for complex, ill-structured problems of varying complexity.

CHAPTER ONE: INTRODUCTION

The Problem

Complex, ill-structured problems have long been identified as archetypal of real-world problems (Berry & Broadbent, 1995; Hagmayer & Meder, 2013). Many of the problems we grapple with on a daily basis are complex, ill-structured problems. This class of problem can have national or international impact, local impact, or personal impact. Examples of problems that could impact many levels of society are contingency planning and the response to natural and man-made disasters. For such infrequent but potentially devastating occurrences, decisions and plans must be made that address the economic, infrastructure, personal, and environmental impact of weather events such as Katrina, man-made disasters such as the BP oil spill, and other crises. On a smaller scale, a college football coach has to consider the team's current strengths and weaknesses when designing plays; he has to consider the capabilities of the team's rivals when deciding on game strategy; and he has to consider the team's long-term strengths and weaknesses when making recruiting decisions. These decisions, whether for a football team or for disaster planning, address complex, ill-structured problems. Subject matter experts, such as coaches or emergency managers, are highly valued for their expertise and their judgment, and are entrusted with evaluating complex decisions and selecting the best course of action.

What do decisions made when planning for disaster response have in common with the decisions made throughout a football season? They are decisions made in response to complex, ill-structured problems. Planning for disasters and designing a football play both require assessing many disparate factors and the dynamic relationships among them to make decisions that are as likely as possible to result in a desirable outcome.

But what makes football and disaster response planning complex, ill-structured problems? The definition of what makes a problem complex has evolved as the types of problems being solved have changed (Quesada, et al., 2005). The definition varies with the specific discipline being considered. Briefly, and for decision theory specifically, complex problems are ill-defined and unstructured problems (Robbins & Hall, 2007) that represent multifaceted, dynamic situations involving a large number of interconnected elements. Complex problems also tend to be polytelic, i.e., implying multiple conflicting goals (Blech & Funke, 2010). Ill-structured problems are a subset of complex problems that are further characterized by having underlying relationships among a problem's elements that are unquantifiable, unknown, or unknowable. This uncertainty in the relationships between the elements of a decision means that formulating objective functions is difficult, and therefore modeling the entirety of these problems is not practical. These factors characterize the essence of complex, ill-structured problems in general. For the purposes of this research, complex, ill-structured problems are considered to be those problems that have uncertain aspects, that cannot be solved with

current optimization methods, that cannot be practically modeled in their entirety, and that ultimately require human judgment to resolve.

The uncertainty in complex, ill-structured problems makes evaluation of the quality of decisions difficult. For example, many of the decisions made to manage a sports team are ill-structured and are of varying complexity. Consider the design of a football play: even with years of experience and running the play many times in practice and in games, the outcome of executing that play is uncertain. Similarly, the uncertainties considered when deciding on a game strategy, which players to play, and which plays to run will certainly impact the outcome of the game. Or even consider the decisions made when designing a strategy for the season and deciding which players to recruit. The prefect recruit may rupture his Achilles tendon and thus impair the team's performance for the season and beyond. So, are the choices of specific plays, players, and recruits "good" decisions even if the overall results are not necessarily desirable? Due to the uncertainty inherent in complex, ill-structured problems, i.e., due to uncertainty in the context of the decision, tying such individual decisions directly to outcomes is often difficult.

Even more difficult is using outcomes to evaluate decisions made for complex, ill-structured problems such as disaster planning. Many decisions are made as possible options or as contingencies and are never implemented. Even though many governments, businesses, and other organizations have plans for dealing with hurricanes, floods, earthquakes, or oil spills, those plans are rarely implemented; and in fact those plans are made hoping they are never implemented. Likewise, the military typically generates

several plans for any operation, but at most only one plan (series of decisions) will ever be executed. As a result, for many decisions like those made for disaster contingency and military planning, there are few outcomes by which decisions can be evaluated. This lack of outcomes, like uncertainty in outcomes, emphasizes the need for a method to evaluate decision quality that does not rely on outcomes.

Even though outcomes are not particularly suitable for evaluating complex, ill-structured decisions, outcomes are used in the evaluation of decision quality. They are used as proxies for evaluating decision quality because the evaluation of decision quality directly is difficult. This is a sound approach for problems, such as laboratory experiments, where the uncertainty in the context of the decision can be controlled; or if not controlled, the uncertainty can be quantified using simulations. This is not the case for complex, ill-structured problems. The uncertainty in the relationships among the elements of a decision addressing this type of problem are so uncertain, unknown, or unknowable that the uncertainty in the context cannot be controlled or captured via simulation. The uncertainty in the context is often so great that there is no clear objective function that can be used to evaluate decisions, and often there so no clear path for developing an objective function (de Silva, et al., 2003). This level of uncertainty in the context and relationships means that, given current methods, there is no way to determine that a single, best decision for a complex, ill-structured problem exists (Davern, et al., 2008).

The lack of a single, best decision creates a significant problem in the evaluation of decision quality in all types of decisions; there is no global standard of decision

quality. Without some kind of standard of decision quality, whether global or specific to a problem, there is no scale with which to determine how good a decision is. The lack of a single best solution means that in addition to there being no global scale of decision quality to reference, there is also no "ground truth" specific to a given problem. Without a "ground truth," a specific standard of decision quality cannot be established. In cases where there is no "ground truth," researchers are usually driven to evaluate a decision using a criterion such as the outcome of the decision, not the decision itself (Philips-Wren, et al., 2004); but as has been discussed, the evaluation of decision quality using outcomes is not viable for complex, ill-structured problems. Because of this lack of a "ground truth" and the inability to use outcome-based evaluation, and in contrast to simpler, well-structured problems that have been relatively well explored, evaluating the quality of decisions addressing complex, ill-structured problems is difficult and remains relatively unexplored (Bennett & Bennett, 2008, Fischer, 2011; Esereyel et al., 2013).

Given that a standard of decision quality is not available and that using outcomes is not a variable method of assessing decision quality, a direct assessment of decision quality would be a valuable contribution. Figure 1 illustrates a typical decision process from planning to outcome/implementation and at what point in a decision process a direct assessment of decision quality would occur. A typical decision process is initiated with a decision problem. This problem is also associated with an anticipated context. The context on which the decision-makers base their decision is the context which is anticipated to exist when the decision is implemented. Decision-makers use an anticipated context as a means of limiting the uncertainty in the context; the decision-

makers understand that reducing the uncertainty between the anticipated context and the context in which the decision is actually implemented will be more likely to lead to a desirable outcome.

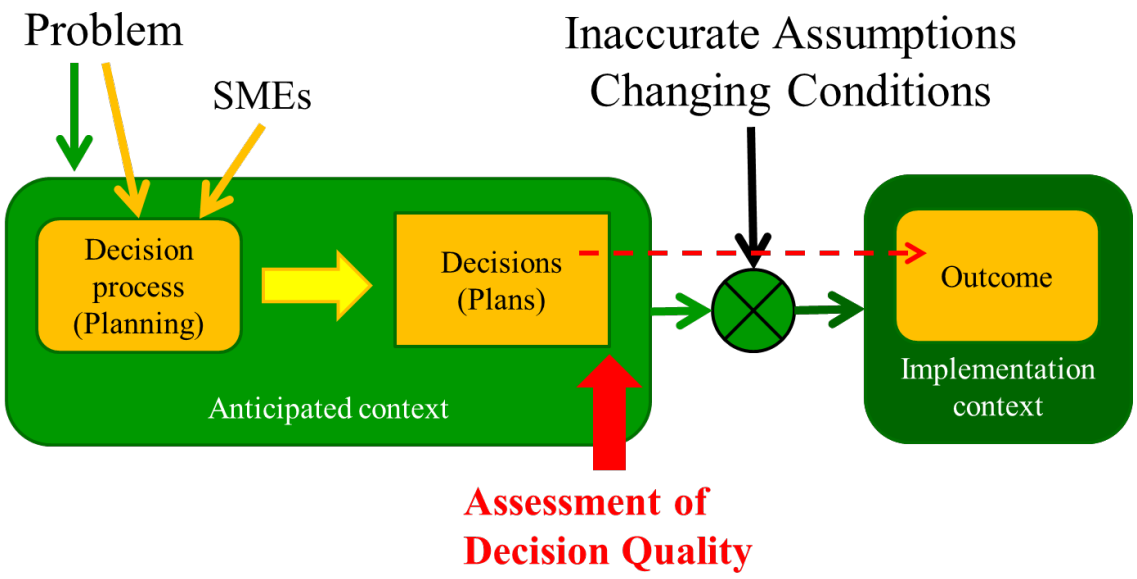

Figure 1: Assessment of Decision Quality in a Decision Process

The problem also has associated with it Subject Matter Experts (SMEs) who are decision-makers knowledgeable in the specific problem being addressed. They make decisions and develop plans designed to provide solutions, sets of decisions (plans), or courses of action designed to produce desirable outcomes. In a complete process, no more than one decision is implemented and one outcome generated. But, due to inaccurate assumptions and changing conditions, the implementation context is not identical to the anticipated context. Therefore, the linkage between the decision and the

outcome is not perfect, and the desirability of the outcome may not be closely linked to the quality of the decision (dashed arrow).

Because the desirability of the outcome may be uncertain or because the decision may not be implemented, the assessment of the quality of decision-makers' decisions, plans, or courses of action must be evaluated at the end of the decision-making process. This assessment must directly evaluate the quality of the decisions and be able to discriminate among the quality of those decisions. Because the decision-making requires the expert and subjective judgment of the decision-makers, the assessment of the decisions must be made using the judgment of SMEs and will be primarily subjective.

Based on the need to be able to discriminate among the qualities of decisions, the general need for a method to evaluate decision quality in complex, ill-structured problems, the requirement for a primarily subjective evaluation method, and recognizing the difficulty in using outcomes as proxies for decision quality, the following research hypothesis and method were used to assess whether a structured, subjective method could be used to evaluate decision quality.

Research Hypothesis

Given the considerations discussed above, one possible method of directly evaluating decision quality is through a structured approach that subjectively captures the relationships among the elements of a decision. Multi-Criteria Decision Analysis (MCDA) is an area of study that advocates the use of structured methods to identify and quantify relationships as part of a decision-making process. Although quantifying relationships in complex, ill-structured problems is extremely difficult, a structured

method (Chen, et al., 2011), a Decision Quality Evaluation Method (DQEM), could be developed from standard MCDA methods and adapted to capture the understanding of subject matter experts (SMEs) who are knowledgeable of the elements and relationships associated with a problem of interest. This DQEM would employ (1) a detailed decomposition of decision quality characteristics and (2) a unique scoring procedure to capture the SMEs' knowledge and aid them in subjectively evaluating the quality of decisions. Therefore, the hypothesis this research is designed to address is:

> *The direct evaluation of the quality of decisions made to address complex, ill-structured problems can be improved through the use of a structured subjective decomposition of decision quality characteristics.*

Several assumptions are inherent in using a structured approach to evaluate decision quality. These assumptions are well documented and accepted in the MCDA approach to decision-making. These assumptions have been adapted for use in the evaluation of decision quality and provide the basis for the development of the DQEM. The assumptions that form the foundation of this research are:

For a complex, ill-structured problem and an associated context, there exists some combination of measurable subjective characteristics that define what constitutes a good decision in response to that problem. Elicited from Subject Matter Experts (SMEs), these characteristics are those that, when part of a decision, make the decision robust and more likely to lead to desirable outcome. Because decisions having these characteristics are perceived to more likely lead to desirable outcomes, these characteristics are considered the characteristics of a "good" decision. But, due to the ill-structure of the

problems targeted in this research, the presence of good characteristics means that the decision may be "good" only with respect to the decision-making context.

The subjective characteristics of a "good" decision can be decomposed into subjective sub-characteristics that support the general characteristics. Further, these sub-characteristics can be further decomposed into additional sub-characteristics that also support the general characteristics. Also, subjective measures can be defined to support the evaluation of each sub-characteristic. Finally, specific subjective criteria can be generated that define the scope of each measure.

Using stated criteria, the subjective measures can be independently evaluated by a SME and a numeric score assigned to each measure. SMEs are necessary to conduct an evaluation of decision quality because no decomposition can capture all the nuances of the relationships among the factors of the decision. Although, the subjective evaluation of decision quality relies on the knowledge and judgment of SMEs, the criteria generated by the decomposition will serve to focus the SMEs on those factors that are most pertinent to the overall decision. To support the assignment of scores representing the quality of the decisions associated with each measure, Likert scales can be tailored to each measure that relate the achievement of specific criteria to specific scores on the Likert scale.

The scores associated with the evaluation of the decision quality measures can be aggregated to generate an overall score of decision quality. This aggregation is typical in MCDA using objective measures and is applicable to the subjective evaluation. Once Likert scale scores are associated with each measure, the individual measure scores can

be aggregated to generate an overall decision quality score. Methods to aggregate data are well documented in the literature.

Research Approach

In order to test the research hypothesis, several components were constructed. First a structured, subjective method was developed which would allow SMEs to evaluate the decision quality characteristics of the decision-makers' decisions. Second, a means of scoring was developed which would translate the SMEs' evaluations into numeric scores while minimizing the variance introduced by the scoring technique itself. Third, a means of aggregating the SMEs' scores for scores for individual decision quality characteristics was devised. Once the case studies were developed, real world decision-makers making real world decisions were identified. Fortunately, the Army Geospatial Center (AGC) contracted with the George Mason University C4I Center of Excellence to evaluate the value of adding Geospatial Decision Support Systems (GDSSs) to the Military Decision Making Process (MDMP). This afforded the researcher access to real-world problems, decision-making processes, decision-makers, and SMEs needed for the two case studies in which the structured, subjective approach was assessed.

The method assessed in this research applied and extended the principles of Multi-Criteria Decision Analysis (MCDA) to a subjective evaluation of the quality of decisions. The method, the Decision Quality Evaluation Method (DQEM), uses detailed decomposition of the characteristics of "good" decisions generated through an interactive process based upon achieving consensus among SMEs. Generating the decomposition was an iterative process in which the SMEs decomposed decision quality characteristics

into sub-characteristics until they reached consensus on measures that could be evaluated independently of other measures. This interactive process served to capture some of the SMEs' understanding of the relationships among the elements of the decision and served to both expose and resolve differing biases in the SMEs' opinions.

Once the decomposition of decision quality characteristics was complete, SMEs generated scoring criteria. The scoring criteria were specific to each decision quality measure and served to define and limit the scope of each measure. Using the criteria, Likert scales were tailored for each measure. Similar to Behaviorally Anchored Rating Scales (BARS), these tailored Likert scales served three purposes. First, they captured additional information on the SMEs' understanding of the relationships among the decision elements. This was accomplished by specifying the accomplishment of specific criteria required to achieve each score on the Likert scale. This process also achieved the second purpose of relating the SMEs' subjective evaluations to numeric scores. Lastly, the tailoring of the Likert scales served to reduce the variance in scores that would be present due to individual SME interpretation of traditional Likert scales.

Once the scoring mechanisms were developed, research indicated that weighted aggregation of the SMEs' scores might provide a better representation of their understanding of the relative importance of each decision quality characteristic. Other research suggested that weighted averages did not perform better than simple averages when aggregating data. Therefore, both weighted and simple averages were used to aggregate the data.

The DEQM, including the decomposition of decision quality, the tailored Likert scales for scoring, and the aggregation mechanisms coupled with standard experimental design techniques would be sufficient to meet AGC's goal of evaluating the usefulness of the new GDSSs; but in order to evaluate the effectiveness of the DQEM in aiding SMEs evaluation of decision quality, some measure of the ability of the SMEs to evaluate decision quality needed to be developed. And, since there is no general standard of decision quality and since complex, ill-structured problems do not have a "ground truth" solution, accuracy could not be used as a measure of the SME's evaluations for this class of problem. Using scoring criteria and Tailored Likert scales, the SMEs could translate their evaluations into scores; but without a standard, either global or specific to the problem against which decisions could be compared, the accuracy of the SME's evaluations has no meaning. And, without a standard, the only information that could be gleaned from the SMEs' evaluations would be the relative quality of the decision-makers' decisions. Therefore the SMEs' ability to consistently, evaluate decision quality could be a useful alternate measure of the effectiveness of the DEQM. The measure of consistency used in this research was the inter-rater reliability of the SMEs' evaluations.

To measure the agreement among the SMEs' evaluations, this research uses the correlation between the SMEs' subjective overall evaluation (SOE) scores and the correlation between the SMEs' averaged decision qualities characteristic scores (DQCSs). Statistical evidence that these correlations had increased would indicate that the DEQM was effective in aiding the SMEs in making evaluations of decision quality that were more reliable.

This definition of reliability lends itself to a second measure of the effectiveness of the DQEM. As the SMEs' evaluations become more reliable, their evaluations should be able to better differentiate among the qualities of various decisions. Further, if using the DQEM allows the SMEs' evaluations to differentiate sufficiently among the quality of various decisions, then a ranking of the SMEs' scores for the decision-makers decisions should be identical. An analysis using Spearman's Rank Correlation Coefficient was used to measure the agreement in the rankings of the SMEs' scores.

Since there is no way to compare the quality of decisions without actually evaluating the decision, and since there is no other means of evaluating the decision quality associated with complex, ill-structured problems other than DQEM, it was impossible to structure the assessment of the DQEM using traditional experimental design in which one condition used the DQEM and one did not. Therefore in order to assess the value of the DQEM, an alternate basis for comparison was used. Since the DQEM depends on a detailed decomposition of decision quality and since including more relevant detail in the decomposition should better capture the subjective SMEs' understanding of the factors and relationships in the problem, the comparisons of the SMEs' ability to reliably evaluate decision quality was made between their scores at three points in the decomposition corresponding to three different levels of detail in the decomposition. Evaluating at three points during the decomposition allowed for trends in the effectiveness to be more evident than if only two evaluations of decision quality were used. Due to these considerations, the specific hypothesis tested by the analysis of the three sets of data is the following:

As the decomposition of decision quality better captures the SMEs' understanding of the problem through the inclusion of greater relevant detail, the evaluations of decision quality would become more reliable.

The goal of this research was to assess the reliability of a structured, subjective approach to improve the SMEs' ability to evaluate the quality of decisions associated with complex, ill-structured problems through an approach that directly evaluated the quality of decisions. The contributions of this research to the scientific knowledge base are detailed below.

Contributions

This research made the following contributions to the field of decision support by developing and applying a structured evaluation method that:

- **Directly evaluated decision quality** using a Decision Quality Evaluation Method (DQEM) that focuses on characterizing the essential elements of decision quality such that, in aggregate, the relative overall quality of decisions can be evaluated without the use of outcomes and proxies for decision quality. The method uses a procedure that decomposes decision quality hierarchically into decision quality characteristics, sub-characteristics, measures, and criteria. Tailored Likert scales were used to capture the relationship among decision quality characteristics, to translate subjective evaluation into numeric scores, and to reduce the variance in scoring due to scale ambiguity.

- **Applied multi-criteria decision analysis (MCDA)** techniques to construct a **structured and detailed** decomposition of **subjective decision quality characteristics** designed to be used by independent SMEs to evaluate the quality of decisions
- **Used the consensus opinions of SMEs** to **capture the SMEs' subjective understanding** of the complex relationships among the factors affecting the quality of decisions without the need to resort to a time-consuming, tedious process that elicits values quantifying these subjective factors. The use of an SME consensus was also used to identify and resolve differing SME biases.
- **Defined tailored Likert scales** that further captured the SMEs' understanding of the relative importance of specific criteria when evaluating the quality of decisions. These Likert scales allowed SMEs to independently translate the achievement of criteria to specific scores, thereby reducing the variation in their scores due to individual interpretation of the Likert scales.
- **Developed a method (the DQEM)** that was designed such that third party SMEs could use the decomposition and tailored Likert scales to independently guide and score evaluations of similar decisions.

Further, this research **demonstrated the effectiveness of the DQEM** in two case studies evaluating the impact of differing decision-making processes on complex real world decision-making. The effectiveness of the DQEM was demonstrated in two ways:

first, by evaluating the inter-rater reliability of the SMEs' evaluations of decision-makers' decisions; and second, by evaluating the similarity of ranking of the SMEs' evaluations.

Finally, use of the DQEM in these two case studies **demonstrated the flexibility and adaptability of the DQEM.** The case studies demonstrated that the DQEM could be successfully applied to the evaluation of decision-making in ill-structured problems of different complexities. Specifically, Case Studies One and Two evaluated decision-making processes that were based on the Military Decision Making Process (MDMP) but were designed to assist decision-makers with different skill sets to make decisions for problems of significantly different complexity that required more extensive and more complex decisions, and that yielded more sophisticated outputs.

Finally, the detailed description of the implementation of the DQEM in Case Study 1 **provided a guide for the construction of evaluations using decision quality as the primary measure**.

The remainder of this dissertation is organized as follows:

- Chapter 2, Background and Literature Review
- Chapter 3, Method
- Chapter 4, Results for and evaluation of Case Study One
- Chapter 5, Results for and evaluation of Case Study Two
- Chapter 6: Conclusions and future work

CHAPTER TWO: BACKGROUND AND LITERATURE REVIEW

The literature relevant to this research is diverse and inter-disciplinary. The main focus of the research is concerned with decision quality, a topic directly addressed in the areas of psychology and decision theory, as well as discussed in various areas where understanding decisions are important: medicine, business, the military, and Decision Support Systems (DSS) design to name a few. The literature in the following areas was reviewed in the courses of this research:

- Complex problems and Ill-structured problems. Problems in this class are those for which it is the most difficult to define and assess decision quality. Not coincidently, this is the class of problems which DSSs, SDSSs, and GDSPs are being developed to support. Understanding the type of problems that DSSs, SDSSs, and GDSPs would be used to support was critical to the development of the DQEM.
- Decision Quality. Because the goal of evaluating decision quality is to improve decision-making, an understanding of what makes a decision "good" was necessary. Specifically, an understanding of the strengths and weakness of the prevailing approaches to defining and determining decision quality with respect to complex, ill-structured problems was needed. The strengths and

weaknesses of the prevailing approaches helped to guide the development of the DQEM. Important topics in this area include:

  - Process-oriented approach
  - Outcome-oriented approach
  - Decision-outcome linkage

- Evaluating decision quality and DSSs. Closely linked to decision quality are the methods used to evaluate it. The strengths, weaknesses, and applicability of current methods used to evaluate decision-making in complex, ill-structured problems generally and methods designed to specifically evaluate decision-making using DSSs to addressing complex, ill-structured problems needed to be understood. Specific attention was paid to methods that addressed:

  - Decomposition of decision quality characteristics;
  - Methods of scoring decision quality;
  - Aggregation of characteristic scores into a composite score of decision quality.

Research on decision-making is multi-disciplinary. Decision-making has been addressed by researchers drawing from the areas of psychology, philosophy, decision science, management science, computer science, business, environmental management, economics, political science, and emergency management to name just a few.

Nevertheless, there has been limited research into using decision-quality as a primary measure, decision-quality in highly-complex and ill-structured problems, and evaluating decision quality for these types of problems.

Complex and ill-structured Problems

In virtually all disciplines in which decisions are made, some decisions are classified as more complex than others. Although many of the most interesting problems seem to be classified as complex problems, there is no generally accepted definition of what constitutes a complex problem and no standard that can be used for determining the level of complexity of a given problem (Clark & Richards, 2002; Hagmayer & Meder, 2013). Most of the literature that addresses complex problem solving does not directly address what constitutes a complex problem (e.g. (Shin, et al., 2003; Braddock, et al., 1999). Researchers do generally agree that more complex problems are more difficult to solve than less complex problems (de Silva et al., 2003). While determining whether a problem should be considered complex or not seems to be a relative rather than an absolute issue (Berry & Broadbent, 1995), there seem to be factors that are common in most definitions of a complex problem.

Of the many factors used in the literature in the attempt to define the characteristics common to all problems, researchers commonly identify uncertainty of outcome as an inherent characteristic of problems with which decision-makers must contend (Pomerol & Adam, 2003; Clark & Richards, 2002; Frensch & Funke, 1995). With respect to the complexity of a problem, uncertainty in the structure of the problem

and uncertainty in the data seem to be significant factors that influence the complexity of a problem (Stabell, 1994; Frensch & Funke, 1995b; Buchner, 1995; Huber, 1995; Kerns, 1995; Berry & Broadbent, 1995; Xu et al., 2007; Blech & Funke, 2010). Some factors that relate to a problem's structure and data that contribute to a problem being classified as complex include the following:

1. the problem contains uncertainty of structure and/or data (Casey & Austin, 2002; Vahidov & Fazlollahi, 2004)
2. the problem involves a large number of variables (Berry & Broadbent, 1995; Huber, 1995; Kluwe, 1995)
3. the problem's variables are interconnected (Berry & Broadbent, 1995; Kerns, 1995; Xu, et al., 2007; Funke, 2010)
4. the problem's variables are inter-disciplinary (Clark & Richards, 2002)
5. the problem's variables are not transparent; that is, the effect of the variables on the outcome and other variables is not known or not observable (Stabell, 1994; Huber, 1995; Buchner, 1995; Xu et al., 2007; Funke, 2010)
6. the problem is polytelic; it has a large number of conflicting goals ((Densham, 1991); (Stabell, 1994; Berry & Broadbent, 1995; Casey & Austin, 2002; Xu et al., 2007; Blech & Funke, 2010)
7. the problem requires varying cognitive approaches to solve (Kluwe, 1995)
8. there is a time lag between decisions and effects (Funke, 1991; Berry & Broadbent, 1995; Huber, 1995; Xu et al., 2007)

9. the problem can be decomposed into simpler problems (Kleinmuntz, 1990; Kerns, 1995)

Although there is little research into the relative effect of these factors on complexity, these factors are commonly cited as important general factors that contribute to making a problem complex.

Several of the factors that can make a problem complex also contribute to the problem being well- or ill-structured. Hubert Simon discussed what constitutes an ill-structured problem:

> *An [ill-structured problem] is usually defined as a problem whose structure lacks definition in some respect. A problem is an [ill-structured problem] if it is not a [well-structured] problem. (Simon, 1973)*

In Simon's discussion, the first characteristic of ill-structured problems that he noted was that there is initially no definite criterion to test a proposed solution. Because the solution criterion is vague for ill-structured problems, ill-structured problems generally have multiple valid solutions (Jonassen, D., 2000; Shin et al., 2003; Murphy, 2004; Axelrod & Cohen, 1999) go further to state that "complex problems may not have a single "right" answer…." Crossland et al. when discussing the probabilistic nature of the structure of problems expressed Simon's assertion this way:

> *A well-structured problem has a high probability that a single best solution exists, whereas an ill-structured problem has a low probability of a single best solution.* (Crossland, et al.,, 1995)

Notably, this definition does not define thresholds or the means to evaluate the probability of the existence of a single best solution. When determining whether a problem is ill-structured, the intransparency of some or all of the variables, the impact of variables on other variables, and the impact of variables on the outcome seem to indicate that a complex problem is not well-structured (Fernandes & Simon, 1999). The higher the degree of intransparency (that is, the less one knows about the problem), the more ill structured the problem is (Jonassen, & Hung, 2008). Researchers, likewise, agree that uncertainty concerning a problem's structure and variable interactions is one of the factors that make a problem ill-structured (Frensch & Funke, 1995; Spering et al., 2005; Vahidov & Fazlollahi, 2004).

Though few researchers explicitly relate complexity to ill-structure, Axelrod and Cohen state that "By a complex problem we mean one that….may not have a single "right" answer…" (Axelrod & Cohen, 1999). The implication is that there is come relationship between complexity and ill-structure. Shin et al. (2003), when asserting characteristics of ill-structured problems provides basis to relate complexity and ill-structure. Shin et al.'s characteristics can be associated with the factors affecting the complexity of a problem, listed previously, and indicated by the number(s) following each characteristic below. Ill-structured problems typically (de Silva et al., 2003):

- Fail to present one or more of the problem elements (1)(5)
- Have vaguely defined or unclear goals and unstated constraints (1)(6)c
- Possess multiple solutions, solution paths, or no solution at all (1)(2)(3)(5)(6)(8)

- Possess multiple criteria for evaluating solutions (6)
- Contain uncertainties about which concepts, rules, and principles are necessary for the solution and how they interrelate (1)(2)(3)(4)(5)(6)
- Offer no explicit means for determining the appropriateness or quality of solutions, and(4)(6)(7)
- Require decision-makers to make judgments about the problem and often defend them by expressing personal opinions or beliefs about their interpretation of a problem (7).

Although there seems to be a relationship between complexity and ill-structure, there are problems that exhibit some of characteristics of complex problems, but are not ill-structured. For example, problems that can be modeled with computer programs may deal with large numbers of variables (2), contain interconnected variables (3), have inter-disciplinary variables (4), and can be decomposed into simpler problems (9) (de Silva et al., 2003). Even though problems that exhibit these characteristics may be highly complex, computer programs by their very nature typically do not lend themselves to address uncertainty in the structure of a problem and thus problems that can be modeled with computer programs typically do not exhibit the characteristics of ill-structured problems. Highly complex problems, on the other hand, will exhibit many of the factors of complexity; and will probably also exhibit many characteristics of ill-structure. Although there is little research explicitly relating complexity and ill-structure, the relationship between the factors of complexity and the characteristics of ill-structure

seems to indicate that more complex problems, i.e., highly-complex problems, are likely to be ill-structured.

Taken together, the characteristics common to ill-structured problems make assessment of aspects of decision-making in highly complex problems difficult (Jonassen & Hung, 2008). Researchers generally agree that problems involved in many traditional problem-solving experiments seem to be less complex than real world ones (Berry & Broadbent, 1995; Zsambok, 1997). Because of this generally lesser complexity, the decision-making typically addressed in the literature tends to be decision-making in complex, but well-structured problems. There are two related reasons for this: (1) The existence of a single best solution in well-structured problems provides a reference to which decisions can be compared. (2) The measures of quality, which are related to the best solution, are easier to define (de Silva et al., 2003). Because ill-structured problems do not have a single best solution, determining the relative merit of solutions is difficult (Mysiak, et al., 2005). Without a single best solution, constructing experiments and interpreting the results is difficult. For this reason, discussions in the literature today still mainly address decision-making in well-structured problems (Bennet & Bennet, 2008).

Even though decision-making in complex, ill-structured problems is not well addressed in the decision theory literature, the need to address this class of problem has been noted in the DSS literature. Gorry & Morton (1971) noted a relationship between the structure of a problem and the use of a DSS:

> *A DSS has been defined as a computer system that dealt with a problem where at least some stage was semi-structured or unstructured.*

> *A computer system could be developed to deal with the structured portion of a DSS problem, but the judgment of the decision-maker was brought to bear on the unstructured part, hence constituting a human-machine, problem-solving system.* (Gorry & Morton, 1971)

Even though the need to address the use of DSSs in complex, ill-structured problems was identified in the early stages of computer-based DSSs, this need persists:

> *...in general, a new paradigm for decision making is needed within decision support systems. This paradigm must address decision-making in more complex [and ill-structured] contexts than have been attacked in the past by DSS research.* (Courtney, 2001)

Although a search of the literature on both decision theory and the use of DSSs indicate that highly complex, ill-structured problems are of interest, only limited research into decision-making in this class of problem has been generated. Factors that affect the complexity of a problem and characteristics of ill-structured problems have been identified, but these same factors and characteristics make research into highly complex, ill-structured problems difficult. As seen in the following section, the study of decision-making in this class of problem has been compounded by the difficulty in defining decision quality and a lack of standards against which decisions can be evaluated.

Decision Quality

The term decision quality, like complexity, is pervasive throughout all areas of decision research. A survey of current thinking in the area of decision quality yields

several insights. First, regardless of the specific focus area of the research, the concept of decision quality is fundamental to decision science. Although the concept is fundamental to decision-making, there is little discussion of it, relative to the total discussion of decision-making. Furthermore, like complexity, there is no single, accepted definition of decision quality (Keren & Bruine de Bruin, 2003; Mysiak et al., 2005; Yates, et al., 2003). This does not mean that there has been no attempt to define decision quality. Ironically, the difficulty in defining decision quality lies in the many theories developed to describe decision-making. Keren and Bruine de Bruin succinctly stated the decision quality quandary:

> *...the notion of decision quality poses some basic, but difficult questions: Are decisions ''bad'' if their outcomes are disappointing? Are decisions ''good'' if we are pleased with the results? What about decisions that are poorly defined, have large uncertainties, or have outcomes that lie far in the future? No strictly correct answers to such questions exist, but there are two schools of thought on the matter.* (Keren & Bruine de Bruin, 2003)

Outcome- and process-oriented approaches

Early in the debate concerning decision quality, the economist and Nobel laureate, Herbert Simon (Simon, 1976) wrote about the two categories into which theories on decision quality seem to fall. He distinguished between procedural rationality and substantive rationality in decision-making processes. The terms procedural and substantive were adapted from the law where the former refers to the legal process and

the latter to the outcome. Simon applied the terms to decision-making to differentiate between the rationality of the process of decision-making (procedural) and the rationality of the observable outcome. He further noted that decision-making can be judged by both rationalities and that the judgments can differ.

Simon's thoughts typify the two schools of thought concerning decision quality that are still prevalent in the literature: (1) The outcome approach to decision quality that believes good [higher quality] decisions produce good [higher quality] outcomes. (2) The process approach purports that good [higher quality] decision processes yield good [higher quality] decisions (Keren & Bruine de Bruin, 2003). Implicit in these approaches is the use of proxies for decision quality. In general, assessments of decisions using both approaches do not actually assess the decisions themselves. In both approaches proxies are used as measures of decision quality: (1) The quality of the outcome is a proxy for decision quality in the outcome approach. (2) In the process approach, the degree of adherence to a rational process is a proxy for decision quality. Why is decision quality not directly assessed? Mysaik et al. said it succinctly, "In our opinion, it is…the lack of general consensus about what constitutes decision quality that makes the *evaluation* of [decision quality] difficult (Mysiak et al., 2005). One reason for a lack of consensus may be that various classes of decisions can be defined with each requiring different judgment criteria e.g., (von Winterfeldt, 1980). Likewise, for each class of decision there would seem to be many possible criteria that could be employed when defining decision quality and, depending upon the decision [and context], the quality criteria for that decision will be differently evaluated (Jacoby, 1977).

Both the outcome and process approaches continue to be common in the literature and valid arguments in favor of each approach are put forth. The following quotes demonstrate that the discussion continues. First a comment on the outcome approach:

> *Conceptually, the simplest and most tangible benefit of a DSS is the ability to help or drive its user(s) toward making better decisions. These decisions are better in the sense that, once implemented, they have such effects as reducing costs, using assets more efficiently, increasing revenue, reducing risks….* (Pick, 2008)

The objective outcome is addressed in the literature from many areas of study including psychology (Keys & Schwartz, 2007), decision science (Pick, 2008), business management (Kanungo, et al., 2001), environmental management (Mysiak et al., 2005), marketing (Lilien, et al., 2004), medicine (e.g. Vatali, et al., 2003), to list a few. There seems to be a reason that this approach is used in such diverse areas. The argument for the outcome approach agrees with the general empirical decision-making experience. The outcome process argument goes something like this: since it is generally acknowledged that the goal of research into decisions is to improve decision-making (Howard, 1988), and since the goal of decision-making is to generate an outcome that improves upon the current situation (Yates et al., 2003), then the only logical measure of the quality of the decision is the benefit derived from the outcome.

Further, it is easy to see why the outcome approach has gained acceptance; evaluation by outcome is reinforced in real world decision-making. Decision-makers tend to be evaluated on the outcomes of their decisions (Lipshitz, 1989). Evaluating

decisions by their outcomes is so prevalent that managers have identified the outcome as the most important criterion of decision quality (Zakay, 1984). Even if the decision-making group labeled their decision as "good" at the time it was made, a bad outcome will imply that the decision-maker(s) made a bad decision (Lipshitz, 1989). Given that real life seems to support the outcome approach, it is reasonable that arguments in favor of this approach are more naturalistic than those for the process approach.

There is abundant research that demonstrates that relationships exist between the decision and the outcome (Pool et al., 2003). However, the outcome approach is commonly applied to problems that neither meet the criteria for complex problems nor exhibit the characteristics of ill-structured problems. When evaluating these problems, the context may be artificially limited such that the context in which the decision is made and the context in which the outcome in evaluated are as identical as possible. Problems that are neither highly complex nor ill-structured can conceivably have contexts that are well understood and remain constant throughout the decision process and its implementation (Keys & Schwartz, 2007). When the contexts are the same, i.e., the factors affecting the decisions and outcome are identical; the decision is thought to directly affect the outcome e.g. (Pool et al., 2003). In fact, analyses of the outcomes that achieve statistically significant results are generally interpreted to indicate a causal relationship between the decision and its outcome. These same statistically significant results may also indicate that the context remained relatively constant. Even so, the consistency of the contexts is usually NOT addressed in the literature and seems to be an assumption made during the design of experiments. Commonly, a stronger relationship

between a decision and the outcome seems to exist when the problems are less complex and less ill-structured. In the more complex real world, the results of decision-making research using the outcome approach are not as clear cut (Berry & Broadbent, 1995). This real world uncertainty is a significant element in the argument for the process approach.

The basis for the process approach is that uncertainty and unknowable factors are involved in complex decisions (Keren & Bruine de Bruin, 2003). Edwards et al. succinctly stated the process approach adherents' view of uncertainty:

> *A decision is…a bet, and evaluating it as good or not must depend on the stakes and the odds, not on the outcome"* (Edwards, et al., 1984)

The principal argument for the process approach goes something like this: because there is uncertainty in the outcome of a decision, higher quality decisions should be able to be made if that uncertainty is reduced. Minimizing uncertainty requires modeling a decision with the appropriate structure and context. Adherence to a formal decision process should permit creation of a decision model that closely reflects reality and conforms to the decision-makers goals; and the more closely the model does this, the better the decision will be. Decision quality from the point of view of supporters of the process approach can be summarized as follows:

> *Decision quality is constructed from the building blocks of procedural rationality. ''Quality'', as used here, refers to group and individual decision-making processes that are consistent by design with*

> *organizational values, objectives, and belief systems, as well as empirical evidence.* (Borchers, 2005)

Like the preceding quote, the arguments in favor of the process approach focus in on the process of modeling a decision. Essentially the arguments can be reduced to,

> *If the process has quality (e.g., based on the ''best available'' science), then the decision has quality, and this will favor the emergence of desirable outcomes* (Borchers, 2005)

A fundamental weakness in the process approach lies in the ability to model the structure of the problem. Since it is easier to determine the structure and context of well-structured problems than those of ill-structured problems, it follows that a model of a well-structured problem will more accurately reflect the problem than would a model of an ill-structured problem. The often unstated assumption of the process approach is that given a good decision process that produces an accurate model, the uncertainty of the decision will likely be reduced, and in the long run, will be more likely to result in good outcomes. But, like the outcome approach, as problems become more complex and consequently more ill-structured, the modeling process may be less able to generate accurate models; and thus the decisions that result from these models are likely to be of lower quality.

The primary argument against the process approach is that it does not explicitly relate the decision process to the outcome. Lipshitz (1989) noted that "decision theory defines 'good' processes and that process evaluation is unaffected by outcome

information." Applying the process approach to management, one can see how "a procedurally rational manager is one to whom the outcomes of a decision are irrelevant to its quality" (Borchers, 2005). This total concentration on process to the exclusion of outcome would seem to be counterproductive for managers who are evaluated on results. For example, claiming the surgery was perfect would not generally be seen as a success if the patient dies. Taking the process approach to an extreme, **(**Pick, 2008) asserts that, "from the point of view of adherents to the process approach, even if a [process change] does not lead to better decisions, the decision process may be improved." Such a focus on a process that does not improve decision-making does not seem to achieve the primary goal of decision research: improving decision quality.

Decision-Outcome Linkage (DOL) in complex, ill-structured problems

The nature of the uncertainty inherent in decision-making in complex and ill-structured problems is generally not discussed in the literature. This lack of discussion of the nature of uncertainty precludes discussion of the relationship of a decision to its outcome.

The adherents of the outcome approach assume a direct linkage between the decision and its associated outcome and use the quality of the outcome as a proxy for the quality of the decision. Alternately, the adherents of the process approach assert that there is too much uncertainty in the outcome and ignore the outcome to instead focus on the process as a proxy for decision quality. These two approaches have conflicting views on the Decision-Outcome Linkage (DOL). Neither approach discusses the nature of the

DOL nor why the strength of the DOL seems to change with the complexity and structure of the problem.

In order to understand the decision-outcome linkage, the nature of decisions and outcomes must be understood. Howard summarized the need to differentiate between the decision and the outcome:

> *The most important distinction needed for decision analysis is that between decision and outcome….A good outcome is a future state of the world that we prize relative to other possibilities. A good decision is an action we take that is logically consistent with the alternatives we perceive, the information we have, and the preferences we feel. In an uncertain world, good decisions can lead to bad outcomes, and vice versa. If you listen carefully to ordinary speech, you will see that this distinction is usually not observed. If a bad outcome follows an action, people say that they made a bad decision.* (Howard, 1988)

Howard's definitions of decision and outcome demonstrate the unstated linkage between a decision, the outcome, and the context under which the former is made and the latter is evaluated. He explicitly defines two different contexts relative to a decision; the current (or predicted future) context in which the decision is made and some future context in which an outcome occurs. This concept of differing contexts is central to understanding the DOL. Howard's comments also contain other stated or implied concepts that are relevant to a discussion of the DOL: (1) That there is a non-definite relationship between decision and outcome. (2) That both decisions and outcomes have

contexts that are in part subjective. (3) The DOL itself is subjective. Howard also noted that the evaluation of decisions and outcomes has subjective components which will be important to the discussion of the evaluation of decision quality.

The non-definite relationship between a decision and the outcome of its implementation is alluded to by several authors. The definition of a decision put forth by Yates et al. also implies a somewhat different relationship between a decision and the outcome:

> *The following definition of decision as a synthesis of how the term is actually understood and used across the myriad disciplines that study decision-making…A decision is a commitment to a course of action that is intended to produce a satisfying state of affairs. Thus, quality is part and parcel of the very idea of a decision.* (Yates et al., 2003)

There are two concepts implicit in this statement. First, by the use of the word "intended:" Yates et al. imply that the outcome of the decision will be evaluated at some time after the decision was made. Second, a decision is made with the *intent* to produce a satisfying state of affairs indicating that the DOL may be less definitive than the cause and effect relationship supported by the outcome approach. Yates et al. (2003), more explicitly refers to time in the relation to the DOL when he defines a good decision as follows:

> *One defensible (and common) definition of a good decision is that it is the selection of the best alternative available at the time the decision is made.* (Yates et al., 2003)

Here they explicitly identify time as a qualifier when evaluating the quality of a decision. But, time itself is probably not the factor Yates et al. were trying to identify. They probably intended to say the following:

> *One defensible (and common) definition of a good decision is that it is the selection of the best alternative available in the context in which the decision is made.*

Yates et al. are implicitly linking the time the decision is made and the context in which it is made. They are also implying that the outcome is evaluated at some time other than that at which the decision was made. Similarly, Keys & Schwartz (2007) explicitly link the evaluation of the outcome to the context in which it is evaluated. Neither Yates et al. nor Keys and Schwartz generalize a relationship between the contexts of a decision and the outcome.

Taking another approach, Tyler (1983) explicitly differentiates the time a decision is made to that of the outcome. He further implies that there is a difference in the context between at the time of a decision and at the time of the outcome when he argues the following:

> *At the time of a decision not all the factors affecting the outcome of a decision can be known to a decision-maker, all decisions potentially have*

variety of different outcomes, and, no decision can guarantee either success or failure."

Tyler's comment also implies that the difference in contexts generates uncertainty in the outcome. Huber expands the potential effect of uncertainty in complex, ill-structured problems and implies that the decision-maker cannot know context of the outcome:

> *Whether or not a consequence occurs is not in the hands of the decision-maker, but depends on chance, nature, luck etc. The probability of various outcomes is more or less known to the decision-maker.* (Huber, 1995)

Although not explicitly stated by any one of the authors cited above, they all seem to be converging on the conclusion that time has an effect on the context of the outcome. Further, it can be inferred that uncertainty of an outcome is affected by context in which it is evaluated (Keys & Schwartz, 2007). As early as 1975, Fischoff understood that factors outside the context of a decision would affect the context of the outcome. He presented an example from the study of psychology:

> *A well designed therapeutic program may fail because of the tenacity of the client's problem or unanticipated and uncontrollable changes in the client's world. Thus, "good therapy" does not necessarily imply "good outcome". [Conversely,] many people who apparently benefit from treatment would have improved anyway, due to changes in their life*

circumstances or outlook. Thus, "good outcome" does not necessarily imply "good therapy. (Fischhoff, 1975)

From the previous arguments, it can be concluded that some of the uncertainty inherent in complex, ill-structured problems stems from the difference in the contexts of a decision and its outcome. This uncertainty due to context has two contributing factors: (1) uncertainty in the factors that comprise the context of a decision, and (2) the added uncertainty in the context of the outcome because of its displacement in time. The uncertainty due to the problem structure is derived from unknowable, unquantifiable, or unmodelable factors. For a given problem, some portion of the factors affecting the context of a decision will be in common with those in the context of the outcome. In less complex, well-structured, problems, the correspondence of these factors that comprise the context of a decision and its outcome should approach unity. Therefore, the level of uncertainty due to these common factors should impact a decision and its outcome more or less equally. On the other hand, the context of complex, ill-structured problems is harder to define than in well-structured problems; and the correspondence of these common factors between the decision and its outcome may be lower. Due to this lower correspondence, the structure of the problem may have unequal effects on the uncertainty in the contexts of a decision and the outcome. The possibility of these disparate effects on the contexts of a decision and the outcome leads one to conclude that the overall effect of structure is to potentially increase the uncertainty in complex, ill-structured problems more than that in less-complex, well-structured problems.

As was discussed above, the context of an outcome has some relation to the time it is evaluated. Conceivably, some of the uncertainty at the time the outcome is evaluated is due, in part, to a deviation from the original context caused by unpredictable factors (Fischhoff, 1975). A reasonable conclusion at this point would be that, as the time displacement between the decision and its outcome increases, the effect of unpredictable factors may cause the context of the outcome to diverge increasingly farther from the context of the decision. Since the uncertainty in the outcome is due, in part, to the lower correspondence between the contexts, then as the time difference between a decision and the outcome increases, the contexts have more opportunity to diverge, and the uncertainty in the outcome has the potential to increase. It follows, that the uncertainty of the outcome increases with both the complexity of the problem and the time displacement between a decision and the implementation of that decision. Therefore, for highly complex, highly ill-structured problems, such as military planning, the outcome could be distinctly disassociated from the initial decision.

Few other researchers address time in relation to the DOL. Keren and Bruine de Bruin (2003) are two of the very few who have mentioned the effect of time on the DOL when they proposed the question, *"What about decisions that…lie far in the future?"* (Keren & Bruine de Bruin, 2003). The effect of time on the DOL was only mentioned as a problem that needed to be addressed, and the bulk of their discussions centered on the pros and cons of the outcome and process approaches. Other research into the temporal aspects of decision-making typically focuses on the judgments under time pressure or sequential decisions. These studies suffer from the typical deficiencies associated with

research into complex, ill-structured problems. They tend to limit the complexity, limit the problems to well-structured ones, limit the time between the decision and the outcome, and limit the contextual factors of the decision and outcome in order to investigate one aspect of the decision. Unlike the laboratory setting, many real world temporal decision tasks are dynamic, requiring the decision-maker to choose a course of action under considerable time pressure; and the outcome is critically dependent on the decision-makers' subsequent actions (Brehmer, 1995). Examples of such tasks are fighting a forest fire, managing a patient in intensive care, fighting a battle, and managing a company, to name a few, dramatic examples. The commonly more limited laboratory experiments can replicate neither the complete contexts of these decision nor the temporal aspects associated with their outcome.

From the discussions above it seems apparent that neither the process-oriented nor the outcome-oriented approaches are adequate proxies for the direct assessment of decision quality in highly complex, ill-structured problems. The process-oriented approach relies on a rational process that can model the complexities and structure of a problem and relate these to the decision-maker's goals. However, the uncertainty in the structure and goals will generally preclude the creation of a sufficiently representative model of reality and this undercuts the assumption that the rational process will necessarily yield good decisions. The outcome-oriented approach, on the other hand, relies on the "goodness" of the outcome as a proxy for the quality of the decision, but typically for high complex, ill-structured problems, the uncertainty in the strength of decision-outcome linkage is not considered. A significant contributing factor to the

strength of the DOL, the temporal aspect, is rarely addressed in the literature. The DOL in highly complex, ill-structured problems in the real world is generally weaker than in experiential scenarios. The effects of the temporal aspect of the DOL are not well understood.

Evaluating Decision Quality

Consistent with the literature's lack of discussion and lack of consensus on a definition of decision quality, there is little consensus on methods to evaluate decision quality, specifically decision quality improvement as a result of using DSSs. The reason decision quality is difficult to assess and the general solution to the quandary of how to evaluate decision quality have been simply stated as:

> *It is very difficult to measure decision quality, as it is impossible to get inside people's heads and find out exactly what they are thinking. Instead of measuring decision making directly, therefore, it must be inferred from performance, whereby the decision-maker's actions are assumed to reflect the choice that they have made.* (Stanners & French, 2005)

Though simply stated, difficulties associated with assessing decision quality are plentiful. That these difficulties have resulted in a lack of progress in evaluating decision quality was noted as early as 1986 by (Aldag & Power, 1986), "Little research has been done to test the effects of computerized decision aids on the quality of decisions." Nevertheless, the evaluation of DSS has been discussed in the literature. The importance

of the development of a uniform and comprehensive scheme to measure and evaluate [DSS] effectiveness as identified as a major future research issue as late as 2003 (Forgionne, et al., 2003; Chen et al., 2011). The experts surveyed by Forgionne also identified the development of measures of decision quality as a significant challenge for the future.

An early effort to address the problem of evaluating decision quality with respect to DSS and complex problems was conducted by Keen & Morton who noted that "there is no best methodology to approach the problems, the criteria for choosing the best decisions are not clear" (Keen & Morton, 1978). But, they also opined that "hard" [objective] measures of decision quality (e.g., income, market share or the like) would be more accurate indicators of decision performance than "soft" [subjective] measures. This preference for "hard" measures demonstrates the bias toward outcome based evaluation that was, and still is, a primary assumption in most discussions of decision quality.

Shrada et al., in their 1988 survey of 13 studies on the effectiveness [used interchangeably with decision quality] of DSSs from 1970 to 1987, indicated that no two studies had the same definition of quality; and methods of determining effectiveness varied with each problem. Of the 13 studies, none met the criteria to be considered complex problems. Of note, 4 of the 13 studies (Joyner & Tunstall, 1970; Aldag & Power, 1986; King & Rodriguez, 1978; Cats-Baril & Huber, 1987) used problems that were to some extent ill-structured; and all four of these used subjective evaluation of the decisions by SMEs. Though not stated, this usage of subjective measures implied that,

even early in the research into decision quality, subjective evaluation was seen as more appropriate to the evaluation of ill-structured problems than objective evaluation.

In the same study, Shadra et al. demonstrated the bias toward outcome based evaluation in their study of a series of eight decisions quarterly. No effort was made to evaluate the quality of the individual decisions even though the use of "prevailing economic conditions" meant that the context changed for each decision; and the four objective measures used only assessed the outcome at the end of the string of decisions. Neither the quality of the individual decisions nor the varying decision contexts was discussed (Peters, et al., 2008).

The evaluation of Spatial DSSs (SDSSs) suffer from the same deficiencies as the evaluation of DSS in general. Typical of evaluations of SDSS (Dickinson & Calkins, 1988), proposed a general method of evaluating the decisions made with a SDSS which concentrated on the metrics time saved and error reduction. Time saved is a common metric cited as a benefit of using a SDSS and DSS in general, and error reduction is an early attempt to quantify the quality of the decisions made using a SDSS. Error reduction in this case was the reduction in misidentification of optimal placement of facilities based on geographic information. This problem probably does not meet the definition of a complex, ill-structured problem. It had a single best solution: the problem could easily be modeled, and a simple optimization could determine the optimal location given the data provided. Even though the problem they investigated was not complex or ill-structured, Dickenson and Calkins (1988) were unusual in their discussions about decision quality with respect to SDSS. They were concerned with the quality of the decision made with

the SDSS, whereas most evaluations of SDSS at that time were commonly concerned with the quality (accuracy) of the data retrieved and information presented. Much of the literature on SDSS evaluations continues to be focused on the accuracy of data retrieval and accuracy of the information presented, not on the quality of the decisions made (William R. King, 1983; Armstrong & Densham, 1990; Tarantilis & Kiranoudis, 2002; Frank, 2008).

Cats-Baril & Huber (1987) conducted an early study of the effectiveness of DSS for ill-structured problems. Their problem required the generation of a career plan and was evaluated using objective criteria (productivity), subjective evaluation by SMEs (quality of the plans), and subjective evaluation by the decision-makers (confidence in plan quality, satisfaction with the DSS, and change in attitude toward career planning and toward computers). Although the study used SMEs to subjectively evaluate decision quality, limited complexity of the study required relatively straightforward evaluations by the SMEs, and no decomposition of the decision quality characteristics was discussed. The study was successful in that it revealed that the use of a heuristic-based [modern] DSS and interaction with the DSS had positive effects on decision quality, productivity, and attitude toward computers, and negative effects on user confidence, satisfaction, and attitude toward the problem.

Crossland et al., (1995), in a more recent study, investigated the impact of a SDSS on a complex, well-structured problem. The experiment was structured as a between-subjects study with 142 college students as decision-makers. The independent variables were System and Complexity and dependent variables were Time and Accuracy. The

tasks required little judgment; and although the experiment did demonstrate the time required for the tasks was reduced and the accuracy of derived information was improved when using a SDSS, the experiment yields little insight into the effects of SDSS on complex, ill-structured decision-making.

More recently, Yates et al. noted that the methods employed to evaluate SDSSs had not progressed significantly:

> *Unfortunately, in many practical situations, there is little hard evidence that the techniques and devices, that is, decision aids,…have, in fact, yielded substantial, demonstrable improvements in how people decide.* (Yates et al. 2003)

Why have at least 25 years of experiments, evaluations, and assessments not yielded demonstrable improvement in how people decide? A clue to the cause may be in the definitions of decision quality. As we have seen, there are two main approaches to defining decision quality; the outcome approach and the process approach. Both approaches can be argued logically, but they also have fundamental flaws. The outcome approach depends totally on outcomes that are to some degree uncertain with respect to the decisions that produce them. On the other hand, the process approach acknowledges the uncertain relationship between a decision and the outcome, but ignores the outcome in favor of evaluating the quality of the process leading to the outcome. Both approaches have been shown to have drawbacks when applied to complex, ill-structured problems.

The methods used to evaluate DSSs are generally tied to a definition of decision quality. Proponents of the outcome approach evaluate decision quality based on the quality of the outcome and supporters of the process approach evaluated decision quality based on adherence to a specific decision model. Yet some decision research has yielded insight into how decision-makers view decision quality. Yates et al. (2003) demonstrated that in decision-makers' eyes, "*decision quality* is a coherent construct that extends far beyond the conception that is implicit in many decision aids and in much of decision scholarship generally." "The state of research into decision-making also highlights the need for a broader notion of a "good" decision than has been customary in decision research and suggests the shape that such a conception might take (Schneider & Shanteau, 2003). In order to find an effective method to evaluate DSSs, the current methods must be discussed.

In 1993, Frisch & Jones discussed five theoretical models currently in use to assess the quality of decision-making: utility theory, prospect theory, generalized utility theory, regret theory, and security-potential/aspiration theory. They noted that no single model was generally accepted and that consequently there was no accepted definition of decision quality or a standardized method to evaluate it. Yates et al. (2003) further summarized the four most referenced perspectives on evaluating decision quality:

- The *decision analytic* perspective emphasizes abstract rationality, such as consistency with the axioms of utility theory or probability theory (Baron, 1988; Dawes, 1988; Edwards et al., 1984).

- The *normative* perspective emphasizes the correspondence between a decision-maker's evaluation of an alternative and an evaluation based on a rule such as an additive value function (Payne, et al., 1988).

- The *decision process* perspective emphasizes adherence to processes that are *naturalistic* and that arguably ought to be expected to enhance the decider's satisfaction with chosen alternatives (Frisch & Clemen, 1994; Janis & Mann, 1977).

- The *accuracy* perspective emphasizes the distinction between decision utility and experience utility (Frisch & Jones, 1993).

Of the above perspectives to determining decision quality, the first two, decision analytic and normative, do not seem to be appropriate to ill-structured problems. The uncertainty concerning the structure of the problem and the intransparency of variables would make the determination of utility curves, probabilities, and an appropriate additive value model infeasible. If using the decision analytic perspective, one would be unable to model the problem sufficiently well to be able to determine whether the model was a realistic representation of the problem. Likewise, the inability to model the problem would mean that there could be no normative result to which to compare the actual outcome. The *decision process* and to some extent the *decision analytic* perspectives are process-oriented approaches that rely on the quality of naturalistic and rational processes respectively as measures of quality. Only the *accuracy* perspective is an outcome-based

approach, but this approach makes some attempt to describe the quality of decisions based the on correspondence between the expected outcome and the actual outcome. .

The *accuracy* approach, (Frisch & Jones, 1993), relies on two concepts: *decision utility* and *experience utility*. *Decision utility* refers to the decision-maker's evaluation of the potential benefits of an alternative at the time the decision is made. *Experience utility* refers to the actual benefits derived from the implementation of the decision, i.e., the quality of the outcome. The *accuracy* approach emphasizes the correspondence between decision and experience utility; the higher the correspondence the better the quality of the decision (Yates et al., 2003). According to Frisch and Jones, there are two explanations for the correspondence not being perfect: (1) The decision-maker does not perfectly predict the utility of the outcome. (2) The decision-maker does not take into account some factor *present at the time of the decision* which has an impact on the experience utility. Both of these explanations give subjective reasons the decision-maker failed to accurately assess the context of the decision, but they do not address the impact of a changing context. The context of the outcome is *assumed* to be the same as that of the decision, but this assumption is not assured with complex, ill-structured problems.

The *accuracy* approach, as presented in the literature, does not address an actual difference between the context in which the decision is made and the context in which it is implemented. Therefore, when evaluating the correspondence, context changes due to external factors are ignored. Although this approach highlights the difference between the quality of the decision and the outcome, it only addresses the differences in subjective utility, but not differences in the actual context between the decision and its

implementation. The authors do leave the definitions of decision utility and experience utility open enough that external changes in context could be incorporated into the *accuracy* approach to the evaluation of decision quality.

Given the various approaches to evaluating the quality of decisions, Yates et al., in their survey of the literature, distilled these approaches into criteria defining high quality decisions. Most of the definitions of decision quality focus mainly on one criterion, but most do acknowledge one or more of the other criteria. A high quality decision should meet one or more of the following criteria (Yates et al., 2003):

- The *aim* criterion: The decision meets the decision-maker's *explicitly* formulated aims (e.g. decisions reached using a DSS meet design, doctrinal, and/or mission requirements).
- The *need* criterion: The decision satisfies the *actual or implicit* needs of the beneficiary, needs that may not correspond to the decision-maker's aim(s) (e.g. decisions reached using a DSS actively support planning).
- The *aggregated outcomes* criterion: All of the actual outcomes of the decision, including ones beyond particular aims and needs, are better than the status quo or the beneficiary's aspiration level (e.g. the outcomes of the decisions reached using a DSS either improved a unit's tactical situation or improved the unit's tactical situation more than expected).
- The *rival options* criterion: The outcomes of the decision are superior to those that would have resulted from any and all available competing alternatives

(e.g. the decisions reached using a DSS resulted in the best possible improvement in the tactical situation).

- The *process costs* criterion: The costs of arriving at the decision are minimal [relative to the benefit] (e.g. time required to learn to use the DSS was minimal).

The *aim* criterion probably is the easiest of the criteria against which to evaluate a decision. It is also probably the most limited in determining whether a decision is good. This is the criterion commonly found in military contracts for the development of DSSs (Kadish et al., 2006). The construction of an evaluation to assess this criterion is relatively easy because there is documentation from which standards can be derived. With documented standards, the measures with which the quality of a decision will be measured are more likely to be objective as opposed to subjective. Subjective measures may be needed if the standards rely on terms like "better" instead of threshold acceptance values that can be measured objectively. Even though the evaluation of the quality of a decision may be relatively easy to determine using this criterion, there are two significant disadvantages to using this criterion: (1) There may not be documented standards against which outcomes may be compared. (2) Any documented requirements that do exist may not adequately reflect the actual needs of the user.

The *aggregated outcomes* and *rival options* are both based on using the relative benefits of outcomes to assess the quality of a decision. The need to assess the relative benefits of all other possible outcomes in order to evaluate a decision precludes using

these criteria to evaluate decisions for complex, ill-structured contexts for two reasons. First, since modeling ill-structured problems is extremely difficult, if not impossible, due to the uncertainty in the structure and the intransparency of the variables, evaluating the relative benefits of all the outcomes of an ill-structured problem would likely be impossible. Second, many decisions related to complex, ill-structured problems may never be implemented; and therefore outcomes cannot be compared. Both of these criteria are used extensively in decision and risk analysis (Clemen & Reilly, 2001) for problems that are well-structured.

Unlike the previous two criteria, the *need* and *process costs* criteria do not use the relative benefit of outcomes to evaluate the quality of decisions. Because both of these criteria are focused on the benefits of the decision, not on a comparison of outcomes, they can be useful in evaluating complex, ill-structured problems for which outcomes are not available for comparison. Likewise, the *need* criterion should be more useful in the evaluation decision quality in complex, ill-structure problems than the *aim* criterion since there are no established criteria of decision quality against which decisions can be evaluated.

Even though the *need* criterion seems appropriate for complex, ill-structured problems, there are obstacles to its use. Because the *need* criterion bases its evaluation on the *actual* or *implicit* needs of the beneficiary, developing "needs" could prove difficult. The *need* criterion will require the subjective judgment of SMEs in order to establish the needs of the beneficiary, i.e., the evaluation criteria. Since these needs are based on subjective judgment, for complex, ill-structured problems; it is unlikely that the

SMEs will be able to establish threshold values for use as evaluation criteria. Additionally, the evaluation of whether or not a decision meets these criteria will also require subjective evaluations. The probable lack of threshold criteria and the need for multiple subjective evaluations will make evaluating decision quality in complex, ill-structured problems against the *need* criterion difficult.

The advantage of using the *need* criterion is that it can support the evaluation of the only direct measure of the effectiveness of a DSS, which is the quality of the decisions made. As has been argued previously, the use of outcomes to evaluate the quality of decisions is fraught with potential pitfalls. The *need* criterion, on the other hand does not rely on outcomes, but on the evaluation of the benefits important to the decision-maker, specifically the quality of his/her decisions. In the context of complex, ill-structured problems for which outcomes may not be available for analysis, the availability of a criterion that directly evaluates decision quality and that does not rely on outcomes would be of much greater use than criteria that require multiple outcomes. Even though the implementation of an evaluation of decision quality using the *need* criterion will not be easy, this criterion would seem to support the direct evaluation of decision quality. Because it does support the evaluation of decision quality, the *need* criterion is probably the primary criterion for evaluating decision quality in complex, ill-structured problems.

Like the *need* criterion, the *process costs* criterion, slightly modified, can be of use when evaluating decision quality when using a DSS. The *process costs criterion,* as stated in Yates et al., would be of minimal value when evaluating a DSS since

establishing and evaluating a minimum cost threshold will most likely be an exercise in arbitrary guesswork for a SME. A more useful statement of the *process costs* criterion might be that the cost of arriving at a decision is minimal *with respect to the improvement in the quality of the decision*. With respect to using a DSS, this could be restated as would the decision-maker use the DSS to a make his/her decision? Evaluating this version of the *relative process costs* criterion would certainly require a subjective evaluation on the part of the decision-maker but would not require setting an arbitrary cost standard.

No matter what basic criterion or criteria are used to evaluate DSSs, there seem to be general aspects of the evaluation of DSSs that need to be resolved. As a result of their study of the evaluation of DSS, March & Smith, (1995) describe the general purpose of evaluating DSSs as the evaluation of "operationally (the ability [of the DSS] to perform the intended task or the ability of humans to effectively use the [DSS]." This purpose seems to be a combination of the *aim* and the *need* criteria. In commenting on the lack of effective methods to evaluate DSSs, they further note that the "Methods for this type of evaluation are *not* unlike those for justifying or testing theories. However, the aim is to determine 'how well' an [DSS] works, not to prove anything about how or why the [DSS] works." Their implicit conclusion seems to be that most current evaluations of DSSs are not as rigorous as methods used to test theories. They go on to describe some aspects of basic experimental design that should be incorporated into DSS evaluations:

> *Once metrics are developed, empirical work may be necessary to perform the evaluation. Constructs, models, methods, and instantiations must be*

exercised within their environments. Often this means obtaining a subject group to do the exercising. Often multiple constructs, models, methods, or instantiations are studied and compared. Issues that must be addressed include comparability, subject selection, training, time, and tasks. (March & Smith, 1995)

Of the elements of experimental design mentioned by March and Smith, they pay particular attention to the importance of the criteria (metrics or measurements) used to evaluate DSS. Specifically, they imply that the lack of good metrics actively hinders the effective evaluation of DSSs.

Evaluation requires the development of metrics and the measurement of performance according to those metrics. Metrics define what we are trying to accomplish. They are used to assess the performance ....Lack of metrics and failure to measure DSS performance according to [an] established criteria result in an inability to effectively judge research efforts. (March & Smith, 1995)

The difficulty in generating these criteria due to the unique nature of each DSS was stated by (von Winterfeldt, 1980), **"**various classes of decisions can be defined, each requiring different judgment criteria." March and Smith agree:

Not only must a system be evaluated, but the evaluation criteria themselves must be determined for the system in a particular environment. (March & Smith, 1995)

The importance of the criteria which DSSs used to address complex, ill-structured problems cannot be overstated. Criteria that do not directly address decision quality are likely to result the in mixed results (Sharda, et al., 1988).

Assuming that effective criteria appropriate to complex, ill-structured problems can be defined, acquiring data that can be used to analyze decision quality has its own difficulties. de Silva et al. (2003) pointed out the following:

> *For decision-making events that occur frequently, the consequences of instances of decisions taken without the aid of the [DSS] can be compared to those instances of decisions taken with the aid of the [DSS]. For decisions with frequent instances, historical data may be available for evaluation. For decisions without frequent instances there is no historical data and validation of its output…becomes extremely difficult….[comparing with existing DSS] is not a satisfactory method of validation, as other tools that are compared with the [DSS] may not have the same output functions or decision support goals as the [DSS].* (de Silva et al., 2003)

The problems associated with acquiring data are even more complicated in a military context. Bolia et al. addressed the general concept confronting the evaluation of decision quality with respect to a DSS and related it to a military environment:

> *…the quality of the decision is measured by the quality of the immediate outcome. If the quality of the immediate outcome is*

> *measureable, the problem is solved. On the other hand, if there is no immediate outcome, or the immediate outcome is not itself measurable, then we are not any closer to a solution. In complex environments such as the battlefield, both situations are likely to occur.* (Bolia, et al., 2004)

Here, Bolia et al. make several explicit comments and implications on the evaluation of decision quality in complex, ill-structured problems. First, although Bolia et al. are using the outcome approach to evaluate decision quality, by using the qualifier "immediate" they are confirming the contention that time affects the ability of decision quality to be assessed using an outcome. The implication is that outcomes that are not immediate will not be useful in evaluating decision quality. Second, in addition to the effect of delayed outcomes, they also identify a significant problem that is commonly confronted when evaluating military planning problems: the plan may never be executed. In this case, any argument for evaluating decision quality using outcome breaks down. Third, the possibility that the quality of the outcome may not be measurable implies that military planning problems are likely ill-structured.

Since actually executing military plans in a rigorous experimental setting so that immediate outcomes can be determined is somewhat problematic, some other standard must be determined by which decisions can be judged. Surrogate metrics for military mission accomplishment have included loss ratios, casualties inflicted, area taken, or other quantifiable results of military action (Hayes & Wheatley, 2001).

Instances of this type of data collected from the actual execution of plans would be anecdotal at best. The contexts of individual plans will be dissimilar enough that the

data from individual outcomes would be suspect. Data of this type could be obtained from combat simulations based on plans generated by military planners. To generate data of this type sufficient to yield statistically significant results would require, at a minimum, multiple runs of simulations designed to stochastically evaluate the results of the many possible series of friendly decisions and enemy counter decisions. For highly complex, ill-structured military planning problems, just modeling the structure of the problem to account for all possible unit actions and enemy interactions would be prohibitive in time and cost. Even if simulations could be used to generate data relative to mission effectiveness, this is not the same as determining decision quality (Bolia, et al., 2004). As discussed before, the many uncontrollable factors which impact the execution of military plans cause the relationship between the quality of military decisions, as evidenced by military planning, and the outcome of the subsequent operations to be statistically noisy.

Summary

There is extensive literature on research into decision-making, methods for evaluating the effectiveness decision-making, and evaluating decision quality. Research on these topics can be found in many disciplines. Nevertheless, there has been limited research into decision-quality in highly-complex and ill-structured problems, into using decision-quality as a primary evaluation measure, and into evaluating decision quality for this type of problem. A significant factor in this lack of research into the evaluation of decision quality in complex, ill-structured problems is due to the nature of the problems themselves.

Although complex, ill-structured problems are the problems on which much current research into decision quality is focused, there are no standard definitions of either complexity or ill-structure. Many characteristics that contribute to problems being complex and ill-structured have been identified and discussed, but there is no clear consensus on the impact of these characteristics on decision-making and decision quality. This lack of a consensus and the subsequent difficulty in defining standards of decision quality has led to approaches for assessing decision quality that evaluate proxies for decision quality instead of the actual quality of decisions. None of these methods have been used to directly assess decision quality in complex, ill-structured problems.

The most common approaches to evaluating decision quality, the process and outcome-based approaches, do not seem to be appropriate for assessing decision quality in complex, ill-structured problems. The outcome-based approach relies on a clear and direct relationship between the decision and the outcome of its implementation. Unlike simpler, well-structured problems, this direct decision-outcome linkage is not clear for complex, ill-structured problems. The factors identified in the research that contribute to problems being classified as complex and ill-structured are also those factors that prevent a clear and direct linkage between decisions and their outcomes in this type of problem. This lack of a clear and direct linkage between a decision and its outcome makes the outcome-based approach of little use for evaluating decision quality in this class of problem. In opposition to the outcome-based approach, the process approach acknowledges that there is uncertainty in outcomes and, therefore, focuses on the fidelity of the procedures used to arrive at decisions. Since the process-based approach asserts

that the quality of decisions depends on the quality of the decision-making process, it makes no attempt to actually determine the quality of decisions generated by a given process. Because of weaknesses in each approach, neither is appropriate for assessing decision quality in complex, ill-structured problems.

Since the process and outcome-based approaches are inadequate, a method that directly assesses decision quality is necessary to assess the decision quality in complex, ill-structured problems. Even though there is much literature discussing the evaluation of decision quality, discussions, experiments, and evaluations primarily rely on using various proxy-based methods based on the outcome approach; and there is little discussion of the direct evaluation of decision quality. This lack of use of the direct assessment of decision quality seems to be due to two factors. First, there are no standards of decision quality against which decisions can be compared; and, second, for complex ill-structured problems, there is no single best decision that could provide a reference against which potential decisions could be compared. The lack of either a standard or a problem-specific base-line for comparison suggests that the evaluation of decision quality must rely on a subjective assessment of the characteristics of decision quality using problem-specific criteria, and that the evaluations of decision quality would only be useful for comparing the relative quality of decisions made in the same context.

The research presented in the remainder of this thesis is the development and assessment of a method, the Decision Quality Evaluation Method (DQEM), which overcomes the difficulties associated with outcome- and process-based approaches to the evaluation of decision quality in complex, ill-structured problems. The DQEM uses a

strategy of extensively decomposing decision quality characteristics to generate measures that can be used to directly measure characteristics of decision quality. The method further defines procedures for the subjective assessment and scoring of decision quality characteristics by SMEs. These procedures, together with procedures for aggregating the scores for individual decision quality characteristics, define a method that can be used to consistently and reliably evaluate relative decision quality in complex, ill-structured problems. Although each decomposition of decision quality, scoring procedure, and aggregation is specific to the type of problem addressed, the usefulness of these procedures is demonstrated in the assessment of relative decision quality for two different complex, ill-structured problems.

CHAPTER THREE: THE DECISION QUALITY EVALUATION METHOD

Overview

The Decision Quality Evaluation Method (DQEM) is a method designed to directly evaluate decision quality in complex, ill-structured problems using a structured, subjective approach. It is based on the principles of multi-criteria decision analysis (MCDA) and extends and adapts those principles for the subjective evaluation of decision quality. The primary difference between traditional MCDA and the DQEM is that unlike MCDA, the DQEM is a tool for decision evaluation not for decision-making. Although the two uses are closely related, the DQEM is separate from the decision-making process and evaluates the decisions produced through the use of a decision-making process. Since there is no standard of decision quality (discussed in chapter 2) against which to compare a single evaluation of decision quality, the DQEM's primary usefulness lies in improving SMEs' ability to discriminate among the quality of different decisions. Improving this ability also implies an improvement in the SMEs' ability to evaluate the relative quality of decisions. Also, because it is not part of the decision–making process, SMEs can use the DQEM to use the observed differences in decision quality to evaluate the benefits of changes to that decision-making process e.g. Case Studies One and Two.

Because outcome-based and process-based evaluation of decision quality in complex, ill-structured problems is problematic, in order to assess decision quality it is necessary to directly measure the quality of the decisions made. Typically, when decision quality is assessed, a single score that represents the overall quality is generated (Schweiger & Sandberg, 1989; Adelman, 1992; Amason, 1996; Hough & Ogilvie, 2005). Complex, ill-structured problems, such as military planning problems, are so complex that a single judgment cannot serve as the evaluation of the entire decision or series of decisions, i.e., a plan. A single score cannot possibly encompass the complex assumptions, estimates, and judgments that were made when making the decision. Even so, an overall score representing the quality of a decision is usually desired as a means of comparison: and the score must reflect the individual judgments and decisions that comprise the overall plan. In order to generate such an overall score of decision quality, the DQEM uses a structured decomposition of the subjective characteristics of decision quality and a tailored scoring procedure to directly develop quality scores for individual decision characteristics. These characteristic scores are then aggregated into an overall score that can be used for comparison with the overall scores of other decisions made to address the same problem.

Although there is no generally accepted definition of decision quality (Yates et al., 2003), it is generally accepted that decision quality can be decomposed into characteristics that describe that decision (Yates et al., 2003; Schneider & Shanteau, 2003; Frisch & Jones, 1993). This principle is fundamental to MCDA and to the ability to directly evaluate the quality of decisions. Complex decisions, including military

planning, are the result of the analysis of many factors; so many factors that without an evaluation structure, an evaluator cannot remain fully cognizant of each factor's impact on the decision (Serfaty, et al., 1997). The decomposition of the decision quality allows evaluators to focus on the simpler characteristics that correspond to specific judgments within the overall more-complex decision. The decomposition also establishes common factors that guide the evaluators' assessments instead of relying solely on individual evaluators to comprehensively evaluate all the characteristics of the decision.

In the literature, decomposing problems into decision quality characteristics is part of the analysis of decision quality (Schweiger, et al., 1989; Peters, et al., 2008; Chen, et al., 2011); but this decomposition is generally limited in scope and is limited to easily measurable (objective) characteristics. This lack of decomposition seems to be primarily due to the problems in the literature being of lesser complexity than the problems encountered in areas such as military planning; and this lesser complexity does not require extensive decomposition in order to capture the entirety of the problems or decisions. The DQEM, on the other hand, relies on the decomposition of decision quality into a detailed hierarchy of decision quality characteristics, sub-characteristics, and measures in order to capture the myriad relationships among the elements of decisions that are typical of complex, ill-structured problems.

MDCA, the basis for the DQEM, has been defined as an aid to decision-making through a process which seeks to integrate objective measurement with value judgment (Belton & Stewart, 2002). Further, MCDA seeks to make the need for subjective judgments explicit and transparent. DQEM differs from this definition in several ways:

(1) It is a decision quality evaluation tool not a decision aid. (2) It does not so much model the decision as model the characteristics of a good decision. (3) It does not attempt to integrate objective measurement; although since its basis is in MCDA, any MCDA techniques could easily be integrated into the DQEM. (4) The DQEM does not just manage subjectivity or identify what subjective judgments need to be made. It is expressly designed to capture SMEs' subjective understanding of a problem and make use of their judgments as part of the evaluation process.

Typically, MCDA techniques are employed by decision-makers to analyze decision options and arrive at the best possible decision. One of MCDA's strengths is that part of the MCDA process facilitates decision-makers' learning about and understanding the problem; about organizational priorities, values, and objectives; and through exploring these in the context of the problem guides decision-makers in identifying a preferred course of action (Belton & Stewart, 2002). The DQEM makes use of this strength by incorporating portions of MCDA techniques that enhance understanding the impact of the elements of a problem and the relationships among these elements. In contrast to the typical use of MCDA, the DQEM employs some of these techniques to evaluate decisions that have already been made. The same techniques that allow decision-makers to analyze options allow evaluators to differentiate among the quality of decisions. And, because the DQEM evaluates previously made decisions, the SMEs who evaluate the decisions can be independent of the decision-makers who made them.

When decision-makers use MCDA techniques, many of these techniques rely on objective measures to model the decision. Obtaining objective measures generally requires eliciting numeric values from decision-makers that are intended to quantify the impact of the elements of the decision and the relationships among them on the decision. The numeric values obtained are models of the decision and are used in various manipulations designed to predict the preferred course of action. The DQEM, on the other hand, is used after the decision has been made; and since no manipulations are required to estimate the impact of various options, the DQEM focuses on capturing the subjective understanding of the problem and not on quantifying that understanding. Concentrating on the subjective relationships does not preclude the use of objective measure in the DQEM; because the DQEM is based on MCDA techniques, objective measures can easily be incorporated into the primarily subjective scoring process of the DQEM.

The DQEM was developed in response to the general need for a method to directly address decision quality as exemplified in Peter et al. (2008). The DQEM is general enough to be applied to the assessment of decision quality in a variety of contexts including the assessment of the impact of Decision Support Systems (DSSs) and the evaluation of decision-making processes. The DQEM was developed and first applied in a series of experiments to evaluate the usefulness of Geospatial Decision Support Systems (GDSSs) in military planning problems. Since the GDSSs were to be evaluated within the context of the Military Decision Making Process (MDMP), the goal of the experiments was to evaluate the effect of modifying the MDMP by adding the GDSSs to

the process. The need for the DEQM became apparent because, although the problems were to be real military planning problems using real decision-makers and real GDSSs, none of the decisions would ever be implemented; and no outcomes would be generated to use as proxies for decision quality. Therefore, the evaluation of decision quality had to be made using the output of the MDMP, and a method was needed to directly evaluate the quality of the decisions made by the decision-makers.

The development of the DQEM was conducted in three stages: (1) the development of a general method based on decision-making theory and MCDA evaluation methods, (2) the development of the scoring procedures that incorporated tailored Likert scales, and (3) the evaluation of the DQEM in military planning scenarios evaluating the change in the MDMP through the employment of GDSSs. Each of these stages is described in the following sections.

Development of the DQEM

The DQEM adapts accepted multi-criteria decision analysis (MCDA) procedures for use in the evaluation of decision quality taking into account constraints and factors peculiar to directly evaluating decision quality instead of using proxies as measures of decision quality. Specifically, the DQEM goes beyond simple application of MCDA techniques found in the literature that are used to evaluate decision quality: it extends these techniques by incorporating methods that specify the subjective hierarchical decomposition of decision quality characteristics into sub-characteristics, the development of measures used to subjectively assess each sub-characteristic, and the derivation of detailed scoring criteria for each measure. The DQEM also specifies the

use of Likert scales tailored to each measure that captures the relative importance of each criterion in the evaluation of the sub-characteristic. These methods allow Subject Matter Experts (SMEs) to directly assess the quality of the decisions for use as the primary measure instead of relying on evaluating outcomes or process.

Additionally, the DQEM approach to the assessment of decision-making in ill-structured problems is appropriate for the assessment of decision quality in problems that are more complex than those that are commonly assessed in the literature. The literature on the assessment of decision-making typically considers only well structured problems: problems with a high probability of a unique best solution (Crossland et al., 1995). When a best solution exists, the decision quality characteristics defining the best solution are usually easily definable, and measures of these characteristics can be easily constructed and compared. These measures are generally objective and relatively straightforward to quantify. Due to the relative ease of analyses using objective measures, the tendency is to reduce the problems addressed to well-structured problems for which objective measures are appropriate. Reducing problems into ones that are well-structured, as discussed in Chapter 2, usually means that the complexity of the problems is also reduced. The goal of the DQEM is to aid in the assessment of decision-making in complex, ill-structured problems without the need to change the structure or complexity of the problem.

In order to meet the goals of the DQEM, it was developed using the basic principle of MCDA. Theory behind MCDA is designed to address decisions for complex, ill-structured problems and though there is little discussion of uses of MCDA

other than as a decision aid, nothing precludes the use of MCDA techniques for the evaluation of decision quality. The DQEM, specifically the decomposition of decision quality characteristics and the development of scoring criteria, was developed by combining and modifying two of the schools of thought in MCDA:

1. Traditional *value measurement models* require the use of a decomposition in which numerical scores are constructed in order to represent the degree to which decision options may be preferred to others. Scores are developed initially for each individual criterion and are then synthesized in order to effect aggregation into higher preference levels (Belton & Stewart, 2002).
2. In DQEM, the decomposition does not attempt to model the decision; and as such, the decomposition is based not on possible options but on determining the characteristics of good decisions. Also, scores in the DQEM are generated from the subjective evaluation of decision quality; and the aggregation of scores represents overall decision quality not an overall option preference.
3. Traditional use of *goal, aspiration, or reference levels*; a method in which desirable or satisfactory levels of achievement are established for each of the criteria. This process then seeks to discover options that are in some sense closest to achieving these desirable goals (Belton & Stewart, 2002). The DQEM does not use levels of achievement to choose options, but when combined with scoring criteria, it uses levels of achievement to determine the impact a given criteria has on the "goodness" of a decision.

The DQEM incorporated and adapted these schools of thought as part of the decomposition and scoring processes. Unlike value measurement models, the decomposition in the DQEM does not attempt to model the decision; and as such, the decomposition is based not on possible options, but on determining the characteristics of "good" decisions. Since scores in the DOEM are generated from the subjective evaluations of these characteristics, the aggregation of scores represents overall decision quality not an overall option preference. Likewise, the DQEM doesn't uses levels of achievement of goals, aspiration, or reference levels to evaluate options but combines levels of achievement of evaluation criteria with scoring criteria to determine the impact of criteria on the "goodness" of a decision.

The decomposition of decision quality characteristics used in the DQEM is based on general MCDA methods that include the development of a value hierarchy. Like most MCDA methods, the DQEM uses a structured hierarchy. However, for DQEM, the value hierarchy serves a different purpose. In MCDA, the value hierarchy models the *decision-maker's values*, and is used to help the decision-maker choose an alternative that best achieves his values. In DQEM, the decision quality hierarchy models *decision quality criteria defined by SMEs*, and is used to evaluate the quality of a decision after it has been made. This capturing of the SMEs' subjective understanding can be accomplished because the decomposition in the DQEM is more detailed than is usually found in the MCDA literature. The use of a decomposition in many MCDA techniques generates conflicting goals: on one hand, a more detailed decomposition can more accurately model the problem; but on the other hand, a more detailed decomposition implies that there are

more relationships that must be quantified which would require eliciting more data from SMEs in an attempt to *quantify* the problem's subjective characteristics. Conversely, limiting the decomposition would reduce the difficult task of eliciting data from SMEs and could also limit the model's accuracy. Since the DQEM does not attempt to quantify subjective relationships, there are no competing concerns and there is no impediment to decomposing the decision quality characteristics to the point that the decision is fully characterized with respect to the decision to be made.

Like the decomposition in the DEQM, the scoring procedure is based on a modified MCDA process. MCDA methods that use the achievement of goals, aspiration, or reference levels to score options were adapted to allow SMEs to use levels of achievement of multiple criteria to evaluate decision quality. This ability to use multiple criteria to evaluate decision quality in conjunction with a detailed decomposition can be used to analyze the impact of differences in the quality of specific sub-characteristics as well as evaluating the quality of the overall decision. Through their use in the decomposition and scoring procedures, the principles and methods of MCDA provided a good basis for the development of the DQEM.

Decomposition of decision quality

The decomposition of decision quality and the identification of decision quality characteristics is at the heart of using the DQEM to assess the quality of decisions. The decomposition is required because of the basic characteristics of complex, ill-structured problems. The large number of elements and relationships among them make understanding the impact of specific elements and relationships on the overall decision

difficult without some form of guidance. The decomposition serves three functions with respect to evaluating decision quality: (1) The decomposition provides a structured means to examine and capture SMEs' subjective understanding of the impacts of the elements and their relationships on the quality of the decision. (2) The decomposition, once complete, provides a source of concise guidance for the evaluation of decision quality by independent SMEs. (3) The decomposition provides a starting point for the development of scoring criteria. The first function, providing a structured means to explore the problem, is taken directly from the MCDA; and it is especially necessary since capturing subjective relationships is the core of the DQEM. The second function, providing evaluation guidance, uses the captured subjective relationships to focus SMEs' evaluations on the most important elements and relationships in the decision and allowing SMEs who did not participate in the decomposition to evaluate the decisions. Finally, the third function, providing the basis for developing the scoring criteria, allows the development of scoring criteria and tailored Likert scales that will translate the SMEs' evaluations into numeric scores. All three functions are important to the development of the DQEM.

The decomposition of decision quality characteristics should be done in consultation with SMEs and use whatever documented guidance is available. Since decision quality decomposition is problem specific, it is necessary to recruit SMEs who understand the impact of the elements of the decisions and the relationships among them on the decisions being evaluated. The DQEM decomposition is designed to capture the knowledge and understanding of the SMEs with respect to the problem under evaluation.

The SMEs need not rely only on their memory and judgment, and they are encouraged to use appropriate reference materials to guide the decomposition. Reference material such as published research, in-house studies, and design documents may be used. For military decision-making, which is the subject of both case studies, specific guidance is available on the general application of the MDMP; and guidance on decision-making for specific problems can be found in the published doctrine, situation-specific standard operating procedures (SOPs), and Tactical Training Plans (TTPs). Doctrine usually defines the basic factors that contribute to a decision and SOPs/TTPs provide specific domain considerations. Even if few specific decision quality characteristics can be derived from sources like these, such sources can provide insight into the subjective characteristics of decisions. Subjective characteristics are characteristics that require decision-makers to exercise judgment, and this exercise of judgment is sometimes referred to as the "art" of decision-making. Identifying the decision characteristics used in this art and that are pertinent to specific decisions requires the experience and expertise of SMEs. Once the characteristics of a decision are identified, the most germane of them should form the basis of the decomposition of decision quality.

There are no hard and fast rules on how detailed the decomposition of decision quality should be. Each problem and the decisions associated with it are unique; and because of this, each decomposition will be unique. The goal of the decomposition is to generate a characteristic hierarchy that captures the SMEs' understanding of the impact of the elements and relationships on the decision. The decomposition decomposes decision quality characteristics into multiple, more finely grained sub-characteristics and

measures until the individual measures can be evaluated straightforwardly and independently of other measures. One possible structure for the decomposition and one that is common in MCDA is a hierarchical tree structure. A tree structure has significant advantages for the decomposition, the evaluation, and aggregation of decision quality. Ideally, the goal is to have decision quality sub-characteristics that support only one, more general, characteristic. This relationship between parent characteristics and their child sub-characteristics allows each decision characteristic to be decomposed and scored independently of other characteristics. Although determining the combined effect of a child sub-characteristic that supports more than one parent characteristic on the overall decision quality is possible using other structures, it is more difficult than if a tree structure is used. And since many successful MCDA decompositions are tree structures, the decomposition used in the DQEM is designed to result in a tree structure. The core of the DEQM decomposition is the use of a consensus of SMEs' opinions to define the decision quality sub-characteristics and measures. The decomposition is based on the defining child sub-characteristics that directly support the evaluation of the characteristics from which they are descended. Because the SMEs will not at first agree completely, the decomposition process is iterative. Sub-characteristics may be created, revised, moved in relation to parents, or discarded during the decomposition. In this, the DQEM process is identical to many MCDA decomposition processes; but the focus of the process is very different from that employed by MCDA. The DQEM decomposes the qualities of "good" decisions and focuses on the subjective relationships between parent

characteristics and child sub-characteristics and among the children instead of estimating the potential impact of decision factors and options.

In the DEQM decomposition, special care should be taken not to include child sub-characteristics that do not directly support the parent characteristic. Since the characteristics that combine to form a "good" decision are very subjective, any sub-characteristic that does not support the evaluation of the parent characteristic will introduce statistical noise into the evaluation. As discussed in Chapter Five, the inclusion of extraneous sub-characteristics adversely affected the statistical significance of the results generated for the sponsor and may have had an impact on the assessment of the DQEM.

Because the decomposition of decision quality is primarily subjective, an underlying assumption of using a consensus is that the SMEs are experts in the specific decision to be evaluated. As noted in the MCDA literature, a decomposition can be used to identify and resolve bias in the opinions of the SMEs; and since one goal of most MCDA techniques is to identify subjective aspects of a decision, extensive decomposition is sometimes used to resolve conflicts of opinion. Since the DQEM always uses a detailed decomposition, the decomposing is likely to identify differing SME biases. In a third experiment (not included in this research), the decomposition identified irreconcilable biases among the three SMEs. In this case, the DQEM identified SME biases that were fundamentally different due to significantly different experiences. In this case the differences in their biases were significant enough to require redefining

the experimental problem. Conversely, in the two case studies in this research, the SMEs were able to resolve the differences in their biases discovered through the decomposition.

Given that the differences in the SMEs biases were not great enough to require redefining the problem, how does one know when the decomposition is complete? The decision quality characteristics should be decomposed into as many levels of sub-characteristics as necessary to define the overall decision quality characteristics. As the decision quality characteristics are decomposed, the sub-characteristics become more specific; the granularity of each level is finer than the one above. That final stage of decomposition consists of those sub-characteristics that can be individually assessed with easily evaluated measures. In the case studies, if the SMEs' answer to the question "how can this characteristic be measured" was a list of attributes that could be easily evaluated (in the judgment of the SMEs), then the decomposition of that sub-characteristic is complete. If, on the other hand, the list of attributes contains attributes that could not be easily evaluated (were in reality additional sub-characteristics); then further decomposition was needed. The list of attributes developed for each sub-characteristics are called the measures, and these attributes are the factors that are actually assessed to evaluate decision quality. This process of decomposition and measure identification was repeated iteratively until the SMEs reach consensus on all sub-characteristics and their supporting measures. Though reaching a final decomposition may seem cumbersome, in both case studies, the SMEs were able to reach consensus on the decomposition of the quality of decisions in a relatively short period of time. A partial decomposition from Case Study One is presented in Table 1 and Table 2 provides a numerical comparison of

the decomposition of decision quality from the case studies that gives a rough idea as to the relative complexity of the hierarchical decomposition trees.

Table 1: Example Decision Quality Characteristic Decomposition

Partial Decision Quality Decomposition from Case Study 1	
Decision Quality Characteristics	**Rationale**
Quality of the Routes	A primary goal of terrain analysis is to generate routes that are suitable for the movement of units through the given terrain. The quality of the routes generated will affect the selection of the recommended AoAs and the unit movement plan
Sub-characteristic	
AoAs take a direct route from phase line to phase line	
Measure	**Evaluation Criteria**
Valid start point	Yes/No: The start point is behind Phase Line X? Is there good route from the AA to start point?
Valid end point	Yes/No: The endpoint is beyond Phase Line Y? There is a good route from the endpoint to the objective?
No unnecessary turns	There are no unnecessary turns and there are no turns of $> 45^{o}$ when a straighter route is available?
Independence of routes	Routes have minimal MC is common (exception for first egress MC from AA).
Analyzed for both on- and off-road	At least one route must be analyzed for each on-road and off road. Optimally, at least one route in each BN AOO should be analyzed for each case. Route analysis is appropriate for the route.
Generated travel times for all route	Valid travel times for three vehicles generate for each route (correct vehicles, entire route).

A completed decomposition serves two purposes. First, it provides guidance to independent SMEs for the evaluation of the decision for which it was created. The

complex, ill-structured problems for which the DQEM is intended have large numbers of elements that have uncertain relationships; and these problems have enough elements and uncertain relationships that an unassisted humans cannot reliably evaluate the impact of all the elements and the associated relationships on the overall decision quality. Second, the decomposition of decision quality forms the basis of this guidance to evaluators and leads to the development of evaluation criteria. The potential effect of the guidance provided by the combination of the decomposition and the evaluation criteria was considered crucial enough that the within-subject correlations discussed in the analyses section below are used as a measure of the degree to which the decomposition reflects the SMEs' understanding of the problem.

Table 2: Comparison of Case Study Decompositions

CS-2 Decision Quality Decomposition Comparison		
Case Study	CS-1	CS-2
decision quality characteristics	5	6
1st level sub-characteristics	13	22
2nd level sub-characteristics	16	25
3rd level sub-characteristics	0	4
measures	35	53

Evaluation of decision quality

In order to assess the quality of the decision quality sub-characteristics, measures and criteria must be developed that can be used to evaluate the quality of the each sub-characteristic. The generation of the measures was discussed above and identified the measures as sub-characteristics that could be independently and easily evaluated. In the course of Case Study One, the use of the measures alone did not capture the SMEs understanding of the problem sufficiently well enough for the SMEs to be able to discriminate among small gradations of decision quality. In response to this, a second set of SMEs, independent of the SMEs who generated the decomposition, were asked to first develop evaluation criteria that would support the evaluation of each measure and then to develop scoring criteria that would identify the relative importance of each measure to the quality of the associated sub-characteristic. The evaluation criteria can be considered an extension of the decomposition that further clarified the SMEs' understanding of what constituted "good" quality with respect to each measure and the individual sub-characteristic. In effect, the evaluation criteria were the most detailed attributes upon which the measure would be evaluated. The scoring criteria, on the other hand, were generated by these SMEs to translate the level of achievement of specific measures to numeric data using tailored Likert scales.

Evaluation criteria

The development of evaluation criteria was conducted in conjunction with the development of the scoring criteria and tailored Likert scales (discussed below). The development of evaluation criteria is an extension of the decomposition and was

conducted because a second set of SMEs, who were developing the scoring criteria and Likert scales, determined that an additional level of decomposition would aid in the evaluation of the measures. Table 1 provides an example of decision quality measures and evaluation criteria that were used to evaluate the decision quality of the sub-characteristics in Case Study One. Sub-characteristics may be assessed using one or more measures, and each measure may have one or more evaluation criteria associated with it. The number of measures and evaluation criteria is determined by the SMEs and reflects their understanding of the important attributes that would characterize the sub-characteristic as "good."

Although all the measures discussed in this research were subjective, the evaluation criteria generated to support analyzing the decision quality measures can be assessed either objectively or subjectively. Even though all the measures were developed through subjective decomposition, there may be some evaluation criteria that can be categorized, counted, or have binary answers that take little subject knowledge to be able to evaluate. If an evaluation criterion has a single best value, then it may be able to be assessed objectively. The assessments of the objective evaluation criteria were treated identically to the assessments of the subjective evaluation criteria when assessing the quality of individual measures.

For example, in Table 1, criteria with yes/no responses can be measured objectively by SMEs who understand the criteria. Subjective evaluation criteria are those that require SME judgment to assess properly. Any criterion that requires an assessment of what is good, better, or appropriate requires subjective assessment. Individual

measures may have criteria requiring a combination of objective and subjective assessments. For example, the assessment of a *valid endpoint* requires an objective assessment (beyond Phase line X) and a subjective assessment (good route from the endpoint to the objective). Decision sub-characteristics that have both subjective and objective measures will require overall subjective assessment of the criteria when evaluating the quality of the measure. For example, since the *valid start point* and *valid endpoint* measures have both objective and subjective sub-measures, the overall assessment of both of these sub-characteristics would be subjective.

The generation of evaluation criteria is critical to the evaluation of decision quality. Like the sub-characteristics and measures, care must be taken to ensure that the decision quality evaluation criteria relate directly to the decision quality measures they support; and they must be detailed enough to be easily assessed, to avoid confusion, and to avoid the inclusion of extraneous criteria in the evaluation of decision quality

Scoring criteria and tailored Likert scales

Like the development of the characteristics, the evaluation of characteristic measures to form a coherent evaluation of the overall decision quality is primarily a subjective process. Although objective evaluation criteria can be incorporated into the evaluation, the evaluation of measures of subjective decisions is generally subjective and a means is needed to translate the SMEs' subjective evaluations to numeric scores. The DQEM, in addition to providing a procedure for defining subjective evaluation measures and evaluation criteria, also provides a procedure for structuring the subjective scoring of the decision quality by SMEs so that meaningful results can be obtained. This procedure

consists of methods of determining the SMEs' understanding of the measures under evaluation, reconciling SMEs' concepts of what constitutes "good" and "poor" for each measure in the context of the current decision, and developing scoring criteria that are used to construct Likert scales tailored to each sub-characteristic. This procedure constitutes a significant expansion on the methods available in the literature to evaluate decision quality.

Since the DQEM is primarily concerned with the subjective evaluation of decision quality, a means of translating the SMEs' subjective assessments to numeric data is needed so that the evaluations of diverse measures can be aggregated into an overall score that represents the overall quality of the decision. Likert scales are a well-documented method of quantifying subjective responses, but the typically constructed Likert scales have a significant drawback with respect to the subjective evaluation in the DQEM. The standard Likert scales introduce variation due to scale interpretation. Likert scales commonly use a five-point scale ranging from 1 to 5 where, in general, 1 is poor, 3 is fair, and 5 is good (Figure 2). Though this scale is typical, the descriptors poor, fair, and good are totally inadequate for evaluating subjective measures because the use of such terms as 'good' does not, in fact, ensure a standardized point of reference (Cummins & Gullone, 2000) and introduces a source of variation into the data collected. In order to minimize the effect of this variation on the results of analysis of the data, the usual technique used for selecting items for a Likert scale is to identify examples of things that lead to extreme expressions of the attitude being captured (Brooke, 1996), thus allowing for sufficient separation in responses to yield statistically significant results. Because the

DQEM scoring criteria use combinations of decision quality measures to differentiate among gradations of decision quality, the use of typical categorical labels designed to elicit extreme responses from only one measure would not allow for the use of combinations of decision quality measures. The DQEM needed a Likert scale that could be used to translate multiple subjective criteria into numeric values.

AoAs take a direct route from phase line to phase line					
Avenues of Approach	Poor		Fair		Good
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					
Measures					

Figure 2: Likert Scale with Typical Categorical Labels

In order to perform the translation of the SMEs' subjective evaluations to numeric scores, the DQEM discards the standard Likert scale, and instead employs Likert scales tailored to each sub-characteristic. These Tailored Likert scales serve to translate combinations of "good" evaluations in specific measures to numeric scores. An example of a tailored Likert scale is presented in Figure 3. These tailored Likert scales are similar

to Behaviorally Anchored Rating Scales (BARS) in they use raters (SMEs) to generate behaviors (measures) that are used to anchor each score. Unlike BARS, the measures are based on decision quality criteria not on specific behavior or actions. Also unlike BARS, the scoring criteria (see below) relate the achievement of combinations of specific measures to specific scores.

Like most conventional Likert scales, the target (sub-) characteristic is a positive statement as this reinforces the idea of evaluating the "goodness" of the sub-characteristic. The first difference in the tailored Likert scale is the inclusion of measures and supporting evaluation criteria. The evaluation of these measures will be associated with the numeric scores using the scoring criteria. A shorthand version of the evaluation criteria are provided to remind the SMEs of the evaluation criteria which are described in detail in a separate document. Like the sub-characteristic, the measures and evaluation criteria are phrased as positive statements to further reinforce the evaluation of the "goodness" of the sub-characteristic.

The second and most important difference between the typical Likert scale and the tailored Likert scale is the inclusion of the scoring criteria. Like BARS, the scoring criteria were generated from a SME consensus with the goal of reducing the variance in the scores assigned by SMEs due to the individual interpretation of the categorical labels. The scoring criteria relate achieving "good" evaluations on combinations of measures to the numeric scores on the Likert scale. The SMEs still need to exercise judgment in the evaluation of the measures, but the translation of their evaluations to numeric scores is less ambiguous than with traditional Likert scales.

AoAs take a direct route from phase line to phase line					
Avenues of Approach	**Meets neither 3 nor 4**	**Does not meet both 3 & 4**	**Meets 3 & 4 and 1 of remaining 3**	**Meets 3 & 4 and 2 of remaining 3**	**Meets all 5**
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

Measures

1. **Valid start point (behind PL, in a valid polygon, good route to start point)**
2. **Valid end Point (beyond PL, in a valid polygon, good route to objective)**
3. **No unnecessary turns (turns < 45^0)**
4. **Independence of routes (no common MCs)**
5. **Analyzed for on-road and off-road**

Figure 3: Tailored Likert Scale

The scoring criteria also serve a second purpose, to further capture the SMEs' understanding of impact of the problem's elements and relationships on the decision. Because the SMEs needed to reach consensus on what combination of "good" evaluations for measures is required to merit a given score, the SMEs' judgments of the relative importance of the measures were incorporated into the scoring criteria. For example, of the measures shown in Figure 3, measures 3 and 4 were considered most important so a subjective evaluation that meets both these criteria will always score better than one that meets only one or neither. Of the other three measures, none were considered as important as measures 3 and 4, and none were considered more important

to the evaluation of the sub-characteristic than any of the others. Together with the evaluation criteria, the scoring criteria and the tailored Likert scales provided a means of translating the SMEs' subjective evaluations to numeric scores that both helped capture the SMEs understanding and reduce the variance due to scale ambiguity

Aggregation criteria

The final process in the evaluation of decision quality is the aggregation of scores from the tailored Likert scales into a score that represent the overall decision quality. Since almost all MCDA methods rely on some sort of weighting method to aggregate option scores, the effect of aggregating sub-characteristic scores with both weighted and simple averages were investigated during the development of the DQEM. Since the DQEM decomposes decision quality into a tree hierarchy, the scores for each measure could be easily aggregated using simple or weighted averages. The elicitation of weights for use in the weighted averages was limited to the sub-characteristic level for two reasons. First, the relative importance of the measures was captured with the scoring criteria; and second, the SMEs found estimating the relative importance of such diverse measures difficult. The effect on the overall scores of aggregating the decision quality measure data using both simple and weighted averages are discussed with the results of the case studies in chapters four and five.

Statistical analyses

As mentioned previously, the case studies fulfilled two purposes; first, to assess the effects of adding GDSSs to the decision-making process, and second, to assess the

usefulness of the DQEM with respect to the evaluation of decision quality. Assessing the data collected with respect to these two purposes requires different statistical analyses. The assessment of the decision-making process primarily used a repeated-measures ANOVA to determine whether there was a significant change in the decision-makers' decisions. These analyses are discussed in detail in the following section on incorporating the DQEM into evaluations. The analyses used to assess the usefulness of the DQEM assessed whether the SMEs' evaluations of decision quality were affected by the use of the DQEM. These analyses were required to act on the data in a fundamentally different manner than did an ANOVA; the analysis of the DQEM used correlations among sets of data to determine the significance of the agreement of the SMEs' evaluation and Spearman's rank correlation coefficient to assess the SMEs' ability to differentiate among the quality of decisions. These analyses were used to address the following assessment hypotheses:

1. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the evaluations of decision quality will agree more closely (have higher inter-rater reliability).*
2. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the SMEs will be better able to differentiate among the quality of decisions.*

3. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the ability of decomposition and scoring criteria to capture the characteristics that the SMEs feel are the most important to the evaluation of decision quality will improve.*

All three hypotheses refer to the decomposition of decision quality becoming more detailed as a basis for the analysis of the effectiveness of decision quality. Since there is no way to compare the quality of decisions without actually evaluating the decision, and since there is no other means of evaluating the decision quality associated with complex, ill-structured problems other than DQEM, it was impossible to structure the assessment of the DQEM using traditional experimental design in which one condition used the DQEM and one didn't. Therefore in order to assess the value of the DQEM, an alternate basis for comparison was used. Since the DQEM depends on a detailed decomposition of decision quality, as that decomposition becomes more detailed the SMEs' evaluations should likewise become more reliable. To test these hypotheses, the SMEs' ability to reliably evaluate decision quality was compared at three points in the decomposition. Evaluating at three points during the decomposition allowed for trends in the effectiveness to be more evident than if only two evaluations of decision quality were used.

Pearson's Correlation Coefficient

The first assessment hypothesis directly supports the overall hypothesis given in chapter one:

The direct evaluation of the quality of decisions made to address complex, ill-structured problems can be improved through the use of a structured subjective decomposition of decision quality characteristics.

The first assessment hypothesis uses correlations to measure the agreement between, the reliability of, and the SMEs' overall evaluations of decision quality. The two correlations used are termed between-SME correlations because they are used to measure the agreement between the SMEs' evaluations. Two separate between-SME correlations measure the agreement between the SMEs' Subjective Overall Evaluation (SOE) scores and between the averages of the SMEs' scores for individual sub-characteristic, their Decision Quality Characteristic (DQC) scores. These correlations are defined as

$$Between - SME\ Correlation\ (SOE) = Corr\left(SOE_{1j}, SOE_{2j}\right)$$

and

$$Between - SME\ Correlation\ (DQCS) = Corr\left(DQCS_{1j}, DQCS_{2j}\right).$$

Where SOE_{ij} is the set of SOE scores from SME i, for each plan j, and

$$\overline{DQCS_{ij}} = \sum_k DQCS_{ijk};$$

The SMEs' subjective overall evaluations are the SMEs "gut" evaluations of the overall decision quality. The SMEs' SOE scores were elicited immediately after the SMEs had completed the scoring of individual sub-characteristics but without the SMEs having access to their scores. The scores were elicited at this point because there was no opportunity for the SMEs to evaluate the decisions without using the DQEM. Because the problems were complex and ill-structured, some analysis had to be completed before

the SOE scores could be elicited. Eliciting the SOE scores using the DQEM caused some concern that SOE scores would be contaminated by the SMEs scores for each sub-characteristic; but the problems were so complex, the SMEs evaluated either 35 or 53 measures and scored either 16 or 25 sub-characteristics for each of 10, 16, or 20 decisions, that SMEs felt that they would be unable to adequately recall the specifics of each decision and would not be certain of their SOEs. Feedback from the SMEs indicated that not having access to their scores for the sub-characteristics did allow them to give an overall estimate of the quality of each decision.

The averages of the SMEs' DQC scores were much more straightforward to generate. These scores were generated by using either weighted or simple averages of the scores recorded on the tailored Likert scales. The weights used to compute the weighted average of the DQC scores were elicited from the SMEs who developed the evaluation and scoring criteria and the tailored Likert scales and were based on the first level decomposition sub-characteristics. These weights were applied only when aggregating these sub-characteristics into the overall DQC score.

Both the SOE scores and the simple and weighed averages of the DQC scores were used when computing the between-SME correlations, and the results and conclusions drawn from the analysis of the results are presented with the case studies in chapters four and five. The SOE and averaged DQC scores were also used to compute within-SME correlations that were used to evaluate the third assessment hypothesis presented above. The within-SME correlation is defined as:

$$Within - SME\ Correlation_i = Corr\left(SOE_{ij}, DQCS_{ij}\right)$$

Where SOE_{ij} is the set of SOE scores from SME i, for each plan j, and

$$\overline{DQCS_{ij}} = \sum_k DQCS_{ijk};$$

Unlike the first assessment hypothesis, the third assessment hypothesis does not directly support the overall hypothesis given in Chapter 1 but instead is used as a measure of the ability of the decomposition of decision quality to capture the SMEs' understanding of the problem. The problems in the case studies require the SMEs to evaluate either 35 or 53 sub-characteristics and the DQEM assumes that these sub-characteristics capture the SMEs subjective understanding of the impact of the elements and relationships of a problem. Either 35 or 53 is a large number of criteria on which to evaluate a decision, and research has shown that when making decisions without some form of guidance that decision-makers use the most important 5-9 criteria (Zsambok, 1997) as the basis for their decisions. The within-SME correlations make use of this fact to assess whether the assumption that the decomposition represents the SME understanding is true.

If the within-SME correlations are not high, then the implication of the low level of agreement between a SME's SOE and averaged DQC scores is that there is some characteristic of decision quality that the SME either included or did not include in his SOE that was either not captured or erroneously captured in the decomposition. Conversely, a high within-SME correlation could be interpreted as an indication that the criteria most important to a SME had been included in the decomposition. The within-SME correlations are only an indication that a SME's most important criteria have been included in the decomposition; but since these criteria are likely to be in the first or second level of sub-characteristics, further decomposition of the sub-characteristics

would continue to support the evaluation of these sub-characteristics. Therefore high within-SME correlations would support the third assessment hypothesis.

In order to further characterize the data used to test the first and third assessment hypotheses, the statistical significance of the within- and between-SME correlations was determined by calculating p-values for each individual correlation as well as the significance of the changes in those correlations. The p-values for the individual correlations were determined by calculating the t:

4.1 $$t = \frac{r}{\sqrt{(1-r^2)/(N-2)}} \text{ for } N > 6$$

(Cohen & Cohen, 1983; Pitman, 1937). The calculated value of t can then be used to calculate the p-value from a student's t distribution. The p-values for the individual correlations test the null hypothesis that correlations are due to random chance. The p-values of the individual correlations were calculated using the web utility by (Lowery, 2012).

The significance levels of the changes in correlations between scorings were determined by calculating the associated z for each correlation using a Fisher's r-to-z transform,

4.2 $$z' = (1/2)[ln(1-r) - ln(1+r)]$$

Then, from Cohen & Cohen, 1983, the z-values are compared using

4.3 $$z = \frac{z_1' - z_2'}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

The p-values can then be determined from standard tables. When comparing two correlations, the p-values test the null hypothesis that the two correlations are the same, i.e., $r_1 = r_2$ The p-values for the changes in correlations were calculated using the web utility (Preacher, 2002).

Bonferonni correction

Because the analyses of the within- and between-SME correlations were done in three successive stages in the first case study, the effect of multiple comparisons must be addressed when analyzing the significance of the correlations. When doing multiple comparisons among statistical tests, including correlations, the increased possibility of Type I errors must be considered; the more samples that are compared the more likely it is that at least one of the correlations will be high due to random variation in the data. Of the numerous methods for correcting significance levels, the Bonferroni correction is the most conservative (Abdi, 2007; Dunn, 1961). Using a Bonferroni calculator (Uitenbroek, 1977) to determine the individual significance level which corresponds to the combined level of significance of $p = 0.05$ for three samples, the significance level should adjusted be $p = 0.01695$ for each individual comparison (given an unknown correlation). In the first cases study, p-values were considered statistically significant at the Bonferroni corrected level of 0.01695. Since there were only two sets of evaluation scores in the second case study, the Bonferonni correction was not needed and p-values were considered significant at the $p = 0.05$ level.

Spearman's rank correlation coefficient

Unlike the first and third assessment hypotheses, correlations are not used to evaluate the second assessment hypothesis; instead Spearman's Rank Correlation Coefficient is used to compare the ranks of the SMEs' evaluations of decision quality. As discussed above, Pearson's correlation coefficient can be used as an indication of the reliability of the SMEs' evaluations as a whole*;* but even though Pearson's correlation coefficient uses paired data, the between-SME correlations are an indication of the overall reliability among the set of SMEs' evaluations and not an indication of the reliability of any pair of evaluations or of any given evaluation. Therefore, Pearson's correlation cannot be used as an indication of the SMEs' ability to discriminate among specific decision qualities. But, if the between-SME correlations indicate that the SMEs' evaluations are becoming more reliable; it follows that the SMEs' evaluations should be better able to discriminate among the quality of decisions. If the SMEs evaluations are better able to discriminate among decision qualities, then a ranking of their overall decision quality scores should be more similar. SRCC captures the level of similarity in the rankings of the SMEs and therefore can be used as a measure of agreement of their evaluations of each decision and a measure of the SMEs' ability to discriminate among the quality of specific decisions.

In order to compare the rankings of the SMEs' overall evaluation scores, Spearman's rank correlation coefficient was used. The Spearman rank correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables

(Myers, Well, & Lorch Jr., 2010). For a sample of size n, the n raw scores X_i, Y_i are converted to ranks x_i, y_i, and ρ is computed from these:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{4.4}$$

Identical values (rank ties or value duplicates) are assigned a rank equal to the average of their positions in the ascending order of the values. Although it is a modification of Pearson's correlation coefficient, Spearman's rank correlation coefficient is specifically used as an indication of the agreement among the ranking of two sets of data; and as such it can be used as an indication of the SMEs' ability to differentiate among the quality of decisions and used as a test parameter for assessment hypothesis two.

Overall, three assessment hypotheses were developed to support the research hypothesis. Assessment hypotheses one and three directly support evaluating the SMEs' ability to reliably evaluate decision quality using the DQEM. While assessment hypothesis two only indirectly supports the evaluation of decision quality, it is a measure of the DQEM's ability to capture the characteristics of the decision that the SMEs consider most important.

Incorporating DQEM into an experimental structure

The DQEM was incorporated into three summative experiments and several formative evaluations during the course of its development. The first two of these evaluations made the most extensive use of the DQEM and contributed the most to its development, and these two summative experiments are reported as the case studies. The

goal of the experiments in the case studies from the sponsor's point of view was to evaluate the effect of using GDSSs on decision-making in one phase of the MDMP. From a decision theory point of view, the case studies assessed the effect of changing MDMP by augmenting it with the GDSSs. Since none of the decisions evaluated in either case study would ever be implemented and simulation would be too difficult, the outcomes of the decisions could not be used as proxies for decision quality. Similarly, since the case studies involved military planning and the GDSSs would be integrated into the codified MDMP, evaluating the fidelity of both the standard MDMP and the MDMP augmented by the GDSS as per the process-based approach was not feasible. Therefore a method of directly evaluating the quality of the decisions was needed. This need was the proximal motivation for developing the DQEM, and the incorporation of the DQEM into this type of evaluation posed unique challenges.

Because the experiments in the case studies serve a dual purpose, the terminology used in the case studies needs to be clarified. In the case studies, decision-makers are those individuals who participate in the decision-making process and decisions as the output of that process. This includes both the staff members who contribute to generating decision options as well as decision-maker who ultimately decides on a course of action. SMEs are experts in the subject area of the decisions who are not part of the decision-making process and who develop the DQEM and/or use the DQEM to evaluate decisions. The decision-makers are most likely also subject matter experts with some level of expertise; but for the purposes of this research, SMEs are outside of the decision-making process and use the DQEM to preform independent evaluations.

Define the problem

There are several aspects that were critical to defining the problem with respect to incorporating the DQEM into an evaluation. The most important aspects are discussed below and all the aspects are discussed in Appendix 4-1.1

Determine critical aspects of GDSS

The critical aspects of the GDSS were essential to the definition of the problem, and development of the evaluation may be determined. In the case studies, these critical aspects were the functions and output of the GDSS and the types of decisions the GDSS was designed to support. These aspects are not independent and identifying the specific aspects required consultation with the GDSS developers and Subject Matter Experts (SMEs) as well as referencing design requirement and doctrinal documents. The functions and output of the GDSS led directly to decisions the GDSS was designed to support. For example, an overlay of natural and man-made obstacles is necessary as a prerequisite step for generating routes. Therefore a function in the GDSS that identifies obstacles would support route planning decisions. Given the functions of the GDSS and output of the GDSS, SMEs who are familiar with the types of problems that the GDSS is designed to support can ascertain for what decisions what types of problems, and what categories of decision-makers the GDSS would be most useful.

Determine general tasks and mission

The development of the general tasks and the mission were the next steps in defining the problem and resulted in the development of the overall decision quality

characteristics. The general tasks and mission were developed from the critical aspects of the GDSSs, and their development resulted in the definition of the problem. The general tasks are the tasks which would require decision-makers to make the types of decisions identified as critical aspects of a GDSS. The general tasks need not be in one-to-one correspondence with the supported decisions. Several supported decisions can and should if possible be incorporated as part of one general task, but all supported decisions must be incorporated into at least one task. The mission is essentially the problem without a defined context. It encompasses all the general tasks, provides the general context, specifies a decision that needs to be made, and defines the overall decision quality characteristics. For example, all the basic routing functions of the GDSS in Case Study One support one decision, the determination of possible routes in support of Course Of Action (COA) development. This supported decision leads to one general task to generate valid routes for course of action development. A mission that encompasses the general task could be "conduct an analysis of the brigade area of operations and generate potential routes in support for battalion-sized units from the assembly area to the objective." This mission requires the decision-makers to decide on potential routes and provides a general context in form of unit sizes, a start point (the assembly area), and an endpoint (the objective).

The mission also determines the overall decision quality characteristics that form the starting point for the decomposition into sub-characteristics. In the case studies, the mission led to three characteristics that were used to evaluate the usefulness of the GDSS; but only one of which was associated with decision quality. The overall decision quality

characteristic used in the case studies was that using the DGSS would allow the decision-makers to produce higher quality plans; and specifically in case study one, the decision-makers would produce higher quality potential routes. The two characteristics of the usefulness of the GDSS that were not associated with decision quality were that the routes could be produced more quickly and that the decision-makers' understanding of the impact of the terrain on the decisions would not be reduced. The second of these, evaluating the loss of understanding, used a modified version of the DQEM to evaluate the change in the decision-makers' understanding. Specifying the mission led to the development of the top-level decision quality characteristic and left the development of the context as the last component needed to complete the problem definition

Determine context and scenario

The context for the case studies was provided by the scenario. The scenario provides the detailed information necessary for the decision-makers to make their decisions. The overall mission and general tasks only provide a framework for the detail needed to provide a realistic context in which an evaluation could be conducted. The scenario contained the details that fleshed out that framework and provided enough background details so that decision-makers had sufficient information to make and evaluate options. Because the decision-makers should be representative of the target user, these decision-makers will have performed tasks similar to those included in the evaluation in real world situations. They will be aware of the importance of the decisions they make to the overall mission, and the scenario must provide enough detail so that the decision-makers can determine the impact of the tasks on the overall mission. Insufficient detail may

distract the decision-makers and introduce variation in the decision-makers' decisions since, without sufficient information, the decision-makers may have to dedicate time and energy to making assumptions about or inferring information from the information that is provided. Figure 4, is an example of the graphical context provided to decision-makers in the first case study. Overall, the scenario provides the detailed contextual information, while the identification of sub-tasks and associated decision quality measures describe below form the core of the evaluation.

Define decision-makers

The development of the mission, general tasks, and scenario yielded insight into the characteristics of the personnel who are likely to make the type of decisions which the GDSS was designed to assess. The most important characteristic of the decision-makers was that they were experienced with the types of decisions that the GDSS was designed to support. The characteristics that could be factors in selecting decision-makers could be specific qualifications and experience or any other characteristics that would aid the decision-makers in making the decision which the GDSS is designed to support. Using decision-makers who were not experienced with the type of decisions the GDSS was designed to support would probably be a source of variance in the data collected that could otherwise be avoided.

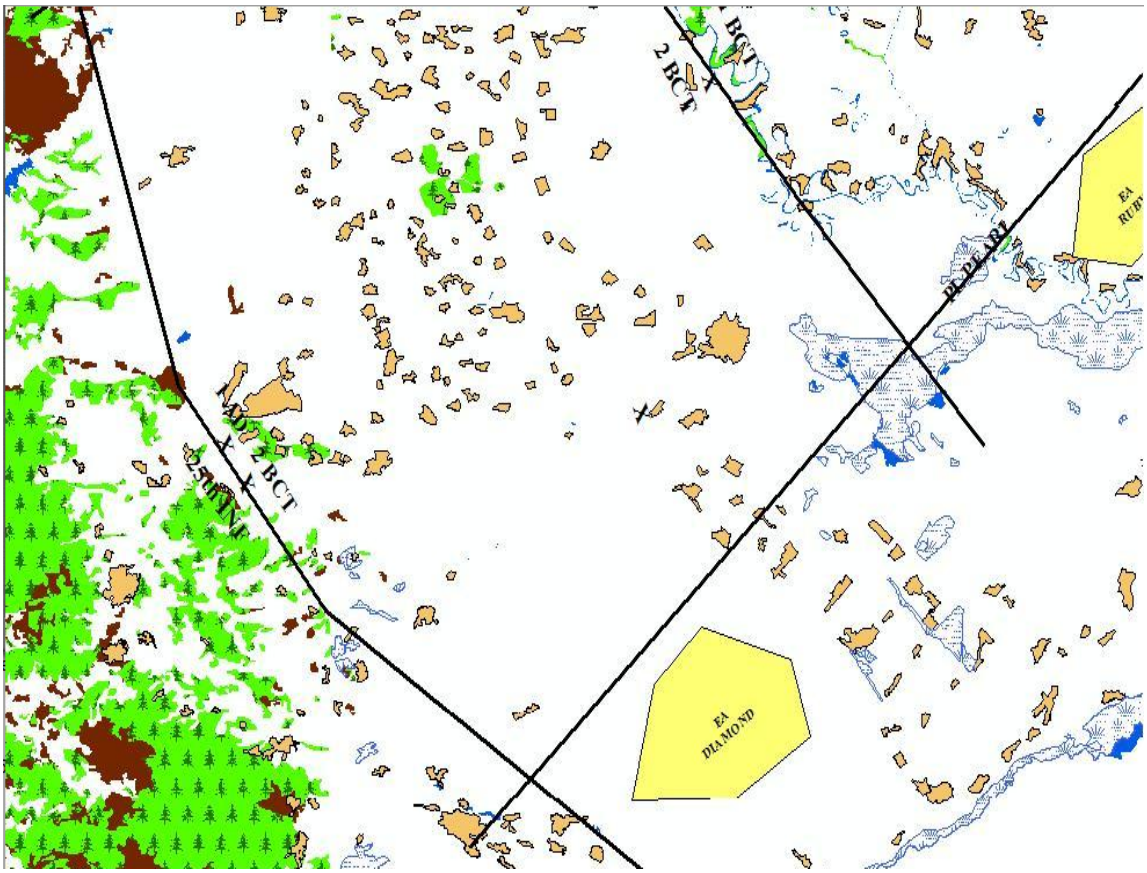

Figure 4: Case Study One operational graphics

Decompose decision quality and define planning tasks

The decomposition of decision quality was closely linked to the development of experimental planning. Since the general task of generating routes is supported by multiple GDSS functions, the corresponding decision quality characteristic, quality of the routes, will be supported by sub-characteristics that are related to the sub-tasks that exercise the GDSS functions and that support the general task. The decomposition of decision quality characteristics and the generation of planning tasks form an iterative process. The identification of sub-tasks that supported the general task required decision

quality sub-characteristics that characterized the decisions required by sub-tasks; and conversely, the definition of sub-characteristics led to the identification of decisions that supported the assessment of the impact of the GDSS and for which sub-tasks were then developed. The decomposition of decision quality extended further than the identification of explicit sub-tasks. In particular, the development of measures and evaluation criteria generally characterized decisions that the SMEs considered to be implicit tasks already identified. The process of decomposition and task generation was iterated until all sub-tasks had been characterized by sub-characteristics and all sub-characteristics had supporting measures.

Case Study Design

Although the DQEM is designed to assess decision quality directly, the implementation of the DQEM in the case studies was based on assessing differences in decision quality that were used to evaluate the impact of using GDSSs on the MDMP. Since a decision quality score is a unitless value that has no meaning unless used as basis for comparison, the case study evaluations used a within-subject design with respect to decision quality. A within-subjects design is one in which the same decision-makers give two sets of responses. In this case, one set of responses will be generated when making decisions with the GDSS (With Case) and one will be generated when making decisions without using the GDSS (Without or Base Case). Throughout this research, the Base Case trials will be referred to as the Without Case, i.e., planning without the GDSS; and the trials in which the decision-makers used the GDSS will be referred to as the With Case.

The evaluation structure described for the case studies is that of a rigorous experiment. The experimental structure was chosen primarily in order to assess the change in decision quality that resulted from augmenting the MDMP with the GDSS and the value of the GDSS to military decision-making. A secondary consideration was providing information on the usefulness of the GDSS and providing feedback to the developers on specific functional and design areas in which improvements were needed. The elements of the general experimental design are available in the literature so minimal attention will be paid to the common techniques, but some attention will be paid to techniques that impact the decision-makers and the data gathering.

It is important to note that the evaluation structure was designed to assess the value of the GDSS to military decision-making. The assessment of the DQEM usefulness was of secondary consideration to the sponsor. The only adjustments that were made to the evaluation structure in order to assess the DQEM were in the data collection. The data collection was modified to generate multiple data collection opportunities to support the assessment of the DQEM. The assessment of the DQEM used all the data collected, but only the last set of data was used for the evaluation of the GDSS. The specifics and details of the construction of the experiment for the first case study are provided in Appendix 4-1, and a summary of the experimental design is provided below.

The GDSS

The GDSS that is evaluated in the case studies consisted of the Battlespace Terrain Reasoning and Awareness – Battle Command (BTRA-BC) suite of geospatial

tools. BTRA-BC contains various tools that can be applied to military planning problems of varying levels of complexity. The individual BTRA-BC tools are referred to as tactical Spatial Objects (TSOs). TSOs are computationally lightweight software engines that transform geospatial data into geospatial information unique to a specific military planning analysis. BTRA-BC TSOs are part of a GDSS whose capabilities include analysis engines, data manipulation routines, and other software products in support of terrain reasoning (USACE, 2003). BTRA-BC generates information addressing (1) Observation, Cover and Concealment, Obstacles and Mobility, Key terrain and Avenues of approach (OCOKA); (2) integrated products defining operational Positions of Advantage; (3) high-fidelity weather/terrain effects on mobility and signature physics; (4) advanced mobility analysis; (5) digital ground and air maneuver potential; and (6) tactical structures relating information produced by the other components (USACE, 2010). BTRA-BC's focus is the development of software analytics designed to create information and knowledge products that capture integrated terrain and weather effects and develop predictive decision tools to exploit those products. The ultimate objective is to empower commanders, soldiers and systems with information that allows them to understand and incorporate the impacts of terrain and weather on their functional responsibilities and processes (USACE, 2009).

The case studies evaluated various BTRA-BC GDSS functions that are grouped by complexity into Tiers. The Tier 2 TSOs that were evaluated in the second case study were more sophisticated and supported more complex problems than the Tier 1 TSOs evaluated in the first case study. Because the Tier 2 TSOs were more sophisticated and

supported more complex decisions than those in the first case study, the problem, the context, the tasks, and the decisions in case study Two were more complex than those in Case Study One. This generally greater complexity also required decision-makers with a different set of decision-making skills. Table 3, below, summarizes the design factors incorporated into case studies one and two.

Table 3: Comparison of Case Study Design Considerations

Case Study	**One**	**Two**
GDSS Functions	BTRA-BC Tier 1 TSOs	BTRA-BC Tiers 1 & **2** TSOs
Level of Decision-maker	Terrain Analysts	Staff Planners
Number of decision-makers (actual/design)	18 / 16	8 / 16
Host Environment	DTSS	Commanders' Support Environment (CSE)
Unit Size	Brigade	Battalion
Mission	Tactical Movement	Tactical Movement to seize an Objective in the presence of hostile forces
Tasks	Terrain Analysis	
	Recommend Avenues of Approach (routes)	Develop Course of Action (COA)
Output	Digital Plan	Digital Plan & Written OPORD
	Terrain Understanding Questionnaire	
	Comparison Questionnaire	
	Post Trial Discussions	

Independent variables and balancing

The case study design employs a balanced, repeated measures design with three independent variables:

- System (with and without the GDSS functionality)
- System Order (whether the first scenario is worked with or without GDSS functionality)
- Scenario Order (whether scenario 1 or 2 is worked first)

and three dependent variables that support the three evaluation hypotheses:

- Decision Quality (quality of decision-makers' plans)
- Time to complete the plans
- Decision-makers understanding of the impact of terrain on the plans

The design of the experiment is summarized in Table 4. One of the independent variables, System, was a within-subjects variable; each decision-maker would work one planning scenario with the GDSS and one scenario without the GDSS. System Order and Scenario Order were between-subjects variables because any given decision-maker can only be part of one ordered sequence for these variables. The decision-makers performed the same tasks on two similar military planning scenarios. One set of tasks was performed with the GDSS functions in addition to those native to the host system (With Case), and the other set of tasks was performed with the host system functions only (Without Case). The two trials were essentially identical except for the use of the GDSS. A within-subjects design is particularly valuable when the number of available decision-makers is limited as in the current case. Results from the sets of tasks can be compared

for each decision-maker, thus reducing decision-maker-specific effects that might add variability to the results.

In order to balance the decision-maker populations, the decision-makers were split into two groups that were evenly balanced as to the ability and knowledge of the decision-makers as determined by biographical information supplied by the decision-makers. The first of these groups performed the set of tasks first without the GDSS and then with the GDSS. The second group reversed the order of tasks. The order of the tasks was randomly selected so that half of the decision-makers performed each of the tasks first. Randomizing the order of the tasks enabled the analysis to control for learning effects. As discussed above, it was expected that this design could generate statistically significant results with 16 subjects in the System variable, our variable of primary interest.

Table 4: Design Elements

Variable	Manipulation	Levels
System	Within Subject	With System Without System
System Order	Between Subjects	With System then Without System Without System then With System
Scenario Order	Between Subjects	Scenario 1 then Scenario 2 Scenario 2 then Scenario 1

In order to reduce contamination of both the With System and Without System cases, the decision-makers were not exposed to the System until immediately before the trial in which the system is used. In light of this limitation, the experimental design

elements in Table 4 required the decision-makers to be divided into four groups. The two groups discussed above were further divided into two sub groups while maintaining the balance of ability and knowledge. The first of these subgroups in each group would perform the tasks in terrain area one (Scenario 1) then in terrain area two (Scenario 2) while the second subgroup reversed the order of terrain areas. This structure, summarized in Table 5, controls for any variance in results due to the experience of the decision-makers, the order of the systems use, or the order of terrain area used. The specific systems and scenarios for each evaluation are discussed in chapters 4 and 5.

Table 5: Experimental Groups

	Scenario Order	
System Order	With System then Without System Scenario 1 then Scenario 2	With System then Without System Scenario 2 then Scenario 1
	Without System then With System Scenario 1 then Scenario 2	Without System then With System Scenario 2 then Scenario 1

Data Collection

As mentioned previously, the data collection requirements of the evaluation of the GDSS were somewhat different from those for the assessment of the DQEM. In both case studies, the decision-makers each generated two responses documenting their decisions; one response for each of the With and Without Cases. These two sets of

responses provided the raw data for both the evaluation of the GDSS augmented MDMP and assessment of the effectiveness of the DQEM. The evaluations of decision quality were provided by independent SMEs who were not involved in the development of decomposition of decision quality characteristics.

AoAs take a direct route from phase line to phase line					
Avenues of Approach	Poor		Fair		Good
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					
Measures					

Figure 5: Example of an Untailored Likert Scale from the First Soring

The evaluation of the decision quality was conducted in three stages that provided the data that was used to determine the effect of increasing the level of detail in the decomposition of decision quality on the effectiveness of the DQEM. Since the decomposition of decision quality characteristics into sub-characteristics was conducted concurrently with the task development, the SMEs evaluating the decision-makers' decisions had the decomposition available to them for their initial evaluations. At this

point the SMEs had not reached consensus on the measures to support the evaluation and the first scoring was conducted using the Likert scales similar to Figure 5

This first scoring generated the baseline against which the succeeding scorings would be compared. Based on the literature, it was not expected that this scoring would yield evaluations that would be reliable enough to be used in the assessment of the DQEM. The of the SMEs' evaluations were not sufficiently reliable to be used to assess the DQEM, and the SMEs extended the decomposition and reached consensus on the measures that supported the evaluation of the decision quality of each sub-characteristic.

AoAs take a direct route from phase line to phase line					
Avenues of Approach	Poor		Fair		Good
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

Measures

1. **Valid start point**
2. **Valid end Point**
3. **No unnecessary turns**
4. **Independence of routes**
5. **Analyzed for on-road and off-road**

Figure 6: Semi-tailored Likert scale from the 2nd scoring

The second scoring was conducted with semi-tailored Likert scales that identified the measures that supported each sub-characteristic but did not identify the evaluation or

scoring criteria (which had not yet been developed). Figure 6 is an example of a Likert scale used in the second scoring. Like the first scoring, the second scoring did not produce evaluations with statistically significant reliability, and the SMEs then developed the evaluation and scoring criteria.

The third scoring was conducted with fully tailored Likert scales similar to Figure 7 that have the evaluation criteria listed with the associated measures and have the standard Likert scale categories replaced with the scoring criteria.

AoAs take a direct route from phase line to phase line					
Avenues of Approach	Does not meet both 3 & 4	Meets both both 3 & 4	Meets 3 & 4 and 1 of remaining 3	Meets 3 & 4 and 2 of remaining 3	Meets all 5
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

Measures

1. **Valid start point (behind PL, in a valid polygon, good route to start point)**
2. **Valid end Point (beyond PL, in a valid polygon, good route to objective)**
3. **No unnecessary turns (turns < 45^0)**
4. **Independence of routes (no common MCs)**
5. **Analyzed for on-road and off-road**

Figure 7: Fully Tailored Likert Scale from the Third Scoring

The data from this third and final scoring was used for both the assessment of the effectiveness of the DQEM and the evaluation of the GDSS augmented MDMP. For the assessment of the DQEM, the evaluation data was compared with the data from the

previous two scorings using the analyses described earlier in this chapter. Unlike the assessment of the DQEM, the evaluation of the GDSS did not require a series of sets of data; and since the evaluation data from this scoring would be the most reliable, only data from this scoring was used in the evaluation of the GDSS. The results of the analyses of all the data is presented in Chapters Four and Five for Case Studies One and Two respectively.

Summary

This chapter's primary focus was to present an overview of and background of the development of the DQEM and describe how the DQEM was integrated into an evaluation structure. The DQEM described the process by which decision quality can be decomposed into characteristics and sub-characteristics. These sub-characteristics define individual aspects of decision quality particular to a given problem, and taken together they can be used to evaluate the overall decision quality. The method also shows how measures can be defined to permit the quality of these sub-characteristics to be evaluated and a decision quality score to be generated from the aggregated measure scores.

This chapter also discussed the unique challenges of integrating the DQEM into a design to evaluate the change in a decision-making process. This evaluation required the development of a mission, a scenario, and general tasks that encompassed the decisions the GDSS was designed to support. Further, the decomposition of decision quality characteristics was developed concurrently with the planning tasks designed to support both the evaluation of the GDSS and the assessment of the DQEM. This chapter also briefly touched on the experimental structure of the case studies before discussing the

collection of data that would support both the evaluation of the effects of augmenting the MDMP with the GDSS and the assessment of the effectiveness of the DQEM. Specific details of the implementation of the DQEM and its integration into Case Study One is provided in Appendix 4-1. The discussion of this implementation is intended to serve as a guide for future evaluations of decision quality and decision-making.

The following chapters describe the two case studies in which the DQEM was implemented. The first case study evaluated decision quality in a less complex planning context, i.e., less complex scenario and GDSS functions, than the subsequent case study. The second case study applied the DQEM to a more complex scenario requiring more complex decision-making. Both case studies were designed to evaluate the effect that using geospatial planning tools have on a decision-making process. These case studies also generated data that was used to assess the impact of the DQEM on the ability of SMEs to reliably evaluate the quality of decisions based on complex, ill-structured problems. Both case studies include detailed discussions of results with respect to both the value of the GDSS to military decision-making and with usefulness of the DQEM in evaluating decision quality.

CHAPTER FOUR: RESULTS AND CONCLUSIONS FROM CASE STUDY ONE

Chapter Three discussed in general terms the implementation of the DQEM and the incorporation of the DQEM into an evaluation structure. This chapter summarizes the specific implementation of the DQEM in Case Study One and the results obtained in that case study as they relate to the assessment of the effectiveness of the DQEM. Because the intent of this chapter is to concentrate on the results generated in the case study and the conclusions that can be drawn from those results, the implementation of the DQEM and the development of the experimental structure are summarized in sufficient detail to provide the background information needed to understand the conclusions drawn in the case study. Since the details of the implementation of the DQEM and the development of the experimental structure may be of interest to other researchers desiring to directly assess decision quality, these details are presented in Appendix 4-1.

Case Study One implemented the DQEM and the experimental structure described in Chapter Three in an evaluation of GDSS functions developed for a project at the U.S. Army Engineer Research and Development Center (ERDC). The goal of the evaluation, from the point of view of ERDC, was to assess the impact on of the use of GDSS functions on decision making. This goal resulted in actually using the DQEM to assess the impact of augmenting the Military Decision Making Process (MDMP) with the

GDSS functions. The goal of this case study, on the other hand, was to investigate the effectiveness of the DQEM in the evaluation of decision quality. The specific GDSS functions evaluated in the assessment were the Battlespace Terrain Reasoning and Awareness – Battle Command (BTRA-BC) suite of geospatial tools (U.S. Army, 2003). The BTRA-BC program, which builds upon a commercial GIS tool (ARCINFO), has resulted in mature components that have been integrated into the Army's Digital Topographic Support System (DTSS), a system that provides topographic engineering support to topographic technicians as they assist military planners (Herrmann, 2002). DTSS provides geospatial data generation, collection, management, information processing, and services. The BTRA-BC GDSPs expand the capabilities of DTSS through the creation of information and knowledge products that enhance soldiers' understanding of terrain and weather as it impacts their functional responsibilities. The BTRA-BC functions assessed in this study include the identification of obstacles, the production of a Modified Combined Obstacles Overlay (MCOO), the generation of Mobility Corridors (MCs), the combining of MCs to form routes, and the identification of Choke Points (CPs). While this assessment provided essential information to evaluate the contribution of the BTRA-BC tools in particular and GDSSs in general, to the military decision making process, it also provided data on the effectiveness of using the DQEM to evaluate decision quality.

Scope of Case Study One (CS-1)

The primary goal of this case study was to investigate the effectiveness of using the DQEM to evaluate decision quality in complex and ill-structured problems. A

secondary goal was to describe the implementation of the DQEM in such a way that this chapter could be used as a guide for others to use in designing evaluations of decision quality (Appendix 4-1.1). A third goal was to generate the lessons learned from this implementation of the DQEM that could be used to refine the DQEM for use in more complex problems. Case Study Two (Chapter Five) incorporates these lessons learned.

As discussed in Chapter Three, the evaluation of the GDSS functions required the definition of the problem and the decisions to be evaluated. The process of defining the problem included using the knowledge and judgment of SMEs to determine the decision the GDSS functions were designed to support, to develop general tasks that would lead decision-makers to make those decisions, to develop a mission that encompassed all the decisions and tasks, and to generate a scenario that proved the context in which the decisions would be made. Once the problem had been defined, theses SMEs iteratively constructed a hierarchy of specific tasks that would make use of the GDSS and require the appropriate decisions in conjunction with the decomposition of decision quality into hierarchy of sub-characteristics. The mission, scenario, and tasks were then incorporated into the experimental structure described in Chapter Three. Once data were obtained, a second set of SMEs developed the evaluation and scoring criteria and constructed Likert scales tailored to each sub-characteristic. They used these Likert scales to evaluate and score each decision-makers' decisions, and these evaluation scores were analyzed using the analyses described in Chapter Three to produce the results presented in this case study.

Case study development

The discussion in this section pertains primarily to the evaluation of the usefulness of the GDSS functions in decision-making and are provided as the context in which the assessment of the DQEM was conducted. The basic experimental structure was discussed in Chapter Three and case-study specifics are summarized here. Appendix 4-1 provides a detailed description of Case Study One and the incorporation of the DQEM.

Problem Definition

The starting point of the problem definition, identifying the GDSS functions and the decisions they are designed to support are summarized in Table 6. The decisions supported by the GDSS functions are listed in the appropriately named column. From these decisions the following mission, which would typically require these decisions to be made, was developed:

> Mission: Conduct an analysis of the terrain in brigade area of operations and generate potential routes in support of the movement of battalion-sized units from the assembly area to the objective.

The scenario that provides the context supporting the mission was documented in the 2-page operations order in Appendix 4-2. This context was supported by operational graphics and the underlying digital terrain data presented by the GDSS. A sample of the operational graphics that supplemented the operations order are shown in Figure 4.

Table 6: CS-1 − GDSS Functions

TSO	TSO Functions	Output of TSO	Decisions Supported	Geospatial Environment
Obstacles	Identify areas of highly restricted and restricted terrain due to terrain factors	Generate Combined Obstacle Overlay (COO) with areas of highly restricted and restricted terrain identified	Determination of areas through which vehicular travel is not possible	Open terrain with natural and man-made obstacles to vehicular movement. The terrain must have MC that vary along the potential routes, terrain that exhibits choke points for variously sized units, and both potential on-road and off-road routes. There should be enough variation in the terrain that multiple routes are possible.
	Identify areas of highly restricted and restricted terrain due to man-made obstacles			
Mobility Corridors (MCs)	Identify Mobility Corridors (MCs)	Generate overlay with MCs categorized by size	Determination of possible routes in support of Course Of Action (COA) development	
Movement Projection	Identify potential routes by primary vehicle type	Generate overlay with potential routes for designated vehicle types		
	Optimize routes for time	Generate overlay indicating fastest routes for designated vehicle types		
	Optimize routes for distance	Generate overlay indicating shortest routes for designated vehicle types		
	Optimize routes given user input barriers	Generate overlay with routes that do not cross user input barriers		
	Optimizes routes for both on-road and off-road	Generate overly routes as specified by user selection of on- or off-road		
	Calculate travel times for vehicle type	Generate travel times for generated routes by vehicle type		
Choke Points (CPs)	Identify areas of restricted movement due to unit size	Generate overly indicating areas along routes where movement will be restricted categorized by unit size		

The specific planning tasks and the detailed decomposition of decision quality characteristics generated by the SMEs can be found in appendices 4-2 and 4-1 respectively. Table 7 provides a summary of the decomposition of decision quality characteristics. Table 8 summarizes the design factors that were incorporated into the evaluation of the GDSS in CS-1 including the appropriate decision-makers, general tasks, and the outputs that reflect the decisions made by the decision-makers. An example of the digital plans with four potential routes that were developed by the decision-makers is shown in Figure 8.

Table 7: CS-1 – Decision Quality Decomposition Summary

Number of Characteristics / Measures per Level

Decision quality characteristics	5
1st level sub-characteristics	13
2nd level sub-characteristics	16
Subjective measures	35

Table 8: CS-1 − Design Considerations

Case Study One	
GDSP	BTRA-BC Tier 1 TSOs
Decision-makers	Terrain Analysts
Number	18
Host System	DTSS
Mission	Tactical Movement
Tasks	Terrain Analysis
	Recommend Avenues of Approach
Output	Digital Plan
	Terrain Understanding Questionnaire
	Comparison Questionnaire
	Post Trial Discussions

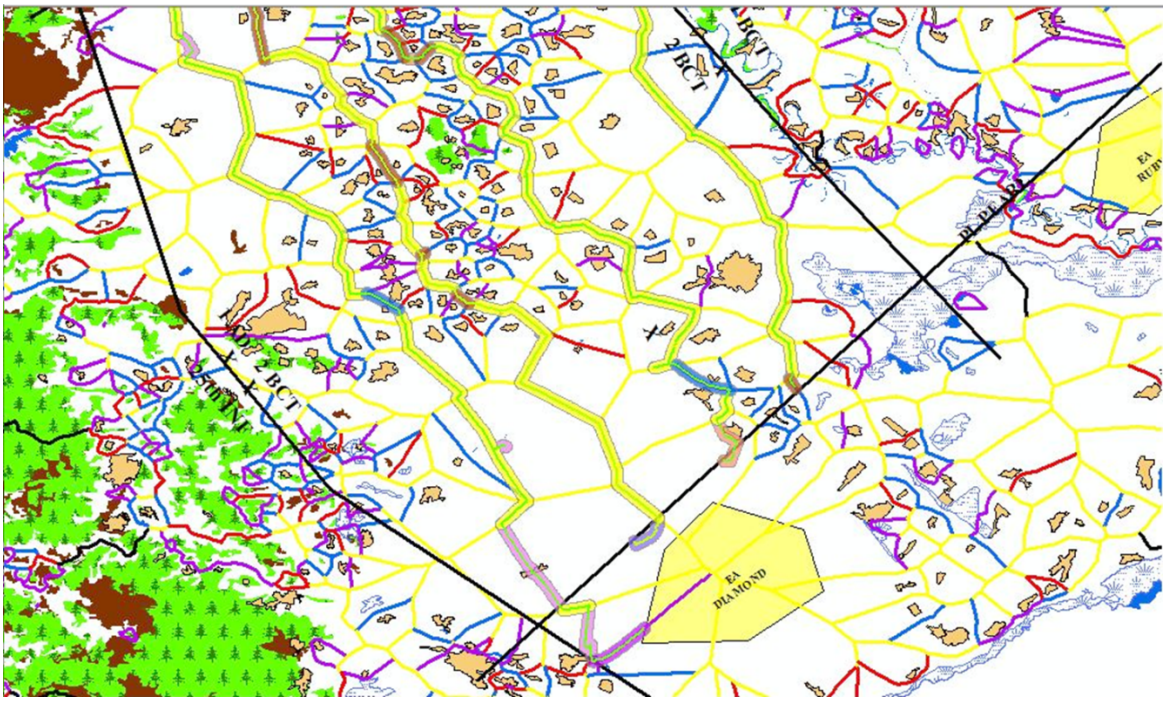

Figure 8: CS-1 − Sample Digital Plan

Because this case study was designed to evaluate of the effect of augmenting the MDMP with the GDSS, the hypotheses and the outputs used to gather data to test them were not limited to just the evaluation decision quality. The SMEs and researchers who developed this case study identified the six hypotheses listed below to evaluate the change in the decision-making process: Of hypotheses 1, 2 and 4, which were evaluated using data gathered from the decision-makers' decisions, only the primary hypothesis, hypotheses 1, is related to the quality of the decision. Hypotheses 3 was defined as a check to see if the decision-makers using the GDSS fell in to the trap of believing the information presented by the system. Hypotheses 5 was used to verify the integrity of the experimental design, and hypothesis 6 was added to capture the decision-makers' subjective evaluation of the use of the GDSS. A detailed discussion of these hypotheses and the results associated with them can be found in Appendix 4-1. The hypotheses for evaluation of impact of the GDSS functions in this case study stated that trained, experienced, military personnel who use the GDSS would:

- *Produce a higher quality plan than personnel not using the GDSS;*
- *Produce the designated plans more quickly than personnel not using the GDSS;*
- *Display as good an understanding of the impact of the given terrain on military decision-making as personnel not using the GDSS;*
- *Produce decisions using the GDSS that are more uniform, i.e., have less variance in the first two of the three categories above (speed and quality), than output generated without the use of the GDSS;*

- *Would not exhibit a learning effect due to experimental design; and*
- *Consider using a GDSP superior with respect to (1) allowing them to complete the tasks more quickly, (2) allowing them to produce higher quality output, (3) allowing then to have a greater terrain understanding, and (4) overall .*

The next section summarizes the statistical analyses used to test these hypotheses.

Analysis

The primary analysis used to determine whether the uses of the GDSS (the System variable) had an effect on the decision-makers' decision-making was a repeated-measures analysis of variance (repeated measures ANOVA) of the decision-makers' responses. Under the assumption that the data are Gaussian (normal) or near-Gaussian (near-normal), a repeated-measures ANOVA should determine whether the decision-makers' average response when using a GDSP are significantly different from their average response when not using the system. As the System variable is within-subject, the repeated-measures ANOVA should be able to determine statistical significance for a smaller main effect than for an effect due to the between-subjects variables, System Order and Scenario Order.

An ANOVA should also provide evidence of any existing interaction among System, System Order, and Scenario Order variables.

Because the small projected sample size (16) may not yield enough samples to allow the data to be treated as having a normal distribution, the normality of the data sets

and the validity of the results of the ANOVA were verified using normal-probability plots and the non-parametric Wilcoxon Signed-Rank Test. Also, other tests were conducted to confirm the validity of and to supplement the ANOVAs. For instance, some objective data, such as binary data, cannot not be treated as normal; and an ANOVA cannot be used to determine if two binary distributions are different. Instead, a chi-square test can be used to estimate the probability that two sets of binary responses came from different distributions. Equal variance tests were used to determine whether the average variation in the data is smaller when using the GDSP under evaluation than when it is not used. The analyses are discussed in detail in Appendix 4-1.

The analyses just discussed only to support evaluation of the effect of using the GDSS and not for the assessment of the DQEM. The assessment of the effectiveness of the DQEM used the analyses discussed in Chapter Three. The analyses included the between- and within-SME correlations of the SMEs' subjective overall evaluation (SOE) scores and averaged decision quality characteristic scores (DQCS) as well as Spearman's Rank correlation coefficient.

Results

The analysis of the collected data was in accordance with the analyses discussed above and in Chapter Three. Detailed results and conclusions are presented below for the assessment of the effectiveness of using the DQEM to evaluate decision quality. A summary of the results of the evaluation use of the GDSS are also provided here, and detailed results and conclusions can be found in Appendix 4-1.1

Results of the assessment of using the DQEM

As stated in Chapter Three, there assessment hypotheses were developed to support the overall research hypothesis. The between- and within-SME correlations and Spearman's Rank Correlation Coefficient were used to analyze the SMEs' evaluation in support the assessment of these three assessment hypotheses:

1. As *the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the evaluations of decision quality will agree more closely (have higher inter-rater reliability).*
2. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the SMEs will be better able to differentiate among the quality of decisions.*
3. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the ability of decomposition and scoring criteria to capture the characteristics that the SMEs feel are the most important to the evaluation of decision quality will improve.*

The three separate scorings of the subjective data permitted an in-depth investigation of the effects of the decomposition decision characteristics and the use of evaluation and scoring criteria on SMEs' ability to reliably evaluate decision quality. The three scorings correspond to three levels of the decomposition of decision quality:

the initial decomposition into decision quality sub-characteristics; the definition of decision quality measures; the further decomposition to define the evaluation criteria that support the evaluation of the measures; and development of the scoring criteria that were used to tailor Likert scales for each sub-characteristic. These three scorings, described below, provide data on the SMEs' ability to consistently evaluate decision quality with respect to a progressively more detailed consensus on the characteristics of "good" decisions. The data in this case study show interesting relationships between the SMEs' subjective evaluations and the level of detail in the scoring criteria.

First Scoring – with decision sub-characteristics

In the initial scoring, the SMEs had reached consensus on the decision quality characteristics, sub-characteristics, and measures; but there had been no discussion of the effect of the specific measures on evaluation of sub-characteristics. Therefore, the SMEs were using only their understanding of the sub-characteristic and their experience for their initial evaluation of the decision quality sub-characteristics.

Second Scoring – with measure consensus

For this scoring, the SMEs had reached consensus on the decision quality measures. The SMEs had not yet reached consensus on evaluation criteria or on the association of performance levels for each measure with Likert scale values (scoring criteria). In this scoring, essentially, the decision-quality characteristics had been decomposed into sub-characteristics, the detailed measures had been developed, but the SMEs had not reached consensus on how the to translate their evaluations into Likert scale values.

Third Scoring – with criteria consensus

This third and final scoring was conducted with decision quality decomposed down to the evaluation criteria level. Further, the SMEs had reached consensus on the combinations of specific measures with "good" evaluations which were required to earn each Likert scale score.

The correlations of the SMEs' evaluations for the first, second, and third scorings are shown in Tables 9, 10, and 11 respectively. In all the tables in this section, the correlations are calculated using both the simple and weighted averages of the SME's DQCSs to correlate with their SOEs scores.

Table 9: CS-1 – Correlations (1st scoring)

		SOE Scores	Simple Average	Weighted Average	Change due to weighting
Within-SME	SME 1	--	**0.896**	**0.935**	0.039
	SME 2	--	**0.837**	**0.840**	0.003
Between-SME		-0.098	0.500	0.346	-0.154

Table 10: CS-1 – Correlations (2nd scoring)

		SOE Scores	Simple Average	Weighted Average	Change due to weighting
Within-SME	SME 1	--	**0.822**	**0.828**	0.005
	SME 2	--	**0.851**	**0.797**	-0.054
Between-SME		0.507	0.667	0.677	0.010

Table 11: CS-1 – Correlations (3rd Scoring)

		SOE Scores	Simple Average	Weighted Average	Change due to weighting
Within-SME	SME 1	--	**0.859**	**0.876**	0.016
	SME 2	--	**0.859**	**0.847**	-0.012
Between-SME		**0.720**	**0.920**	**0.919**	-0.001

The entries in each table are interpreted as follows:

- Within-SME, Simple Average is the within-SME correlation for the specified SME in which the correlation is calculated between that SME's SOE scores and the *simple* average of the same SMEs DQCSs.
- Within-SME, Weighted Average is the within-SME correlation for the specified SME in which the correlation is calculated between that SME's SOE scores and the *weighted* average of the same SMEs DQCSs.
- Between-SME, Simple Average is the correlation that is calculated between the *simple* average of each SME's DQCSs.
- Between-SME, Weighted Average is the correlation that is calculated between the *weighted* average of each SME's DQCSs.
- Between-SME, SOE Scores is the correlation that is calculated between the *SOE scores* of each SME. There are no entries for within-SME, SOE

Scores because the only possible within-SME correlations are between each SMEs' SOE scores and averaged DQCSs.

For each correlation in the previous tables, a p-value was calculated using the procedure from Chapter Three where n = 10 for the first scoring and n=20 for the second and third scorings. These p-values can be found in Tables 12, 13, and 14. For correlations, the p-values test the null hypothesis that the correlation coefficient is the result of random chance. Because three pairs of between-SME correlations are compared in each scoring, Bonferonni correction of p = 0.0167 is used in place of p = 0.05 when determining statistical significantly in order to keep the total probability of Type I errors equal to 0.05.

Table 12: CS-1 – Correlation Significance (1st scoring p-values)

		SOE Scores	Simple Average	Weighted Average	Significance of weighting
Within-SME	SME 1	--	***<0.001***	***<0.001***	0.322
	SME 2	--	***0.002***	***0.002***	0.493
Between-SME		0.788	0.142	0.327	0.363

Table 13: CS-1 – Correlation Significance (2nd scoring p-values)

		SOE Scores	Simple Average	Weighted Average	Significance of weighting
Within-SME	SME 1	--	***<0.001***	***<0.001***	0.480
	SME 2	--	***<0.001***	***<0.001***	0.316
Between-SME		0.225	***0.001***	***.0001***	0.480

Table 14: CS-1 – Correlation Significance (3rd scoring p-values)

		SOE Scores	Simple Average	Weighted Average	Significance of weighting
Within-SME	SME 1	--	***<0.001***	***<0.001***	0.424
	SME 2	--	***<0.001***	***<0.001***	0.451
Between-SME		*<0.001*	***<0.001***	***<0.001***	0.490

For each of the correlations in the previous three tables, a p-value was calculated using the procedure from Chapter Three for the change in each correlation between scorings. These p-values can be found in Table 15. Like the between-SME correlation in each scoring, three pairs of between-SME correlations are compared between the three scorings and the Bonferonni correction of p = 0.0167 is used in place of p = 0.05 when determining statistical significance in order to keep the total probability of Type I errors equal to 0.05.

Table 15: CS-1 –Significance of Changes in Between-SME Correlations (p-values)

		SOE Scores	Simple Average	Weighted Average
1st to 2nd Scoring	Between-SME	***0.017***	0.283	0.151
2nd to 3rd Scoring	Between-SME	0.156	**0.022**	**0.027**
1st to 3rd Scoring	Between-SME	***< 0.001***	***0.010***	***0.003***

From Tables 9 through 15 several conclusions can be drawn regarding the between-SME correlations. In what follows, except as otherwise noted, a 0.05 threshold was used for declaring statistical significance.

- In the first scoring (Table 9),
    - The between-SME correlations are low (below 0.7) and there is insufficient evidence to reject the null hypothesis that the between-SME correlations due to random chance (Table 12).
    - The between-SME correlation of SOE scores is -0.098 indicating that the SMEs' SOE score were essentially uncorrelated.
    - The between-SME scores for the averaged appear to be higher than that of the SOE scores, but with p-values for the difference between the DQCS and SOE scores of 0.240 and 0.410 for the simple and weighted DQCS respectively, there is insufficient evidence to reject the null hypothesis that the correlations are the same.
- In the second scoring (Table 10):
    - There is now strong evidence, $p = 0.001$ (Table 13) to reject the hypothesis that the between-SME correlation are due to random chance.
    - All the between-SME correlations appear to have increased, but the only statistically significant difference was in the SOE scores. There is strong evidence, $p = 0.017$ (Table 15) to reject the null

hypothesis that the first and second scoring correlations are the same for the SOE correlation. There is insufficient evidence to reject this hypothesis for the correlations for the averaged DQCSs

- There is insufficient evidence to reject the null hypothesis that the SOE and averaged DQCS correlation are the same

- In the third scoring (Table 11):
  - There continues to be strong evidence, $p < 0.001$ (Table 14) to reject the hypothesis that the between-SME correlations are due to random chance.
  - All the between-SME correlations again appear to have increased, and there is strong evidence to reject the null hypothesis that the correlations of averaged DQCS scores are the same in the second and third scorings. There is insufficient evidence to reject the null hypothesis that the SOE correlations in the second and third scoring correlations are the same. However, there is strong evidence, $p <= 0.01$ for all correlation changes (Table 15) to reject the null hypothesis that all the third scoring between-SME correlation are the same as the first scoring correlations.
  - There is strong evidence, $p = 0.020$, to reject the null hypothesis that the SOE and averaged DQCS correlations are the same.

- For all Between-SME correlations from all scorings, there is insufficient evidence (Tables 12 to 14) to reject the hypothesis that the simple and weighted average DQCS correlations are same.

Table 16: CS-1 – Significance of Changes in Within-SME correlations (p-values)

		SOE Scores	Simple Average	Weighted Average
1st to 2nd Scoring	SME 1	-	0.262	0.125
	SME 2	-	0.458	0.386
2nd to 3rd Scoring	SME 1	-	0.357	0.305
	SME 2	-	0.467	0.326
1st to 3rd Scoring	SME 1	-	0.361	0.224
	SME 2	-	0.433	0.479

Unlike the results of the between-SME correlations, the within-SME correlations are relatively consistent throughout all three scorings for the correlations of SOE and averaged DQCS scores. All the within-SME correlations are high, $\rho > 0.800$ (Tables 9, 10, and 11), and for all the within-SME correlations there is strong evidence, $p <= 0.002$ (Tables 12, 13, and 14), to reject the null hypothesis that the correlations are due to random chance. For all changes among the within-SME correlations there is insufficient evidence (Table 16) to reject the hypothesis that the correlations of the different scorings are the same.

Table 17: CS-1 – Summary of Rank Correlations

		SOE Scores	Simple Average	Weighted Average
Spearman's Rank Correlation	1st Scoring	0.036	0.299	-0.137
	2nd Scoring	0.495	0.622	0.493
	3rd Scoring	0.532	0.996	0.830

Table 18: CS-1 – Summary of Significance of Rank Correlations (p-values)

		SOE Scores	Simple Average	Weighted Average
Spearman's Rank Correlation Significance	1st Scoring	0.912	0.200	0.565
	2nd Scoring	0.146	0.333	0.027
	3rd Scoring	0.114	< 0.001	< 0.001

The result of the analysis of the correlations discussed above were used to evaluate assessment hypotheses one and three; but in order to assess the ability of the SMEs using the DQEM to differentiate among the quality of decisions, a non-parametric method of assessing the agreement in the SME's ranking of their scores for each evaluation was needed. Spearman's Rank Correlation Coefficient (SRCC) was used to generate the information in Table 17. From the first scoring to the third scoring the SRCC shows a roughly linear increase. The change in the correlation coefficients is

statistically significant ($p < 0.001$) indicating that there is strong evidence to reject the hypothesis that the SRCCs in the first scoring and the third scoring are the same. Likewise, there is strong evidence ($p < 0.001$) to reject the hypothesis that the correlation of the ranks in the third scoring is due to random chance.

In addition to the analyses supporting the Assessment Hypotheses, analyses were performed to determine if biases existed between the SMEs' scores. Student t-tests and ANOVA analyses were performed on the SMEs' scores from each scoring to determine if statistically significant differences (biases) existed between the means and variances, . This data can be found in Table 19. Only the means and standard deviations from the second scoring provided strong evidence (bolded in Table 19) to reject the hypothesis that the means and standard deviations were the same. There was insufficient evidence to reject these hypotheses for the means and standard deviations from the first and third scorings.

Table 19: CS-1 – Means, Standard Deviations, and Bias Significance

Simple Averages				**Weighted Averages**				**SOE Averages**			
	SME 1	SME 2	p-value		SME 1	SME 2	p-value		SME 1	SME 2	p-value
Scoring 1	3.146	3.137	0.907	Scoring 1	3.182	3.149	0.723	Scoring 1	3.500	2.600	0.108
Scoring 2	3.153	2.668	***0.001***	Scoring 2	3.160	2.766	***0.016***	Scoring 2	3.450	2.750	***0.009***
Scoring 3	3.032	3.057	0.383	Scoring 3	3.270	3.336	0.458	Scoring 3	3.300	3.550	0.234
Simple Standard Deviations				**Weighted Standard Deviations**				**SOE Standard Deviations**			
	SME 1	SME 2	p-value		SME 1	SME 2	p-value		SME 1	SME 2	p-value
Scoring 1	0.652	0.323	0.970	Scoring 1	0.640	0.498	0.901	Scoring 1	1.179	0.966	*0.078*
Scoring 2	0.547	0.737	***0.020***	Scoring 2	0.635	0.683	***0.066***	Scoring 2	1.146	1.020	***0.048***
Scoring 3	0.742	0.669	0.908	Scoring 3	0.769	0.668	0.772	Scoring 3	0.657	1.050	0.372

Summary of results of BTRA-BC evaluation

The evaluation of the value of the GDSPs centered on five aspects of plan quality: (1) time to completion, (2) objective plan quality, (3) subjective plan quality, (4) understanding of the terrain, and (5) the decision-maker perception of the GDSPs. The first two were evaluated objectively, the second two were evaluated subjectively by SMEs, and the last was a subjective evaluation by the decision-makers. An ANVOA analysis of the data from the third scoring indicated the following:

Time to Completion – the decision-makers' average time to completion when they used the GDSSs was significantly faster than when the GDSS were not used. There was strong evidence ($p < 0.001$) to reject the hypothesis that the two average times to completion were the same.

Objective Quality – confirmed by a Wilcoxon Signed Ranks test, the decision-makers' average objective quality score when they used the GDSS was higher than when the GDSSs were not used. There was strong evidence ($p < 0.01$) to reject the hypothesis that the average objective quality scores were the same.

Subjective Quality – the decision-makers' average subjective quality score when they used the GDSS was higher than when the GDSS were not used. There was strong evidence ($p = 0.003$) to reject the hypothesis that the average subjective quality scores were the same.

Terrain Understanding – the decision-makers' knowledge and understanding of the impact of terrain was greater when decision-makers used the GDSS than when they

did not. There was weak evidence (p = 0.059) to reject the hypothesis that the decision-makers' average knowledge and understanding of the impact of terrain scores were the same.

A more detailed presentation of these results can be found in Appendix 4-1.2, and the conclusions that can be drawn from these results with respect to the usefulness of the DQEM are discussed in the Conclusion section.

Conclusions

Case Study One addressed the questions of whether the GDSS were valuable to the decision-maker and whether the DQEM was a useful method to use in evaluating decision quality. From the results of the evaluation of the GDSS and the results of the assessment of the DQEM scorings presented above, several conclusions can be drawn.

Evaluation of the Usefulness of the GDSS

Table 20: CS-1 – GDSS results

		Average		Variance	
		with GDSS	without GDSS	with GDSS	without GDSS
Time to Completion		**1.136**	**3.124**	**0.053**	**0.793**
Quality	Objective	**3.849**	**2.920**	0.392	0.371
	Subjective	**3.399**	**2.719**	**0.180**	**0.561**
Terrain Understanding		3.185	2.565	0.741	0.902

The analyses of all four characteristics of a useful GDSS, time to completion, objective and subjective quality, and terrain understanding, strongly support the primary

hypotheses that decision-makers using the GDSS produced outputs (1) faster, (2) with higher quality, and (3) showing as good, if not better, an understanding of the impact of the terrain (the change in understanding did not come near the 0.05 threshold for statistical significance).

Table 20**,** above, summarizes the data for the averages, variances, and significance for each measure. Entries in boldface indicate measures that were statistically significant at or below the p = .05 level. A more detailed discussion of the conclusions of the evaluation of the usefulness of the GDSPs can be found in Appendix 4-1.2 and Powell, et al. 2009.

These conclusions and the supporting results indicate that the DQEM was useful for evaluating the impact on decision quality of using a GDSS to aid in decision-making. Further, results support the hypothesis that the DQEM was useful in evaluating the impact of a change in a decision-making process. The statistically significant results obtained in the GDSS evaluation imply that the SMEs were able to evaluate the decision quality precisely enough using the decomposed decision quality measures that the effects of the primary independent variable (System Used) were statistically significant. These results also imply that the SMEs were able to differentiate sufficiently well among the quality of decisions that differences (or lack thereof) in the quality of the decision-makers responses could be determined. But, the generation of statistically significant results is a function of both the SMEs' ability to discriminate among decision quality and the underlying quality of decision-makers' responses, i.e., even if the SMEs could evaluate decision quality with absolute reliability, their scores could yield statistically significant

results if there were no difference in the decision-makers' responses. Determining the effect of using the DQEM on the reliability of SMEs' subjective evaluations required further analysis of the data. The discussion below focuses on the effect of using the DQEM on the subjective evaluation of decision quality, specifically the effect on the reliability of the SMEs' evaluations.

Assessment of the Effectiveness of the DQEM

The results obtained from the evaluation of the GDSPs support the assertion that an evaluation of decision quality can be used as the primary criterion in evaluating changes in a decision-making process. In addition to evaluating the GDSS, the experimental structure and the data the experiment generated also allowed an exploration of the three assessment hypotheses:

1. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the evaluations of decision quality will agree more closely (have higher inter-rater reliability).*
2. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the SMEs will be better able to differentiate among the quality of decisions.*
3. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the ability of decomposition and scoring criteria to capture*

the characteristics that the SMEs feel are the most important to the evaluation of decision quality will improve.

The data collected to explore these three hypotheses was also used to investigate whether the use of weighed averages of SOE scores and DQCSs would result in more reliable overall decision quality evaluation scores. The discussions below concentrate on how well the experimental results support the three assessment hypotheses.

Because there are no ground truth scores for either the quality of individual characteristics or the quality of the overall decision in the complex, ill-structured problem used in this case study, the four effects mentioned above were investigated through the analysis of the SMEs' within- and between-SME correlations and the ranking of each SME's overall evaluation scores. Between-SME correlations are the correlations between the SOE scores and averaged DQCS scores resulting from the SMEs' independent evaluations of the decision-makers' decisions. The between-SME correlations are a measure of the level of agreement between, the reliability of, the SMEs' evaluations. Within-SME correlations, on the other hand, are correlations between each SME's SOE scores and the averages of their DQCSs. The within-SME correlations are a measure of how well the decomposition of decision quality characteristics reflected SMEs' understanding of the most important elements and relationships involved in the decisions. Spearman's rank correlation coefficient, unlike Pearson's correlation coefficient used in the between- and within-SME correlations, is a non-parametric measure of the similarity of two rankings that is used to assess the SMEs' ability to differentiate among the quality of decisions.

Assessment Hypothesis 1

The evaluation of assessment hypothesis 1 centers around the between-SME correlations that are summarized in graphic form in Figure 9 below. This summary supports the first assessment hypothesis:

> *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the evaluations of decision quality will agree more closely (have higher inter-rater reliability).*

In order to support this hypothesis, the between-SME correlations would have to demonstrably increase over the course of the three scorings. Upon visual inspection of Figure 9, trends in the data seem to support the assessment hypothesis 1 and provide some additional insight into the SMEs' ability to subjectively evaluate decision quality. First, all three between-SME correlations (SOE scores, simple and weighted averages of DQCS), increased over the three scorings. Second, the correlations of average DQCS are higher than the correlation of SOE score in all scorings. Third, there is little difference between the correlations of simple and weighted average DQCSs. The p-values associated with the correlations support the visual analysis of the data.

The first scoring in which the decomposition was to the sub-characteristic level was the SMEs' first evaluation; and they evaluated a subset, 10 decisions, of the decision-makers' decisions. Since the decomposition was at the coarsest granularity, the correlations from this scoring served as the baseline with which the correlations were

compared. In this scoring, the SMEs' evaluations are not in very close agreement. The simply averaged DQCS has the highest correlation, $\rho = 0.500$; and the SMEs' SOE scores have the lowest, $\rho = -0.098$, which indicates they were essentially unrelated even though the SMEs evaluated the same decisions. An analysis of the statistical significance of these correlations yields p-values for each correlation that do not approach significance providing further evidence that these scores are not in close agreement. These low correlations indicate that, with only the decision quality characteristics and sub-characteristics defined, the variation between the SMEs' overall evaluations, both for SOEs and when evaluating individual sub-characteristics, were too great for the evaluations to be useful in evaluating decision quality.

This variation in the SMEs' scores is likely due to two factors: one, the almost total reliance on the SMEs' personal experience in determining what factors contribute to the quality of each sub-characteristic; and, two, the SMES' subjective interpretations of Likert scale categorical labels. First, in this scoring the SME's had not reached consensus on the measures that would support the evaluation of the sub-characteristics; and without a sufficient relevant detail in the decomposition to fully capture the SMEs' understanding of the problem or to provide a common basis for evaluating decision quality, the evaluators were individually defining the measures associated with each sub-characteristics; in essence the evaluators were using different measures in their evaluations. Also, in this scoring the SMEs were using standard Likert scales with vague categorical labels. These imprecise labels required that the SMEs' subjectively translate these labels into scoring criteria, thus inserting a source of variation into the SMEs'

scores. Overall, the evaluations of decision quality from the firsts scoring were not reliable enough to be useful.

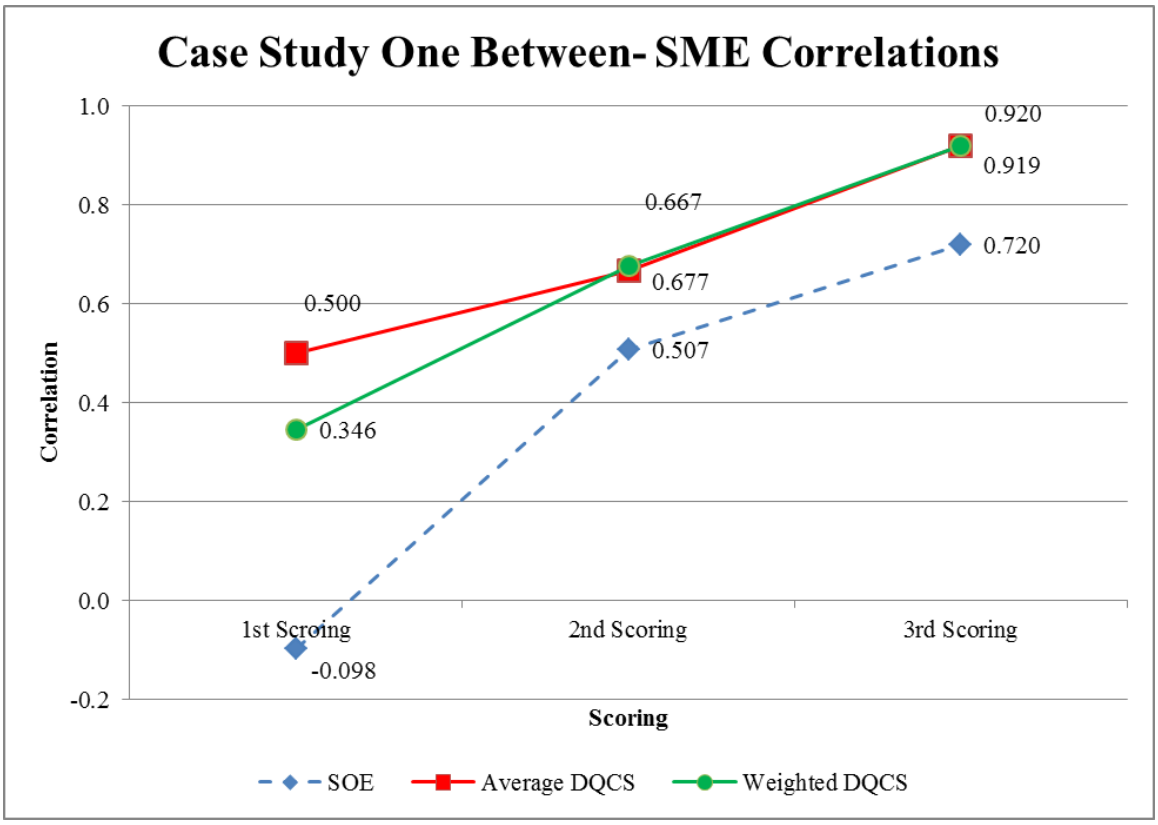

Figure 9: CS-1 – Between-SME Correlations

In the second scoring, the SMEs had reached consensus on the measures that supported the evaluation of the sub-characteristics; but they had not defined the specific evaluation criteria that supported the evaluation and the decision quality measures. Thus, in the second scoring, the level of relevant detail in the decomposition was greater than in the first scoring and the decomposition provided additional information that clarified the

SME's understanding for the problem. In this scoring, all the between-SME correlations had increased from their respective values in the first scoring. Even though all the correlations are still too low ($\rho < 0.700$) to indicate close agreement among the SMEs' evaluations, the p-values associated with the correlations of averaged DQCSs ($p = 0.001$ for both correlations) indicate that these correlations, unlike those from the first scoring, are a true estimate of the agreement of the SMEs' evaluations.

The SMEs' evaluations seemed to be in closer agreement in the second scoring, and the reliability of their evaluations had also improved. Because the only change in the evaluation procedure was the increase in the relevant information in the decomposition, that increase in reliability was due in some measure to the increased granularity of the decomposition providing more guidance to the SMEs as they evaluated 20 different series of decisions (plans). Even though the between-SME correlations were higher in this scoring, the SMEs' evaluations were not reliable enough to adequately differentiate among the quality of decisions (see the section on assessment hypothesis 2). This lack of sufficient reliability likely was due the continued use of standard Likert scale categorical labels introducing statistical noise into the SMEs' scores. This source of variation was corrected in the third scoring.

In the third scoring the SMEs' had reached consensus on the evaluation and scoring criteria. This consensus resulted in the finest granularity for the decomposition of decision quality and led directly to the construction of the tailored Likert scales. In this scoring all the between-SME correlations again increased. The correlations of averaged DQCSs were very high ($\rho = 0.920$ and $\rho = 0.919$ for the simple and weighed averages,

respectively) and $\rho = 0.72$ for the correlation of SOE scores. All the correlations were statistically significant ($p < 0.001$) indicating that there was a relationship among the SMEs' evaluations. The high levels of reliability indicated by the correlations implied that the portion of the SMEs' understanding captured in the decomposition combined with the use of tailored Likert scales reduced variation in the SMEs' evaluations below that in the previous scorings. Because the level of detail in the decomposition was increased (evaluation criteria defined) in the same scoring that the scoring criteria were used to construct the Likert scales, there is no information on the relative effect of these two changes on the SMEs' evaluations. But, together the decomposition and the tailored Likert scales aided the SMEs in the production of very reliable evaluations.

The previous discussion of the three scorings concentrated on the apparent increase in the between-SME correlations. An analysis of the changes in these correlations over the three scorings confirm that the between-SME correlations did increase significantly. The overall changes, from the first to third scorings, in all the between-SME correlations are significant at the Bonferonni corrected level of $p = 0.01695$ (Table 15) This level of significance combined with the increased correlations indicates that the correlations were the highest in the third scoring. This increase in the between-SME correlations indicates that the SMEs' subjective evaluations, both their subjective overall evaluations (SOEs) and their evaluations that were captured by the average of their decision quality characteristic scores (DQCSs), became more reliable over the course of the three scorings.

The only significant change in the SMEs' between-SME correlations of averaged DQCSs occurred between the second and third scorings with p-values of $p = 0.022$ and $p = 0.027$ for the simple and weighted average correlations respectively. These changes were significant at the $p = 0.05$ level and approached the Bonferroni corrected significance level of $p = 0.1695$. This implies that the definition of evaluation and scoring criteria likely had a greater effect on the increased reliability of the SMEs' evaluations than the development of decision quality measures. Since the definition of evaluation criteria could be considered an extension of the decomposition, it seems likely that the development of the scoring criteria and the construction of the tailored Likert scales had a large impact on reducing variation in the SMEs' scores and the improvement in the reliability of their evaluation.

As seen in Figure 9, there appears to be little difference between the simple and weighed average DQCS correlations; but these correlations seem to be greater than the correlation of SOE scores. While there is insufficient evidence to reject the hypothesis that the averaged DQCS correlations are the same, there is strong evidence ($p = 0.020$) to reject the hypothesis that the averaged DQCS and SOE correlations are the same. Given this evidence, two conclusions can be drawn: first, it is safe to say that the correlations of the SMEs' average of the DQCS yielded more reliable overall evaluations of decision quality than did the SOE scores. This could be due in part to the average DQCS using a consistent process for aggregating the DQCSs that takes into account all the sub-characteristics in the decomposition. In contrast, the SMEs' generation of SOE scores is entirely subjective and unique to each SME. There is no guarantee that the SMEs'

mental aggregation will be consistent and, therefore, could be a source of variation in the SOE scores that is not present when using a consistent process such as simple or weighed averaging. Second, there is insufficient evidence that the weighted average DQCS correlations improve the reliability of the SMEs' evaluations. Given this lack of evidence and the general difficulty in eliciting weights from SMEs, the effort required to elicit weights was probably not well spent.

In summary, the results obtained from the three scorings support Assessment Hypothesis 1. First, the consistently higher correlations between the SMEs' averaged DQCS support the conclusion that these scores were better estimators of decision quality that the SME's SOE scores. Second, the consistently increasing agreement between the SMEs' scores though successive scorings support the conclusion that the better the decomposition represented the SMEs' subjective understanding of the problem, the more reliable were the evaluators' assessments of decision quality. Third, the lack of significant change in the between-SME correlations due to weighting the average of the DQCS confirmed the previous conclusion that weighting does not consistently improve the ability of the SMEs' scores to estimate relative decision quality.

Assessment Hypothesis 2

Although the analysis of the between-SME correlations provided strong evidence that the DQEM aids SMEs in reliably evaluating decision quality (Assessment Hypothesis 1), the between-SME correlations are not appropriate to test Assessment Hypothesis 2:

As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the SMEs will be better able to differentiate among the quality of decisions.

Even though the Pearson's Correlation Coefficient uses paired data, it is a gauge of the overall reliability of the set of evaluations, but it provides little information on the agreement of the SMEs' evaluation on specific decisions. In order to compare the SMEs' evaluations on each decision-makers' decision, Spearman's Rank Correlation Coefficient (SRCC) was use to compare the rankings of each SME's evaluations. Figure 10 shows the SRCC for each of the averaged DQCS and SOE scores.

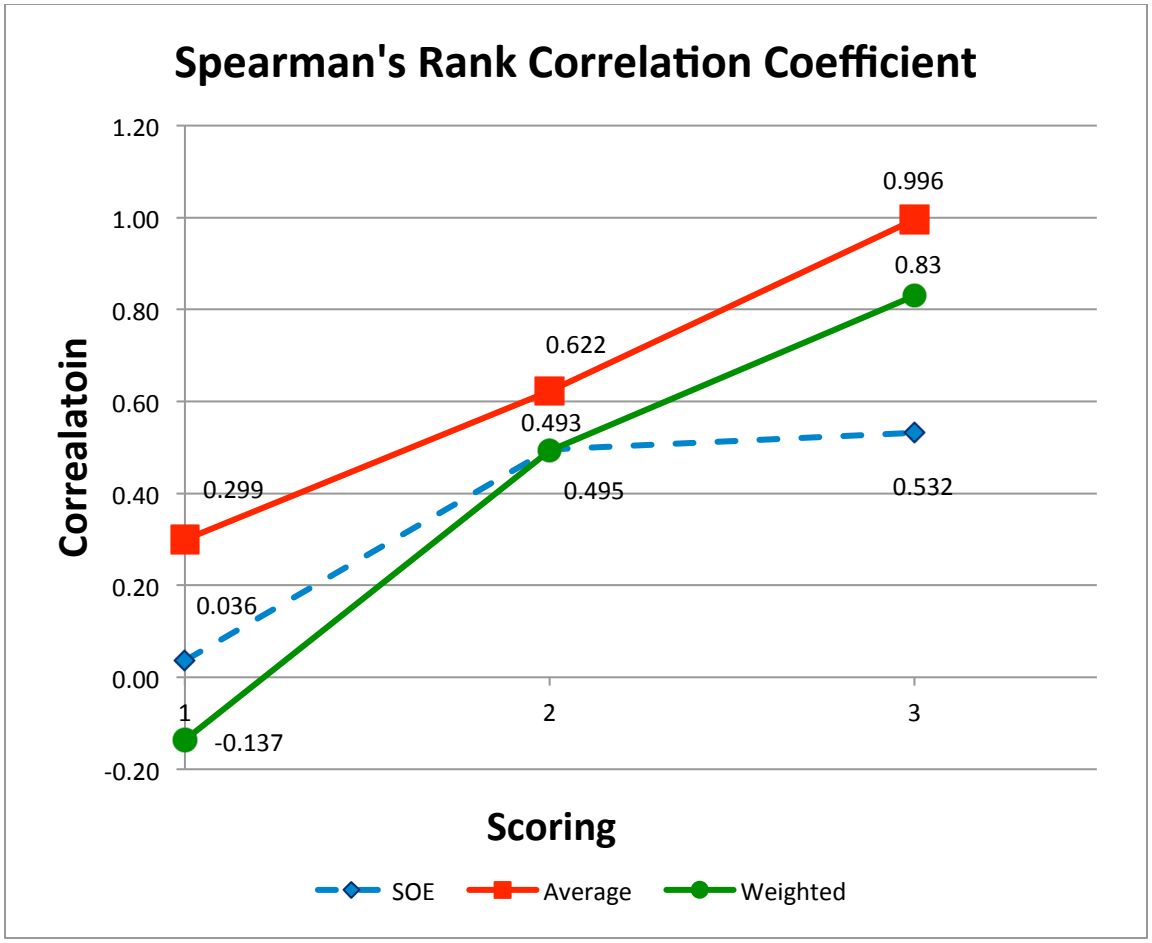

Figure 10: CS-1 – Spearman's Rank Correlation Coefficient

From inspection of Figure 10, several conclusions can be drawn. First, for each set of evaluation scores, the SRCC increases with each scoring. Second, the SRCC for the simply averaged DQCS is greater than either the weighted average of the DQCS or the SOE scores. Third, while the SRCC for the averaged DQCS appears to increase nearly linearly over the three scorings, the SRCC for the SOE scores seems to plateau after the second scoring. This inspection suggests that like the conclusion from the analysis of the between-SME correlations, either the simple or weighted average of the

DQCS would be able to discriminate among the quality of decisions; but given that the SRCC of the simply averaged DQCS is consistently higher than that for the weighted average of DQCS, the simply averaged DQCS would be better able to differentiate among the quality of decisions. The comparison of the SMEs' ranked average DQCS and their simply averaged DQCS in Figures 12 and 13 respectively demonstrates the ability of the SMEs to discriminate among the quality of decisions using the DQEM.

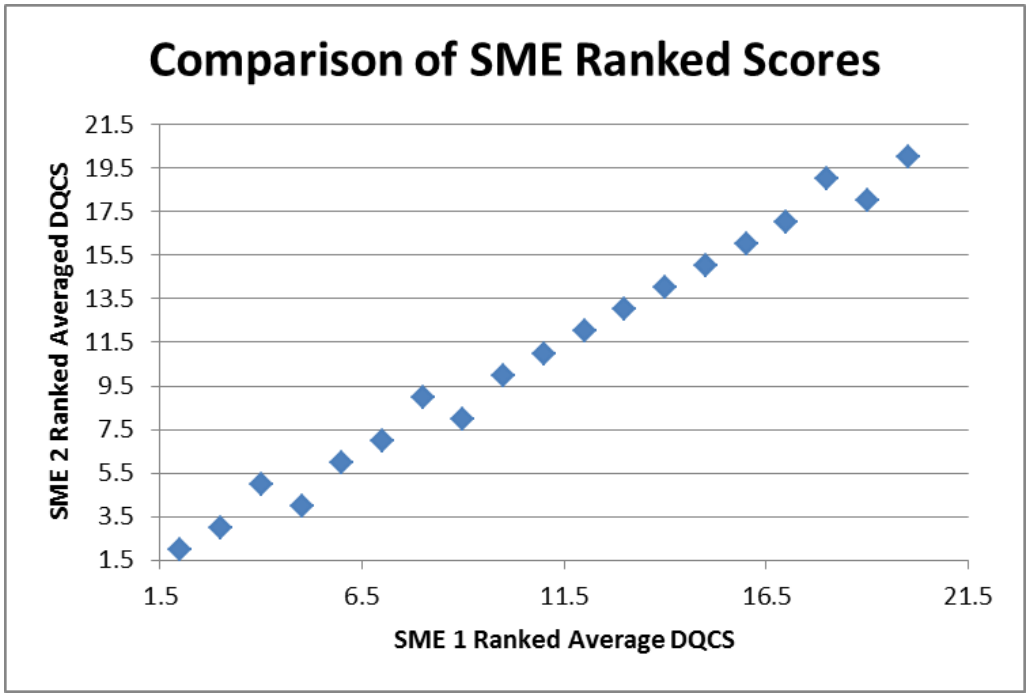

Figure 11: CS-1 – Comparison of Ranked SME Average Scores

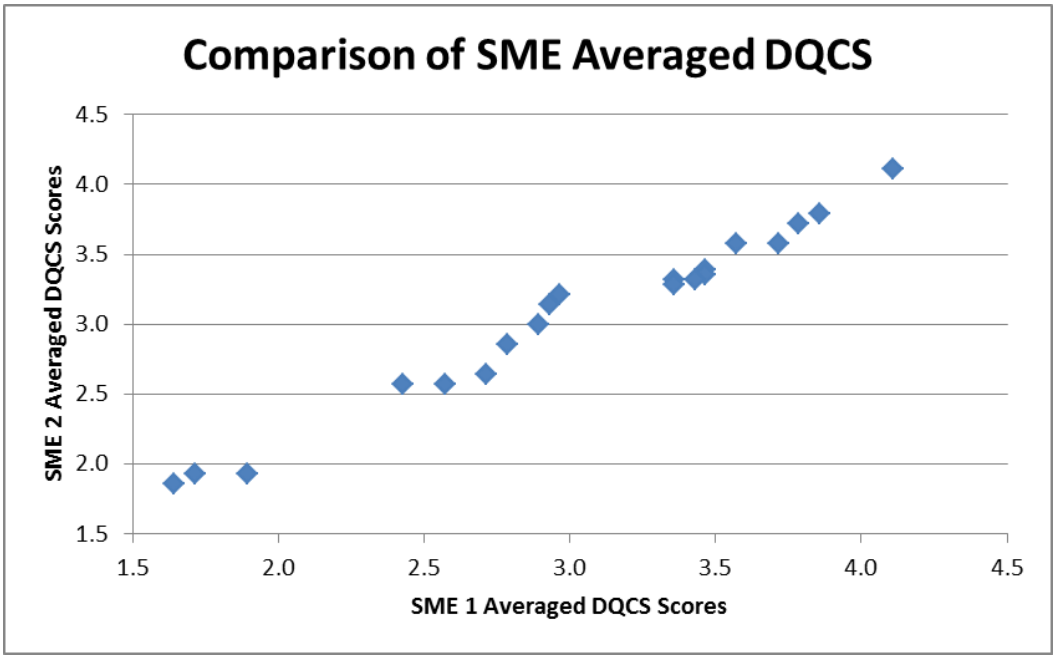

Figure 12: CS-1 – Comparison of SMEs' Averaged DQCS

Assessment Hypothesis 3

As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the ability of decomposition and scoring criteria to capture the characteristics that the SMEs feel are the most important to the evaluation of decision quality will improve.

As stated above, Assessment Hypothesis 3 was intended to discover if, as the decision quality characteristics were further decomposed, the decomposition better captured the decision quality characteristics that the SMEs' felt were important to the evaluation of decision quality. From the data collected in the three scorings, there is insufficient evidence to support the hypothesis that additional decomposition beyond

decomposition to the sub-characteristic level (as in the first scoring) affected the ability of the decomposition to capture the characteristics that the SMEs feel are important. Because of the structure of the assessment, there is also insufficient evidence to support that less detail in the decomposition captures less of the important characteristics.

In all three scorings, the within-SME correlations of both SMEs are consistently within the range of $\rho = 0.797$ to $\rho = 0.935$ with p-values for all correlations for all scorings of $p < 0.0001$. Also, the p-values for the changes in the within-SME correlations between scorings and the p-values comparing the within-SME correlations in each scoring are not statistically significant at the $p = 0.050$ level. Overall this indicates that there is a consistently high level of agreement between each SME's SOE scores and their simple and weighted averages of DQCSs. This consistently high level of agreement may be due decision-makers typically using between five and nine factors when making unaided decisions (Zsambok, 1997). Since the decomposition in the first scoring consisted of more than nine sub-characteristics, and since the within-SME correlations were high in this scoring two conclusions can be drawn: First, the initial decomposition captured the five to nine characteristics that the SMEs felt contributed most to a good decision. Second, further decomposing the decision quality characteristics into measures and evaluation criteria did not impact the degree to which the decomposition reflected the characteristics the SMEs felt were most important.

SME Bias

There does not appear to be a reason that reaching consensus on the decision quality measures should have resulted in differences between the means and standard

deviations in the second scoring. Intuitively, reaching consensus on the measures should have captured more of the SMEs' understanding and was expected to produce less bias. Conversely, the lack of evidence of biases in the third scoring seems indicate that the incorporation of tailored Likert scales in the third scoring seems to have overcome source of the bias in the second scoring. If the use of the tailored Likert scales did reduce the biases, this would be an important factor in improving the SMEs' ability to reliably evaluate decision quality.

Summary

Overall the results gathered from the three scorings support the research hypothesis:

> *The direct evaluation of the quality of decisions made to address complex, ill-structured problems can be improved through the use of a structured subjective decomposition of decision quality characteristics.*

The evaluation of the research hypothesis is supported by the statistical analysis of the subjective evaluations of the quality of decisions made to address complex, ill-structure problems by SMEs. The three assessment hypotheses that were addressed by the statistical analysis were:

1. *As the decomposition of decision quality and the scoring criteria become more detailed, and better reflect and further clarify the SMEs' understanding of the problem, the evaluations of decision quality will agree more closely (have higher inter-rater reliability).*

2. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the SMEs will be better able to differentiate among the quality of decisions.*
3. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the ability of decomposition and scoring criteria to capture the characteristics that the SMEs feel are the most important to the evaluation of decision quality will improve.*

In summary, the data collected supported assessment hypotheses 1 and 2 but did not support assessment hypothesis 3:

- An analysis of the between-SME correlations generated from the SMEs' evaluations supported Assessment Hypothesis 1 in that the analysis indicated that the DQEM including the decomposition of decision quality, a procedure for developing tailored Likert scales and the aggregation of decision quality characteristics using a simple average of decision quality characteristics, resulted in reliable evaluations of decision quality by the SMEs.
- In support of assessment Hypothesis 2, an analysis of the ranking of the SMEs' SOE scores and averaged DQCSs indicated that as the level of detail in the decomposition increased, the SMEs' evaluations were significantly better able to differentiate among the quality of decisions.

- Even though the analysis of the within-SME correlation was not able to support assessment Hypothesis 3, the consistently high correlations indicated that the decomposition of decision quality characteristics was consistently able to capture the characteristics that the SMEs thought were important.
- As part of the analysis of assessment Hypothesis 1, an exploration of the effect of using weighted and simple averages to calculate overall decision quality scores was conducted. In this case study, there was insufficient evidence that using a weighted average to aggregate the scores of individual sub-characteristic evaluations improved the reliability of the overall scores.

Finally, as evidence that the DQEM can be used to successfully evaluate decision quality in complex, ill-structured problems, this case study evaluated the impact of augmenting a decision-making process with a GDSS. The results of this evaluation were that SMEs using the DQEM incorporated into the experimental design were able to sufficiently distinguish among the differences in the quality of decision in order to generate statistically significant results indicating that the augmenting of the MDMP with a GDSS improved the quality of the decision-makers decisions.

Overall, the support for assessment Hypotheses 1 and 2 support the overall research hypothesis. The Decision Quality Evaluation Method (DQEM) used a structured, subjective decomposition of decision quality characteristics to capture the SMEs' understanding of the elements and relationships of a decision, develop a unique scoring procedure, and aggregate the decision quality evaluation scores for individual sub-characteristics into overall scores of decision quality. SMEs were able to use the

decomposition and scoring procedure to reliably evaluate and score the quality of decision quality sub-characteristics. The SMEs' evaluations were sufficiently reliable that the direct evaluation of decision quality was successfully used to evaluate a modification to a decision-making process.

Next step

The usefulness of the DQEM and the experimental procedures described in CS-1 demonstrated the potential benefit of using the DQEM to directly assess relative decision quality as the primary measure in an overall assessment of decision-making. The results generated in Case Study One demonstrated the usefulness of the DQEM in one specific context, with one set of decision tools, with one set of decision-makers, and with one set of decision-making skills. Although the success of the DQEM in directly assessing relative decision quality in a single case study can be extrapolated to usefulness for assessing decision quality in general, the use of the DQEM in the successful evaluation of decision quality in other contexts would lend additional support to the general usefulness of the DQEM for the evaluation of decision quality. To support the general usefulness of the DQEM, Case Study Two describes another evaluation in which the DQEM was used to generate statistically significant results that addressed decision-making in a context different from that of CS-1. The discussion of CS-2 in Chapter 5 summarizes an evaluation which used the DQEM as the assessment tool which addressed decision-making in an ill-structured problem with different complexity, with decision-makers having different decision-making skills, and using SMEs with different expertise.

CHAPTER FIVE: RESULTS AND CONCLUSIONS FROM CASE STUDY TWO

The purpose of the second case study is to demonstrate the flexibility and the use of the DQEM in an evaluation that assesses the relative quality of decisions for a problem that is more complex, has different targeted users, and that evaluated other GDSSs. Case Study One yielded statistically significant evidence that using the DQEM improves the ability of SMEs to more reliably evaluate the quality of decisions made with respect to complex, ill-structured problems. Using the DQEM allowed the SMEs who evaluated the decision-makers' decisions (in the form of plans) to be more reliable in their evaluations and resulted in the differentiation among the relative quality of decisions. Even though CS-1 demonstrated the effectiveness of using the DQEM for that specific problem, it is important to assess the generalizability of these results. For this reason, a second case study was undertaken to assess the usefulness of the DQEM in the evaluation of decision quality in a more complex problem. The discussions of Case Study Two center around the differences between the complexities of problems in the two case studies and the assessment of the DQEM in a more complex problem. The differences that are highlighted in the following case study are used to demonstrate the adaptability of the DQEM to problems of varying complexity and at different levels of decision-making.

Scope of Case Study Two (CS-2)

Like Case Study One (CS-1), the primary goal of the evaluation was the determination of the impact of the GDSS functions on military decision-making for the U.S. Army Engineer Research and Development Center (ERDC). Also like Case Study One, the goal of Case Study Two was to assess the effectiveness of the DQEM in assessing relative decision quality as the primary measure in the evaluation. The GDSS functions used in CS-2 were Tier 2 TSOs which support more complex decisions than the Tier 1 TSOs evaluated in CS-1. These more complex decisions resulted in the use of the DQEM in an evaluation of more abstract and complex decision-making than in CS-1. Integrating the DQEM into CS-2 followed the same steps described in Chapter Four; but the missions, tasks, scenario, and the decision-makers characteristics used in CS-2 needed to be appropriate to the complexity of the decisions under evaluation. Toward that end, the missions, tasks, and scenario in CS-2 supported the generation of Courses of Action (COAs). A military COA is a plan designed to accomplish the mission and contains sufficient detail to be the basis for an Operations Order. The decisions required to generate COAs encompass entirely and expand upon the decisions required in CS-1 which were concerned with generation of specific routes in support of a COA. To support the generation of COAs in CS-2, the decision-makers were required to be more familiar with operational planning and thus more senior to the terrain analysts used in CS-1.

Table 21, below, compares the major design factors incorporated into CS-1 and CS-2. The missions and outputs for CS-2 expand upon those for CS-1. Because the

decisions involved in generating the COA were so complex, a digital (graphic) plan could not adequately convey the basis for the decision-makers' decisions. Therefore, the decision-makers produced an abbreviated written Operations Order (Opord) which aided the SMEs' assessment of the quality of their decisions. As appropriate to the problem, CS-2 used a different study population (trained staff planners rather than terrain analysts) and a smaller unit size (battalion rather than brigade) than CS-1.

Table 21: CS-2 – Comparison of Design Considerations

Case Study	**One**	**Two**
GDSS	BTRA-BC Tier 1 TSOs	BTRA-BC Tiers 1 & **2** TSOs
Level of Decision-maker	Terrain Analysts	Staff Planners
Number of decision-makers (actual/design)	18 / 16	8 / 16
Host Environment	DTSS	Commanders' Support Environment (CSE)
Unit Size	Brigade	Battalion
Mission	Tactical Movement	Tactical Movement to seize an Objective in the presence of hostile forces
Tasks	Terrain Analysis	
	Recommend Avenues of Approach (routes)	Develop Course of Action (COA)
Output	Digital Plan	Digital Plan & Written OPORD
	Terrain Understanding Questionnaire	
	Comparison Questionnaire	
	Post Trial Discussions	

Of note, the availability of staff planners needed as decision-makers for CS-2 was significantly more limited than that of the terrain analysts who participated in CS-1. As a result, only eight decision-makers were available in the time frame in which CS-2 was conducted. This limited availability of highly experienced military personnel for CS-2 and the limited availability of experienced personnel in general emphasize the need for highly cohesive evaluations that can generate statistically significant results with few decision-makers.

Case Study Two Development

As was discussed above, the mission, tasks, and TSOs used in CS-2 encompass and expand on those used in CS-1. The Tier 2 TSOs are designed to support decisions that are fundamentally different from those supported by the Tier 1 TSOs in the previous case study.

The decisions supported in CS-1 were the identification of plan elements while in CS-2 the TSOs support the evaluation of possible plan elements. The Tier 2 TSOs combine the doctrinal requirements, output of Tier 1 TSOs, and user input to rate geographic areas with respect to the minimum doctrinal requirements and user-input thresholds. The TSO output is generated in such a manner that the decision-maker has access to sufficient information to be able to make judgments concerning the merits of the various geographical areas with respect to mission goals. The critical aspects of the Tier 2 TSOs can be found in Appendix 5-2, and examples of outputs of Tier 2 TSO can be found in (Powell et al., 2010).

Problem definition

Like the previous case study, the evaluation of the impact of the functions of the GDSS on the decision-making process required the generation of a problem definition and an iterative process that identified and meshed the products of the decomposition of decision quality (characteristics, sub characteristic, and measures) with missions, tasks, and sub-tasks. Since this second case study is evaluating the same general types of GDSS in a similar type of planning problem, the six general hypotheses from the first case study are still applicable. But unlike these hypotheses, the decision quality decomposition needed to be tailored to reflect the specifics of the mission and tasks and visa-versa. The specifics of the mission and tasks can be found in Appendix 5-1, but the decomposition of decision quality needs further discussion.

Table 22: CS-2 – Comparison of Decision Quality Characteristics

Comparison of Decision Quality Characteristics	
CS-1	**CS-2**
Quality of MCs	Use of routes
Quality of CPs	Use of NAIs
Quality of potential AoAs	Use of EAs
Quality of recommended AoAs	Use of BPs
Quality of Bn Boundaries	Use of APs
	Overall Integration of Information

Table 22 compares the decision quality characteristics determined by the SMEs for CS-1 and CS-2. These decision quality characteristics in CS-2 are phrased as the *use of* instead of the *quality of* because the complexity of the problem necessitated that the evaluations of decision quality be based on both quality of GDSS outputs and their integration into the decision-makers' decisions. Because the selections of decision components are highly interrelated in a decision of this complexity, the "best" decisions will be ones that support the overall mission; and the selected decision components may not necessarily be the ones considered "best" if considered in isolation. To capture this aspect of the decision quality, the *use of* was to be decomposed into sub-characteristics that identify both the relative quality of the plan comments and the how well the components support the overall decision. The additional step of integrating decision components required additional decisions which were captured in the decomposition. Table 23 numerically summarizes the decomposition of decision quality in CS-1 and CS-2 and gives a rough idea as to the relative complexity of the hierarchical decomposition trees. Since the scope of the mission and tasks in this case study was more complex and broader in scope those than in CS-1, the decomposition of decision quality required more sub-characteristics to be defined in order to reach sub-characteristics that could be narrowly enough defined to generate measures. The entire decomposition and the rationale for each sub-characteristic can be found in Appendix 5-3.

The Tier 1 TSO information has been largely incorporated into subjective measures of the Tier 2 TSOs.

Due to the type of planning and the planning skills required by the scenario and tasks, the qualifications desired in the representative decision-makers are listed in Table 24 and are compared with the qualifications of the decision-makers in CS-1.

Table 23: CS-2 – Decision Quality Decomposition Comparison

CS-2 Decision Quality Decomposition Comparison		
Case Study	CS-1	CS-2
decision quality characteristics	5	6
1st level sub-characteristics	13	22
2nd level sub-characteristics	16	25
3rd level sub-characteristics	0	4
measures	35	53

Table 24: CS-2 – Comparison of Decision-Maker Qualities

CS-1	CS-2
Army or Marine Corps E-5 to CWO-3	Army or Marine Corps Officers (O4-O6)
Formal training in terrain analysis (Basic Terrain Analysis School)	Formal training in military planning (command and General Staff College or equivalent)
Experience in a terrain analysis staff position	Experience on a battalion or above staff planning operations
Experience with computer-aided decision support tools	
Familiarity current terrain analysis tools	Familiarity current planning tools

Decision-makers with this experience should be familiar with the planning tasks they would be asked to complete. Since decision-makers with these qualities are in demand in active duty units, finding decision-makers was difficult. Within the timeframe of the evaluation, only 8 decision-makers were available. The results from this case study were generated from the responses of these eight decision-makers.

Experimental structure and procedures

CS-2 had the same basic experimental structure and followed the same basic procedures as described in Chapters Three and Four for CS-1. The experimental structure and procedures were tailored for the number of decision-makers and the specific graphic and written outputs required. The required outputs are summarized as part of Table 21.

Determine scoring criteria

Lessons learned from CS-1 were used in the decomposition of decision quality for establishing the scoring procedures in CS-2. Like CS-1, three scorings were conducted; and only the data from the last scoring was used for the evaluation of the GDSS. But unlike CS-1, only the first two scorings were used in the assessment of the DQEM. Because the of the low correlation and lack of reliability demonstrated in the first scoring in CS-1, repeating a scoring with decision quality characteristics decomposed only to sub-characteristics was deemed not to be an effective use of scarce resources (SMEs). Thus the first two scorings in CS-2 were comparable to the second and third scorings in the previous case study. The first scoring in CS-2 was conducted after reaching

consensus on the decision quality measures, and the second scoring was conducted after reaching consensus evaluation and scoring criteria.

Due to the limited number of decision-makers, a third scoring was conducted to achieve a consensus on the score awarded for each decision quality measure, thus resolving any remaining differences in the SMEs' scores after the second scoring and removing any variation in the data due to the individual SME's scores. Although the between-SME correlations, discussed later, indicated that the SMEs' evaluations were reliable, the limited number of decision-makers suggested that the additional time and effort needed to generate final consensus scores would increase the likelihood of achieving statistically significant results for the evaluation of the GDSS.

Determine aggregation criteria

Since aggregation with weighted averages was not found to be superior in CS-1, the decision quality data in CS-2 was aggregated using simple averages of decision quality characteristic scores (DQCSs).

Results

As discussed above, CS-2 provided data for both the evaluation of the impact of augmenting the military decision making process (MDMP) with the GDSS for military decision-making and the assessment of the effectiveness of the DEQM in improving the reliability of the SMEs' subjective evaluations of the relative quality of decisions. A summary of the results of the evaluation of the GDSS and the results of the assessment of DQEM are presented below.

Results of the assessment of using the DQEM

As stated in Chapters Three and Four, three assessment hypotheses were developed to support the overall research hypothesis. The between- and within-SME correlations and Spearman's Rank Correlation Coefficient were used to analyze the SMEs' evaluation to support the assessment of these three assessment hypotheses:

1. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the evaluations of decision quality will agree more closely (have higher inter-rater reliability).*
2. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the SMEs will be better able to differentiate among the quality of decisions.*
3. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the ability of decomposition and scoring criteria to capture the characteristics that the SMEs feel are the most important to the evaluation of decision quality will improve.*

The two independent scorings of the subjective data permitted an in-depth investigation of the effects of the decomposition decision characteristics and the use of evaluation and scoring criteria on SMEs' ability to reliably evaluate decision quality. The two scorings correspond to two levels of the decomposition of decision quality: (1)

the initial decomposition into decision quality sub-characteristics and the definition of decision quality measures, and (2) the further decomposition to define the evaluation criteria that support the evaluation of the measures and the development of the scoring criteria that was used to tailor Likert scales for each sub-characteristic. These two scorings, described below, provide data on the SMEs' ability to consistently evaluate decision quality with respect to a progressively more detailed consensus on the characteristics of "good" decisions. The data in this case study show interesting relationships between the SMEs' subjective evaluations and the level of the SMEs' consensus understanding of the problem as incorporated into the scoring criteria.

First Scoring – with measure consensus

For this scoring the SMEs had reached consensus on the decision quality measures. The SMEs had not yet reached consensus on evaluation criteria or on the association of performance levels for each measure with Likert scale values (scoring criteria). In this scoring, essentially, the decision-quality characteristics had been decomposed into sub-characteristics; the detailed measures had been developed; but the SMEs had not reached consensus on how the to translate their evaluations into Likert scale values.

Second Scoring – with criteria consensus

This second and final scoring was conducted with decision quality decomposed down to the evaluation and scoring criteria level. The SMEs had reached consensus on the combinations of specific measures with "good" evaluations that were required to earn a specific Likert scale score.

Table 25: CS-2 – Summary of Correlations

	SME 1 Within-SME Correlation	Between-SME correlation		SME 2 Within-SME Correlation
		SOE	Averaged	
1st Scoring	0.606	0.494	0.621	0.302
2nd Scoring	0.812	0.708	0.859	0.855

The correlations of the SMEs' evaluations for the first and second scorings are shown in Table 25. In all the tables in this section, the within- and between-SME correlations are calculated using only the simple averages of the SMEs' DQC scores.

The entries in each table are interpreted as follows:

- SME [1 or 2] Within-SME correlation is the within-SME correlation for the specified SME in which the correlation is calculated between that SME's SOE scores and the *simple* average of the same SME's DQC scores.
- Between-SME, Averaged is the correlation that is calculated between the *simple* average of each SME's DQC scores.
- Between-SME, SOE is the correlation that is calculated between the *SOE scores* of each SME.

For each correlation in the previous tables, a p-value was calculated using the procedure from Chapter Three with n = 16 for both the first and second scorings. These p-values can be found in Table 26. For correlations, the p-values test the null hypothesis that the correlation coefficient is the result of random chance. Because only two pairs of between-SME correlations are compared, a Bonferonni correction was not required.

Table 26: CS-2 – Correlation Significance (p-values)

	SME 1 Within-SME Correlation	Between-SME Correlation		SME 2 Within-SME Correlation
		SOE	Averaged	
1st Scoring	0.056	0.107	0.050	0.234
2nd Scoring	<0.001	0.001	<0.001	<.0.001

For each of the correlations in the previous tables, a p-value was calculated using the procedure from Chapter Three for the change in each correlation between scorings. These p-values can be found in Table 27. Unlike CS-1, only two sets of data are being compared for each correlation, so a Bonferonni correction is not required.

Table 27: CS-2 – Significance of Correlation Changes (p-values)

	SME 1 Within-SME Correlation	Between-SME Correlation		SME 2 Within-SME Correlation
		Subjective	Averaged	
1st to 2nd Scoring	0.136	0.192	0.075	0.007

Since there is data from the equivalent scorings in Case Study One, the results from the equivalent scores can be roughly compared, The comparison can only be rough because the scoring SMEs, the decision-makers, and the decompositions were all different; and all the variance reduction advantages of a within-subject experimental design are lost when comparing different experiments. Tables 26, 27, and 28 compare the within- and between-SME correlations, the significance of those correlations, and the

significance of the changes in the correlations between the equivalent scorings in each case study.

Table 28: CS-2 – Comparison of Correlations

	SME 1 Within-SME Correlation	Between-SME Correlation		SME 2 Within-SME Correlation
		SOE	Averaged	
CS-1 2nd scoring	0.822	0.507	0.667	0.851
CS-2 1st scoring	**0.6056**	**0.4936**	**0.6209**	**0.3020**
CS-1 3rd scoring	0.859	0.720	0.920	0.859
CS-2 2nd scoring	**0.812**	**0.708**	**0.859**	**0.855**

Table 29: CS-2 – Comparison of Correlation Significance (p-values)

	SME 1 Within-SME Correlation	Between-SME Correlation		SME 2 Within-SME Correlation
		SOE	Averaged	
CS-1 2nd scoring	< 0.001	0.225	0.001	< 0.001
CS-2 1st scoring	**0.056**	**0.107**	**0.050**	**0.234**
CS-1 3rd scoring	< 0.001	< 0.001	< 0.001	< 0.001
CS-2 2nd scoring	**< 0.001**	**0.001**	**< 0.001**	**< .0.001**

Table 30: CS-2 – Comparison of Significance of Correlation Change (p-values)

	SME 1 Within-SME Correlation	Between-SME Correlation		SME 2 Within-SME Correlation
		SOE	Averaged	
CS-1 2nd to 3rd scoring	0.407	0.156	0.022	0.889
CS-2 1st to 2nd scoring	**0.136**	**0.192**	**0.075**	**0.007**

Table 31: CS-2 Means, Standard Deviations, and Bias Significance

Simple Averages				SOE Averages			
	SME 1	SME 2	p-value		SME 1	SME 2	p-value
Scoring 1	3.503	3.147	***< 0.001***	Scoring 1	3.875	3.500	***0.054***
Scoring 2	3.147	3.405	*0.092*	Scoring 2	3.245	3.405	*0.751*
Simple Standard Deviations				**SOE Standard Deviations**			
	SME 1	SME 2	p-value		SME 1	SME 2	p-value
Scoring 1	0.579	0.824	0.168	Scoring 1	1.183	1.600	*0.376*
Scoring 2	0.211	0.141	0.288	Scoring 2	0.396	0.933	0.830

In addition to the analyses supporting the Assessment Hypotheses, analyses were performed to determine if biases existed between the SMEs' scores. Student t-tests and ANOVA analyses were performed on the SMEs' scores from each scoring to determine if statistically significant differences (biases) existed between the means and variances. This data can be found in Table 31. Only the means for the simply averaged DQCS scores and SOE scores from the first scoring provided strong evidence (bolded in Table 31) to reject the hypothesis that the means were not the same. There was insufficient evidence to reject this hypothesis for the remaining means and standard deviations from the first and second scorings. Overall, there does not seem to be a consistent bias in either of the SMEs' scores.

Analysis of results

The results of the various analyses are presented in Tables 23 and 24, and there are interesting relationships among the between-SME correlations.

<u>First Scoring (Table 25)</u>

- The between-SME correlations are low (below 0.7).

- For the between-SME correlation of SOE scores, there is only weak evidence ($p = 0.107$) to reject the null hypothesis that this Between-SME correlation is due to random chance (Table 26).
- For the between-SME correlation of Average DQCSs, there is strong evidence ($p = 0.050$) to reject the null hypothesis that this Between-SME correlation is due to random chance (Table 26).
- There is insufficient evidence to reject the null hypothesis that the SOE and Averaged DQCS correlation are the same.

Second Scoring (Table 25)

- The between-SME correlations are greater than $\rho = 0.7$.
- For both between-SME correlations, there is strong evidence, $p < 0.001$ (Table 26), to reject the null hypothesis that these Between-SME correlations are due to random chance.
- All the between-SME correlations appear to have increased, but there is only weak evidence, $p = 0.075$ (Table 27), for rejecting the null hypothesis that the first and second scoring correlations are the same for the average DQCS correlation. There is insufficient evidence to reject this hypothesis for the correlations of the averaged SOE scores.

- **Evaluation of the Usefulness of the GDSS**

The evaluation of the value of the GDSSs centered on three aspects of plan quality: (1) time to completion, (2) subjective plan quality, (3) understanding of the

terrain. The first aspect was evaluated objectively, and the remaining two were evaluated subjectively by SMEs. An ANVOA analysis of the data from the second scoring indicated the following:

Hypothesis 1: Time to Completion. As was expected, there was insufficient evidence that decision-makers completed the tasks more quickly when using the GDSS than when not using the GDSS. This expectation was the opposite of the expectation for this hypothesis in CS-1. Military planning problems of this complexity are typically time constrained, and SMEs were confident that planners would continue to plan for the entire time allotted. Therefore, the primary evaluation criteria was the subjective evaluation of decision quality. A repeated measures ANOVA provided insufficient evidence ($p = 0.573$) that there was, on average, any difference between the Time to Completion for the two conditions.

Hypothesis 2: Decision Quality. A repeated measures ANOVA analyzing the SMEs' evaluation scores for the decision-makers' responses for all the sub-characteristics of decision quality suggests that decision-makers using the GDSS produced higher quality outputs than when using functions of the hosts systems alone; but the p-value of 0.080 did not reach the traditional level for statistical significance. All the sub-characteristics of Quality were developed in conjunction with SMEs; but of the 53 measures, 20 are measures of general plan quality and 33 are specifically related to the GDSS being evaluated. A second repeated measures ANOVA on the overall decision quality score calculated using the SMEs' scores on these 35 measures indicated that, with respect to GDSS functions, there was strong statistical evidence ($p = 0.012$) that the

decision-makers' plan quality was superior when using the GDSS. Therefore, there is strong evidence to reject the hypothesis that the quality of decisions was the same.

Hypothesis 3: Terrain Understanding. As expected, the decision-makers' understanding of the impact of terrain on decision-making understanding when using the GDSS was not worse than when using the host system alone. A repeated measures ANOVA result shows insufficient evidence ($p = 0.271$) that, on average, the Terrain Understanding of the decision-makers differed; therefore, the hypothesis that there was no difference cannot be rejected.

A more detailed presentation of these results can be found in (Powell et al., 2008), and the conclusions that can be drawn from these results with respect to the usefulness of the DQEM are discussed in the conclusion section.

Conclusions

The second case study addressed the questions of whether augmenting the MDMP with the GDSS improved the decision-making process and whether the use of the DQEM improved the SMEs' ability to reliably evaluate decision quality. From the results of the evaluation of the GDSS and the results of the assessment of the DQEM scoring presented above, several conclusions can be drawn.

Evaluation of the GDSS

Although only eight of the desired sixteen decision-makers were available, the results indicate that use of the GDSS functions did improve decision-making. Also, achieving statistically significant results suggests that using the DQEM aided the SMEs in evaluating decision quality such that the overall decision quality scores could

discriminate among the quality of decisions. Due to only having eight decision-makers, these conclusions are based on results that were not as strong as those in CS-1. Even given the small sample size, statistically significant results were obtained for the primary hypothesis:

> *Hypothesis 2 - Decision-makers using the GDSS produced better quality plans. Statistical evidence supports this and there is also strong evidence that they produced better plans with respect to the areas directly supported by the GDSS.*

There was also strong support for:

> *Hypothesis 3 - Decision-makers demonstrated no loss of Terrain understanding due to system automation. The decision-makers' Terrain understanding using the GDSS was equal to or better than when their terrain understanding when not using the GDSS.*

In order to estimate the potential significance of the analyses conducted if the desired total of sixteen decision-makers had been available, eight additional sets of data were simulated. The simulation generated ten sets of eight additional data points from the distributions of both the With and Without conditions for Time to Completion, Quality, and Terrain Understanding. With 10 data sets from a total of sixteen decision-makers (eight real and eight simulated), we repeated the ANOVA analyses. For Quality, eight of ten p-values were less than the traditional $p = 0.05$ indicating that if the experiment were continued with eight additional subjects, statistically significant results would likely be achieved for hypothesis 2 (plan quality). Our simulation results suggest that adding eight additional subjects is unlikely to yield strong statistical evidence

supporting Hypothesis 1 (time to completion) or refuting Hypothesis 3 (terrain understanding).

Overall, the significant results in this case study provided evidence that when employing the DQEM, SMEs could subjectively evaluate decision quality with sufficient reliability to generate statistically significant results. That statistically significant results were obtained using the direct evaluation of decision quality as the primary evaluation measure implies that the SMEs' evaluations were reliable enough to discriminate among the quality of decisions. The ability to generate statistically significant results in CS-2, a problem that was more complex and required different decision-making skills than the previous case study, indicates that the DQEM can be adapted and be useful in evaluating the decision quality associated with problems of varying complexity. Specific to CS-2, even with half of the desired number of decision-makers, the SMEs' evaluation scores had low enough variation that statistical analyses indicated that using the GDSS improved decision making. The simulation of the possible strength of the evidence if sixteen decision-makers had been available suggests that the evidence of improvement in decision-making would have been even stronger had more subjects been available. Generalizing from Case Studies One and Two, the use of the DQEM seems to improve the SMEs' ability to evaluate decision quality for problems of varying complexity.

Assessment of the effectiveness of the DQEM

The results obtained from the evaluation of the GDSSs support the assertion that an evaluation of decision quality can be used as the primary criterion in evaluating changes in a decision-making process. In addition to evaluating the GDSS, the

experimental structure and the data the experiment generated also allowed an exploration of the three assessment hypotheses:

1. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the evaluations of decision quality will agree more closely (have higher inter-rater reliability).*
2. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the SMEs will be better able to differentiate among the quality of decisions.*
3. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the ability of decomposition and scoring criteria to capture the characteristics that the SMEs feel are the most important to the evaluation of decision quality will improve.*

Because there are no ground truth scores for either the quality of individual characteristics or the quality of the overall decision in the complex, ill-structured problem used in this case study, the three hypotheses above were investigated through the analysis of the change of the SMEs' within- and between-SME correlations and the ranking of each SME's overall evaluation scores. Between-SME correlations are the correlations between the scores SOE and averaged DQCS scores resulting from the SMEs' independent evaluations of the decision-makers' decisions. The between-SME

correlations are a measure of the level of agreement between the reliability of the SMEs' evaluations. Within-SME correlations, on the other hand, are correlations between each SME's SOE scores and the averages of their DQCSs. The within-SME correlations are a measure of how well the decomposition of decision quality characteristics reflected SMEs' understanding of the most important elements and relationships involved in the decisions. Spearman's rank correlation coefficient, unlike Pearson's correlation coefficient used in the between- and within-SME correlations, is a non-parametric measure of the similarity of two rankings that is used to assess the SMEs' ability to differentiate among the quality of decisions.

Assessment Hypothesis 1

The evaluation of assessment hypothesis 1 centers around the between-SME correlations that are summarized in graphic form in Figure 13, below. In order to support this hypothesis, the between-SME correlations would have to demonstrably increase over the between-SMEs' two successive scorings. Upon visual inspection of Figure 13, the apparent increase in the between-SME SOE and averaged DQCS correlations seem to support the assessment hypothesis 1 and provide some additional insight into the SMEs' ability to subjectively evaluate decision quality. First, both between-SME correlations (SOE scores and simple average of DQCS) seemed to have increased between the scorings. Second, the correlation of average DQCS seems higher than the correlation of SOE scores in all scorings.

.

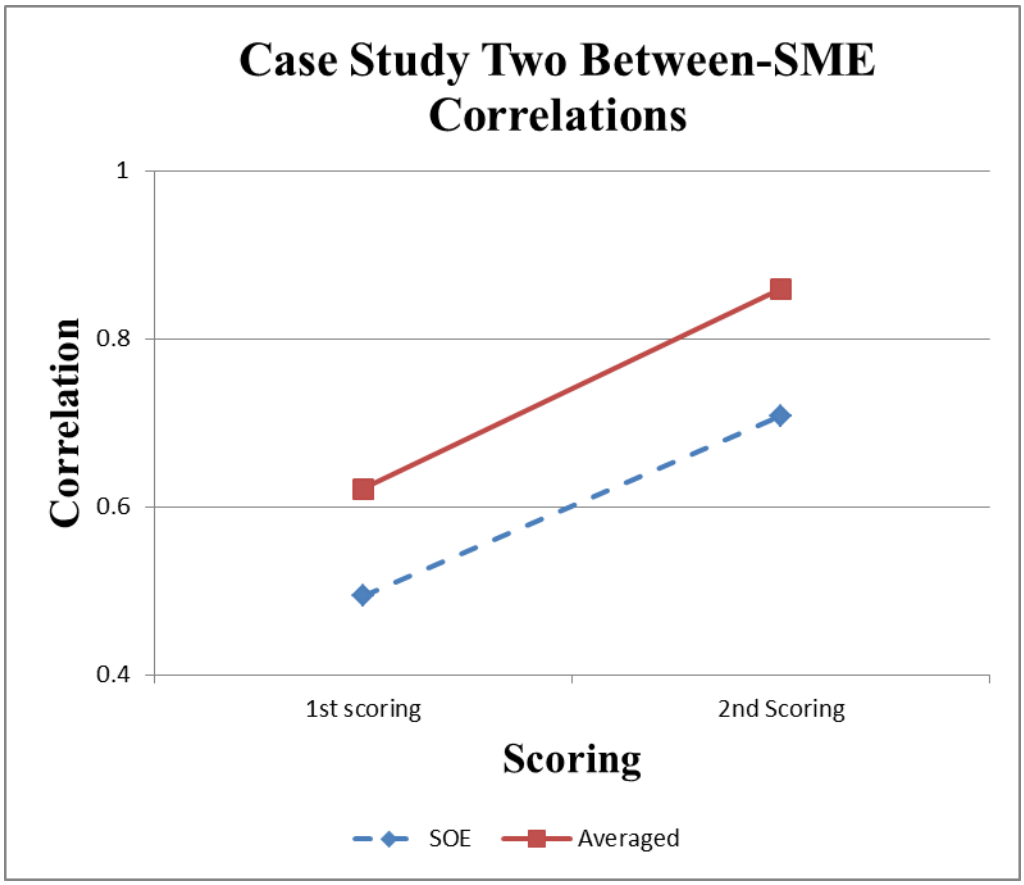

Figure 13: CS-2 – Between-SME Correlations

Based on the level of detail in the decomposition of decision quality, the first scoring in this case study was performed at the same point in the decomposition as the second scoring in CS-1. Like the between-SME correlations in the equivalent scoring in CS-1, the between-SME correlations here are both not high enough ($\rho < 0.700$) to indicate close agreement between the SMEs' evaluations. Also, the significance of the correlation of SOE scores is too low ($p = 0.107$) for this correlation to be a good estimate of the agreement of the SMEs' evaluations. In contrast to the SOE correlation, there is strong evidence ($p = 0.050$) that the correlation of the average DQCS is a reliable

estimate of the agreement of the SMEs' evaluations. Overall though, neither between-SME correlation provides evidence that the SMEs' scores are in close agreement.

In the second scoring, the SMEs had reached consensus on the evaluation and scoring criteria. This consensus resulted in the finest granularity for the decomposition of decision quality and led directly to the construction of the tailored Likert scales. In this scoring both the between-SME correlations had increased. The correlation of averaged DQCSs was high ($\rho = 0.859$) and; like in CS-1, the correlation of SOE scores was somewhat lower with $\rho = 0.720$. Also like in CS-1, both the between-SME correlations were statistically significant ($p <= 0.001$) indicating that the correlations could be considered true estimators of the agreement of the SMEs' evaluations. The high levels of reliability indicated by the correlations reinforced the conclusion from CS-1 that the level of detail in the decomposition combined with the use of tailored Likert scales reduced variation in the SMEs' evaluations below that in the previous scorings which resulted in the SMEs' evaluations becoming very reliable. Because the level of detail in the decomposition was increased (evaluation criteria defined) in the same scoring that the scoring criteria were used to construct the Likert scales, there is no information on the relative effect of these two changes on the SMEs' evaluations. But together, the decomposition and the tailored Likert scales aided the SMEs in the production of very reliable evaluations.

The previous discussion concentrated on the apparent increase in the between-SME correlations, and an analysis of the changes in these correlations between the two scorings did not confirm that the between-SME correlations did increase significantly.

Likely due to the smaller sample size, there is only weak evidence ($p = 0.075$) supporting an increase in the average of DQCS and insufficient evidence that there was a difference in the SOE score correlation. Even so, actual between-SME correlations in the first and second correlation and the change in these correlations are very similar to those in the first case study which suggests that the development of evaluation and scoring criteria had similar effects on the SMEs' evaluations in both case studies.

Assessment Hypothesis 2

Although the analysis of the between-SME correlations provided strong evidence that the DQEM aids SMEs in reliably evaluating decision quality, the between-SME correlations do not provide support for Assessment Hypothesis 2. As in CS-1, Spearman's Rank Correlation Coefficient (SRCC) was used to compare the rankings of the SMEs' evaluation scores to determine the level of agreement of the SMEs' evaluations of each decision-makers' decisions. Figure 14 shows the SRCC for each of the averaged DQCS and SOE scores.

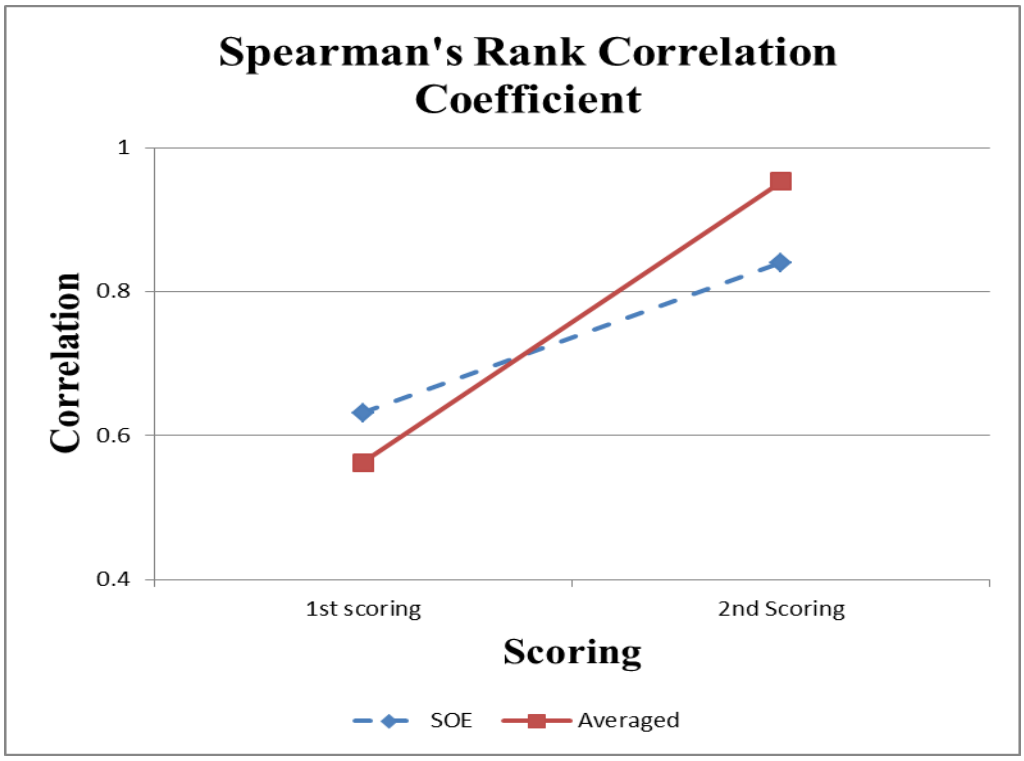

Figure 14: CS-2 − Spearman's Rank Correlation Coefficients

An inspection of Figure 14 indicates that the SRCC for both correlations increased between the first and second scorings. Since by the second scoring the SRCC for the average DQCS is higher than that for the SOE scores, the SMEs' average DQCS should be better able to discriminate among the quality of decisions. The comparison of the SMEs' ranked average DQCS and their averaged DQCS in Figure 15 and 17 respectively demonstrates the ability of the SMEs to discriminate among the quality of decisions using the DQEM.

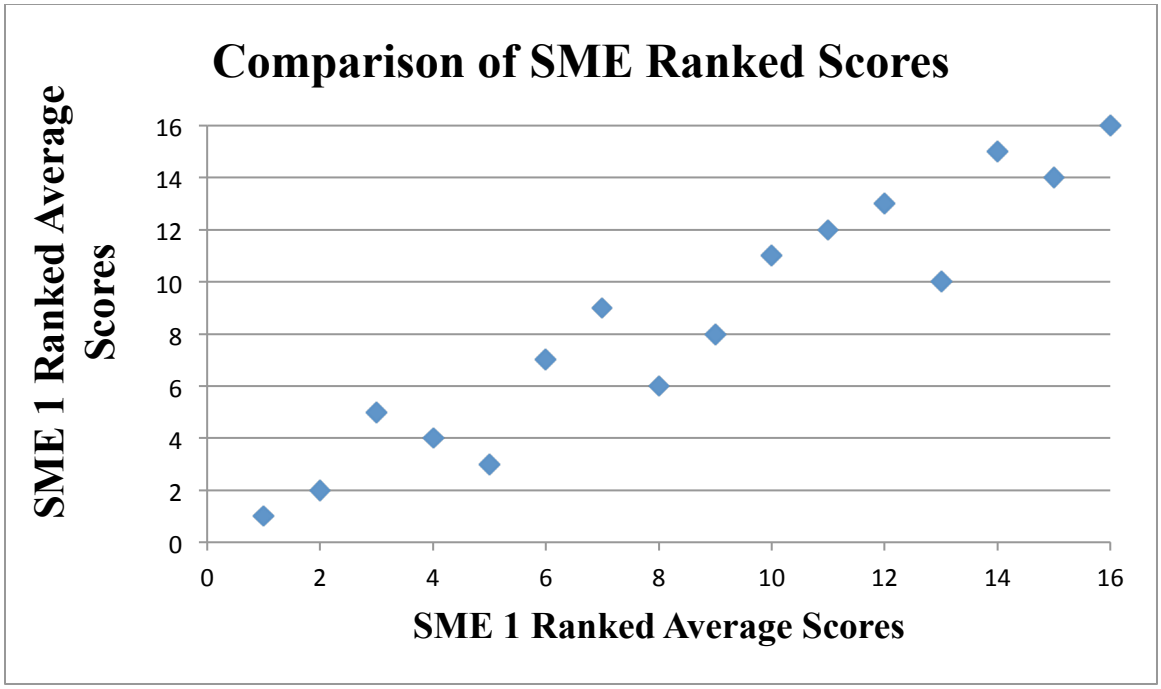

Figure 15: CS-2 − Comparison of SMEs' Ranked Scores

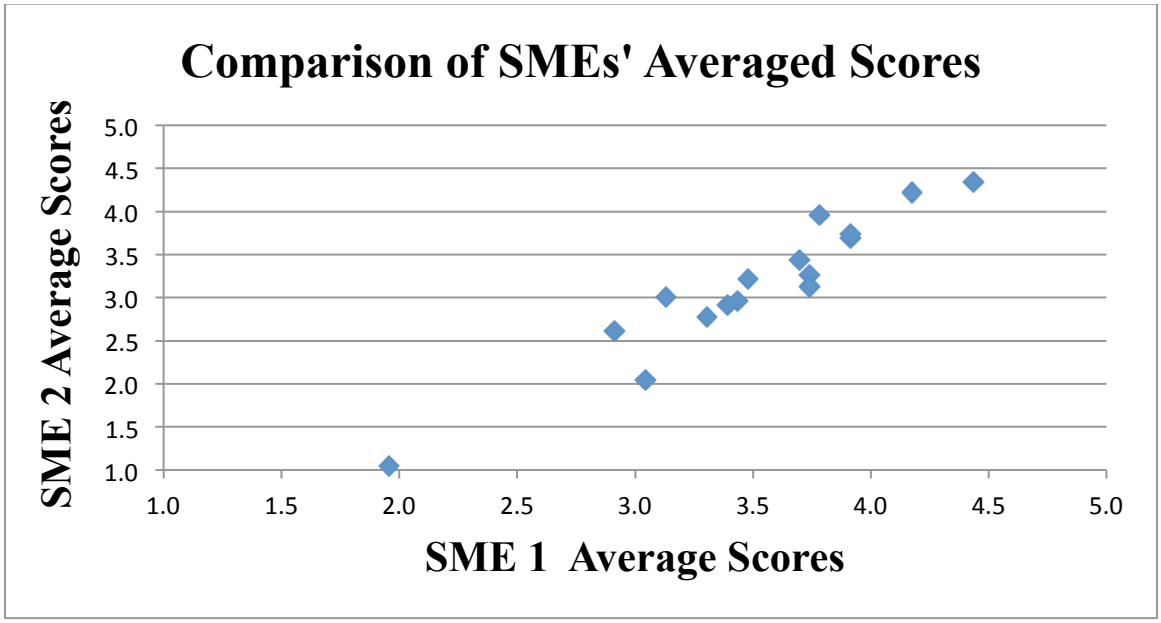

Figure 16: CS-2 − Comparison of SMEs' Average Scores

Assessment Hypothesis 3

Like the evaluation of Assessment Hypothesis 3 in CS-1, the results in CS-2 do not support that decomposition beyond the decision quality sub-characteristic level provides an improvement in the ability of the decomposition of decision quality to better capture the characteristics that the SMEs feel are important. The data collected from scorings provide insufficient evidence that the within-SME correlations of the SMEs' SOE scores changed throughout the scorings.

Summary

Overall the results gathered from the two scorings reinforce the support for the research hypothesis documented in Case Study One. While the results from this case study did not provide support as strong as that from Case Study One, the statistical analyses did provide significant support for Assessment Hypotheses 1 and 2.

The results of CS-2 supported the conclusions of CS-1 that the use of the DQEM seems to improve the agreement of the SMEs' subjective evaluations of decision quality. In successive scorings in which the level of detail included in decomposition of decision quality was increased, the between-SME correlations increased. The analysis of the between-SME correlations and the associated p-values supported and extended the conclusions resulting from the analysis of the within-SME correlations. First, the higher between-SME correlations of average DQCS and the associated significant p-values supported the conclusion that these scores were better estimators of decision quality than the SMEs' SOE scores. Second, the consistently increasing agreement between the

SMEs' scores though successive scorings supported the conclusion that the increased detail in the decomposition of decision quality seemed to aid the SMEs in assessing decision quality. Third, the analysis of the SMEs' rankings of decision-makers' overall decision quality scores indicated that by the second scoring the average DQCS scores were able to discriminate among the quality of decisions. Supporting this conclusion, the results of the evaluation of the GDSS demonstrate that SMEs' average DQCS were able to sufficiently differentiate among the quality of decisions to be able to generate statistically significant results.

In summary the data collected in Case Study Two, like the result of CS-1, supported Assessment Hypotheses 1 and 2 but did not support assessment hypothesis 3:

- An analysis of the between-SME correlations generated from the SMEs' evaluations supported Assessment Hypothesis 1 in that the analysis indicated that the DQEM, including the decomposition of decision quality, a procedure for developing tailored Likert scales, and the aggregation of decision quality characteristics using a simple average of decision quality characteristics, resulted in reliable evaluations of decision quality by the SMEs.
- In support of assessment Hypothesis 2, an analysis of the ranking of the SMEs' SOE scores and averaged DQCSs indicated that as the level of detail in the decomposition increased, the SMEs' evaluations were better able to differentiate among the quality of decisions.
- Even though the analysis of the within-SME correlation was not able to support assessment Hypothesis 3, the high correlations in the second scoring

indicated that the decomposition of decision quality characteristics was able to capture the characteristics that the SMEs thought were important.

Finally, as evidence that the DQEM can be used to successfully evaluate decision quality in complex, ill-structured problems, this case study evaluated the impact of augmenting a decision-making process with a GDSS. The results of this evaluation were that SMEs using the DQEM incorporated into the experimental design were able to distinguish sufficiently among the differences in the quality of decision in order to generate statistically significant results indicating that the augmenting of the MDMP with a GDSS improved the quality of the decision-makers decisions.

Overall, the support for assessment Hypotheses 1 and 2 support the overall research hypothesis. The Decision Quality Evaluation Method (DQEM) used a structured, subjective decomposition of decision quality characteristics to capture the SMEs' understanding of the elements and relationships of a decision, develop a unique scoring procedure, and aggregate the decision quality evaluation scores for individual sub-characteristics into overall scores of decision quality. SMEs were able to use the decomposition and scoring procedure to reliably evaluate and score the quality of decision quality sub-characteristics. The SMEs' evaluations were sufficiently reliable to allow direct evaluation of decision quality to be used successfully to evaluate a modification to a decision-making process.

Overall, the assessment of the use of the DQEM in CS-2 supported the conclusions noted in CS-1. This case study demonstrated that the relative quality of decisions in complex, ill-structured problems could be determined in evaluations that

incorporated the DQEM. This case study also demonstrated the flexibility and adaptability of the DQEM in assessing decision quality in ill-structured problems of varying complexity and context. CS-2, by generating data from two stages of the decomposition of decision quality, provided additional evidence that using the DQEM increased the correlation between the SMEs' DQCS. These increases in the between-SME correlations indicate that the SMEs' evaluations became more reliable and were able to be used to discriminate among the quality of decisions.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

Conclusions

The research reported here spanned several disciplines. It drew from the literature of complex, ill-structured problems, decision quality, multi-criteria decision analysis techniques, decision support systems, and theory on evaluating decision quality. This broad foundation was necessary to support the design and assessment of a general structured, subjective method to directly evaluate the quality of decision-making in complex, ill-structured problems. Unlike current methods, the Decision Quality Evaluation Method (DQEM) evaluates decision quality directly instead of using an outcome or processed based approach. The DQEM's direct evaluation of decision quality is well suited to evaluate decision quality in complex, ill-structured problems including problems for which the context in which the decisions will be implemented is uncertain, problems which require the evaluation of several decisions of which only one can be implemented, and problems for which few if any outcomes are or will ever be available for use in evaluating potential decisions.

The DQEM formalizes and expands on established MCDA techniques that characterize the essential elements of decision quality so that, in aggregate, the overall quality of decisions can be evaluated. The DQEM uses the decomposition of decision quality to generate a detailed hierarchical structure of decision quality characteristics,

sub-characteristics, measures, and criteria that capture the qualities that characterize "good" decisions. The DQEM uses the evaluation and scoring criteria developed as part of the decomposition of decision quality characteristics to tailor Likert scales for the evaluation of each sub-characteristic. The tailored Likert scales are designed to reduce the variation due to ambiguity inherent in typical Likert scale categorical labels. Finally, the DQEM provides for the aggregation of evaluation scores from each of the measures into a single overall score of decision quality. The improvement in the SMEs' ability to consistently assess decision quality when using the DQEM's decomposition, scoring, and aggregation is evidenced by the results of Case Studies One and Two.

The overall goal of this research is summarized in the research hypothesis:

> *The direct evaluation of the quality of decisions made to address complex, ill-structured problems can be improved through the use of a structured subjective decomposition of decision quality characteristics.*

The effectiveness of the use of DQEM, a structured subjective decomposition method, to aid SMEs in the direct evaluation of decision quality was demonstrated in two ways in the case studies: directly by assessing the reliability of the SMEs' evaluations, and indirectly through the generation of statistically significant results in two case studies. The direct assessment evaluated the SMEs' ability to subjectively judge overall decision quality based on the assessment of decision quality characteristics and their ability to differentiate among the quality of decisions. In addition to the direct assessment, the usefulness of the DQEM was demonstrated through the successful evaluation of the impact of GDSSs on the Military Decision Making Process (MDMP) in

both case studies. The case studies additionally demonstrated that evaluations constructed using the DQEM could achieve statistically significant results using a small number of decision-makers.

Direct Assessment of the Effectiveness of the DQEM

The direct assessment of the DQEM utilized three measures of the agreement of the SMEs' evaluations. The first measure, the between-SME correlation coefficients, are Pearson's correlation coefficients calculated between the SMEs' subjective overall evaluation (SOE) scores and between the SMEs' simple or weighted average of their decision quality characteristic (DQC) scores. The DQC scores used specific procedures (simple and weighted averages) for aggregating all the sub-characteristics evaluation scores. The SOE scores, on the other hand, were a single score elicited from the evaluators as a subjective "gut check" of the overall decision quality. The SOE scores were used to capture the SMEs' evaluations of decision quality based on the characteristics they deemed important to the decision; and since they were totally subjective, the characteristics the SMEs' thought important and the subjective weights applied to the characteristics included in their SOE scores were unknown and probably not necessarily consistent. The between-SME correlations were a measure of the level or agreement between the reliability of the SMEs' scores.

The second measure, the within-SME correlation coefficients, were Pearson's correlation coefficients calculated between a single SMEs' SOE scores and their simple or weighted average of their DQC scores. The between-SME correlations were used as an indication of the degree to which the decomposition of decision quality characteristic

(represented by the averaged DQC scores) reflect the characteristics the SMEs felt were important to the decisions (represented by the SOE scores).

The third measure was Spearman's rank correlation coefficient (SRCC). The Pearson's correlation coefficient used in the within- and between-SME correlations can be used as an indication of the reliability of the SMEs' evaluations as a whole*;* but even though Pearson's correlation coefficient uses paired data, the between-SME correlations are an indication of the overall reliability among the set of SMEs' evaluations and not an indication of the reliability of any pair of evaluations or of any given evaluation. Therefore, Pearson's correlation cannot be used as an indication of the SMEs' ability to discriminate among specific decision qualities. But, if the between-SME correlations indicate that the SMEs' inter-rater reliability improves; it follows that the SMEs' evaluations should be better able to discriminate among the quality of decisions. If the SMEs' evaluations are better able to discriminate among decision qualities, then a ranking of their overall decision quality scores should be more similar. SRCC captures the level of similarity in the rankings of the SMEs and therefore can be used as a measure of agreement of their evaluations of each decision and a measure of the SMEs' ability to discriminate among the quality of specific decisions.

The evaluation of the research hypothesis was supported by the measures discussed above through the three assessment hypotheses below:

1. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding*

of the problem, the evaluations of decision quality will agree more closely (have higher inter-rater reliability).

2. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the SMEs will be better able to differentiate among the quality of decisions.*
3. *As the decomposition of decision quality and the scoring criteria become more detailed and better reflect and further clarify the SMEs' understanding of the problem, the ability of decomposition and scoring criteria to capture the characteristics that the SMEs feel are the most important to the evaluation of decision quality will improve.*

The analysis of data collected in Case Study One provided strong support for Assessment Hypotheses 1 and 2 but did not support assessment hypothesis 3. The analysis of the data from Case Study Two supported the results from Case Study One as follows:

- There is strong statistical evidence that all the between-SME correlations increased as the detail in the decomposition increased. Further, in the final scorings, there was strong statistical evidence from all three between-SME correlations that the SMEs' evaluations had become highly reliable.
- There was strong statistical evidence that the averaged DQC scores were more reliable evaluation of decision quality than the SOE scores.
- There was strong statistical evidence from the analysis of the rankings of the SMEs' evaluation scores that the SOE scores and that the analysis of

the between-SME correlations generated from the SMEs' evaluations supported Assessment Hypothesis 1. The analysis indicated that the DQEM, including the decomposition of decision quality, a procedure for developing tailored Likert scales, and the aggregation of decision quality characteristics using a simple average of decision quality characteristics resulted in reliable evaluations of decision quality by the SMEs.

- Also, in support of Assessment Hypothesis 2, an analysis of the ranking of the SMEs' SOE scores and averaged DQC scores indicated that as the level of detail in the decomposition increased the SMEs' evaluations were significantly better able to differentiate among the quality of decisions.
- Even though the analysis of the within-SME correlation was not able to support Assessment Hypothesis 3, the consistently high correlations indicated that the decomposition of decision quality characteristics was consistently able to capture the characteristics that the SMEs thought were important.
- As part of the analysis of Assessment Hypothesis 1, an exploration of the effect of using weighted and simple averages to calculate overall decision quality scores was conducted. In the first case study, there was insufficient evidence that using a weighted average to aggregate the scores of individual sub-characteristic evaluations improved the reliability of the overall scores.

Finally, as evidence that the DQEM can be used to successfully evaluate decision quality in complex, ill-structured problems, the two case studies evaluated the impact of augmenting a decision-making process with a GDSS. The results of this evaluation were that SMEs using the DQEM incorporated into the experimental design were able to sufficiently distinguish among the differences in the quality of decisions in order to generate statistically significant results indicating that the augmenting of the MDMP with a GDSS improved the quality of the decision-makers' decisions.

Overall the support for Assessment Hypotheses 1 and 2, support the overall research hypothesis. The Decision Quality Evaluation Method (DQEM) used a structured, subjective decomposition of decision quality characteristics to capture the SMEs' understanding of the elements and relationships of a decision, develop a unique scoring procedure, and aggregate the decision quality evaluation scores for individual sub-characteristics into overall scores of decision quality. SMEs were able to use the decomposition and scoring procedure to reliably evaluate and score the quality of decision quality sub-characteristics. The SMEs' evaluations were sufficiently reliable that the direct evaluation of decision quality was successfully used to evaluate a modification to a decision-making process.

Case study results

As evidence that the DQEM can be used to successfully evaluate decision quality in complex, ill-structured problems, the two case studies evaluated the impact of augmenting a decision-making process with a GDSS. The results of this evaluation were that SMEs' using the DQEM incorporated into the experimental design we able to

sufficiently distinguish among the differences in the quality of decision in order to generate statistically significant results indicating that the augmenting of the MDMP with a GDSS improved the quality of the decision-makers' decisions. Therefore the results of evaluation of the GDSS also support the research hypothesis.

Factors Contributing to DQEM Effectiveness

The process of developing the DQEM and incorporating it in the two case studies revealed several factors that contributed to the effectiveness of the SMEs' evaluations of decision quality. These factors include using knowledgeable SMEs, using a consensus of the SMEs' opinions in the decomposition and development of scoring criteria, developing a highly detailed decomposition of decision quality characteristics, and the use of tailored Likert scales to reduce variation in evaluation scores.

Knowledgeable SMEs with a comprehensive understanding of the problem and decisions are critical to developing a decomposition of decision quality characteristics that accurately captures the characteristics of "good" decisions. The SMEs' understanding of the elements and relationships that make a problem complex and ill-structured is the source of the knowledge that is captured in the decomposition of decision quality characteristics, the development of evaluation, and strong criteria. Though the DQEM's MCDA-based decomposition procedures were designed to allow SMEs to explore all facets of the decision, the SMEs must be highly knowledgeable so that the decomposition can capture sufficiently detailed expert opinion such that the evaluators can reliably evaluate decision quality. SMEs who were not true experts might not be able to contribute to decomposing decision quality to the level of detail achieved

in the case studies. It is likely that capturing too little of the SMEs' understanding in the decomposition or scoring criteria would require additional judgments on the part of the evaluators that could be a source of variation in their evaluation scores.

Similarly, reaching consensus on the decomposition and the scoring criteria will tend to reduce the impact bias on the development of the decomposition and scoring criteria. One typical use of decompositions in MCDA is to resolve differing biases. This function is inherent in the subjective decomposition in the DQEM and is an aid to resolving the impact of differing SME biases on the decomposition and the evaluation of decision quality. Although using the decomposition will typically detect differing biases, it will probably not detect biases common to several SMEs. So as not to skew the evaluations of decision quality, efforts should be made to ensure that the SMEs are independent of the decision-makers so that overt bias does not contaminate the decomposition of decision quality characteristics and thus affect the evaluations of decision quality.

The DQEM, as implemented in the case studies, decomposes decision quality into a finely granulated hierarchy. As discussed in the results of both case studies, the reliability of the SMEs' evaluations and their ability to discriminate among the quality of decisions improves as the level of detail in the decomposition increases. The level of detail in the decomposition in the last scoring resulted in sufficiently reliable evaluations that showed that the SMEs were able to differentiate among decision qualities. The effect of insufficient detail in the decomposition is evidenced by the lack of reliability and inability to differentiate among decision qualities in all but the last scoring. One of

the strengths of the DQEM is that it extends the decomposition to the development of evaluation and scoring criteria. Developing evaluation and scoring criteria as detailed in Appendix 1-1.1 and described in the case studies should result in decompositions of decision quality characteristics of sufficient detail to produce the reliable evaluations of decision quality demonstrated in the case studies.

The last factor that contributed significantly to the effectiveness of the DQEM was the use of the tailored Likert scales. The tailored Likert scales were designed to reduce the variation in the translation of the SMEs' subjective evaluations to numeric values. The most significant improvement in the reliability of the SMEs' evaluations occurred in the last scoring of both case studies in which the tailored Likert scales were introduced. The development of the scoring criteria which was used to construct the Likert scales and the use of the tailored Likert scales in the SMEs' evaluations had a dramatic impact on the reliability of their evaluations.

In order to achieve the level of reliability the SMEs' evaluations attained in the case studies, all the factors mentioned above need to be considered when using the DQEM to evaluate decision quality. The detailed description of the incorporation of the DQEM into Case Study One, documented in Appendix 4-1.1, can be used as a guide to help ensure that all these factors are considered when implementing the DQEM.

Contributions

This research made the following contributions to the field of decision support by developing and applying a structured evaluation method that:

- **Directly evaluated decision quality** using a Decision Quality Evaluation Method (DQEM) that focuses on characterizing the essential elements of decision quality such that, in aggregate, the relative overall quality of decisions can be evaluated without the use of outcomes and proxies for decision quality. The method uses a procedure that decomposes decision quality hierarchically into decision quality characteristics, sub-characteristics, measures, and criteria. Tailored Likert scales were used to capture the relationship among decision quality characteristics, to translate subjective evaluation into numeric scores, and to reduce the variance in scoring due to scale ambiguity.
- **Applied multi-criteria decision analysis (MCDA)** techniques to construct a detailed decomposition of subjective decisions' quality characteristics designed to be used by independent SMEs to evaluate the quality of decisions
- **Used the consensus opinions of SMEs** to **capture the SMEs' subjective understanding** of the complex relationships among the factors affecting the quality of decisions without the need to resort to a time-consuming, tedious process that elicits values quantifying these subjective factors. The use of an SME consensus was also used to identify and resolve differing SME biases.
- **Defined tailored Likert scales** that further captured the SMEs' understanding of the relative importance of specific criteria when evaluating the quality of decisions. These Likert scales allowed SMEs to independently translate the

achievement of criteria to specific scores, thereby reducing the variation in their scores due to individual interpretation of the Likert scales.

- **Developed a method (the DQEM)** that was designed such that third party SMEs could use the decomposition and tailored Likert scales to independently guide and score evaluations of similar decisions.

This research also **demonstrated the effectiveness of the DQEM** in two case studies evaluating the impact of differing decision-making processes on complex, real world decision-making. The effectiveness of the DQEM was demonstrated in two ways: first, by evaluating the reliability of the SMEs' evaluations of decision-makers' decisions; and second, by evaluating the similarity of ranking of the SMEs' evaluations.

Further, use of the DQEM in these two case studies **demonstrated the flexibility and adaptability of the DQEM.** The case studies demonstrated that the DQEM could be successfully applied to the evaluation of decision-making in ill-structured problems of different complexities. Specifically, Case Studies One and Two evaluated decision-making processes that were based on the Military Decision Making Process (MDMP) but were designed to assist decision-makers with different skill sets to make decisions for problems of significantly different complexity that required more extensive and more complex decisions and that yielded more sophisticated outputs.

Finally, the detailed description of the implementation of the DQEM in Case Study 1 **provided a guide for the construction of evaluations using decision quality as the primary measure**.

The primary contribution of this research was the assessment of the effectiveness of the DQEM. This method is unique in the area of decision quality evaluation; and this research demonstrated that a structured subjective method, such as the DQEM, could be used to evaluate decision quality in complex, ill-structured problems. Further, the analysis of the results from two case studies identified elements in the DQEM that seemed to positively impact the ability of SMEs to reliably evaluate the quality of decisions and differentiate among the quality of several decision-makers' decisions.

Summary

This research assessed a Decision Quality Evaluation Method (DQEM), which was used to directly and effectively assess decision quality in complex, ills-structured problems. The DQEM aids SMEs in the reliable evaluation of the relative quality of decisions. The case studies demonstrated that SMEs using the DQEM could reliably evaluate decision quality in an evaluation of the impact of altering a decision-making process. Nothing in this research suggests that the value of the DQEM is limited to this type of evaluation. In fact, the success of the evaluations in the case studies suggests that the DQEM is flexible and adaptable enough to be applied to a wide range of problems. The DQEM was successfully applied to the evaluation of decision-making in ill-structured problems with various the levels of complexity, requiring different decision-making skills, and in different decision-making contexts; and this success suggests that the DQEM would be appropriate for evaluating decision quality in areas other than those associated with DDSs and military planning.

Future Work

This research demonstrated that the direct evaluation of decision quality in complex, ill-structured problems is not only possible but practical. During the conduct of this research, assumptions (usually conservative) were made that probably affected the ability of the DQEM to assess decision quality.

The DQEM is useful for comparing the relative quality of decisions that have the same context. The DQEM assumes that the contexts are very similar, and the use of an evaluation design that used within-subject independent variables exploited this assumption to reduce the number of decision-makers. The assumption of very similar decision contexts was made because there are no standards of decision quality; there is no means of comparing the quality of decisions made in disparate contexts. Research into standards for decision quality could result in the means to generally compare the quality of decisions. One possible research path would be the generation of a general ontology of decision quality. There exist ontologies designed to assist in decision-making with respect to specific problems, but an ontology that addresses the quality of decisions would be good step toward defining standards by which decision quality can be compared.

In CS-2, there was concern that the SMEs reaching consensus instead of using independent scores could introduce a bias into the scores. The correlations between the SMEs' individual scores and the consensus scores were compared in order to determine if one set of scores had undue influence on the consensus. This method seems to be useful

intuitively, but research into standards for decision quality could lead to a method to estimate the bias introduced by consensus scoring.

Another possible avenue of research is one into the level of homogeneity of decision-makers and SMEs. The need for homogeneity in the decision-makers was understood and the variation due to any lack of homogeneity was controlled by (1) specifying the characteristics of the decision-makers that would be acceptable as decision-makers and (2) by subjectively analyzing biographical data and attempting to equalize the variation in experience, training, and knowledge thought the assignment of a cross section of the decision-makers to each group. Research into a method for quantifying the level of homogeneity in the decision-makers and associating the level of homogeneity to the potential variation and bias in data would help in estimating and reducing the number of decision-makers that could be expected to result in statistically significant results.

Related to research into the homogeneity of the SMEs, it would be interesting to investigate how similar the results of the case studies using different sets of SMEs. In both case studies the SMEs who did the initial decomposition were independent of the SMEs who developed the evaluation and scoring criteria. But, the SMEs who developed the evaluation and scoring criteria were the SMEs who evaluated the outputs of the decision-making process. The SME in the case studies evaluated either 16 or 20 complex plans, and based on discussions with the SMEs, they were not able to recall the specifics of any given plan at the end of an evaluation session. Because of this, the SMEs were assumed to be independent for analysis purposes. If there were SMEs available,

conducting the scoring again using an independent set of SMEs to develop a second decomposition and independent sets of SMEs to evaluate the outputs of the original decision-making process under both decompositions would yield an interesting comparison on the similarity of the decompositions and the evaluations.

Research into a method to quantify the homogeneity of decision-makers could identify the impact of the homogeneity of the SMEs on their ability to consistently evaluate decision quality. In the current research, although the need for homogeneity in the SMEs was not expressly considered, the characteristics desired in the SMEs were identified as being similar to those of the decision-makers. The SMEs sought were personnel with more experience, knowledge, and training than the decision-makers. Since getting the services of highly qualified SMEs is often difficult, the ability to estimate the impact of the homogeneity of the SMEs on the agreement of the SMEs' scores could aid in estimating the likelihood of obtaining statistically significant results. Together with research into quantifying the homogeneity of SMEs, research in this area could improve the effectiveness of the DQEM

APPENDIX 2-1: CALCULATION OF STATISTICAL POWER

From (Kraemar & Theimann 1987), pp.45-49

Using calculations for a Single-Sample Pre-Post Design

Calculating "Glass's effect size", δ, for a mean difference of 0.5 and standard deviation of 0.5:

$$\delta = (\mu_2 - \mu_1)/\sigma$$

$$\delta = (.5)/.5$$

$$\delta = 1$$

Assuming a minimum correlation for the paired samples of 0.7, calculating critical effect size, Δ:

$$\delta = (\mu_2 - \mu_1)/\sigma[2(1-\rho)]^{1/2}$$

$$\delta = (.5)/.5[2(1-0.7)]^{1/2}$$

$$\delta = 1.29$$

and

$$\Delta = \delta/(\delta^2 + 1)^{1/2}$$

$$\Delta = 1.29/(1.29^2 + 1)^{1/2}$$

$$\Delta = 0.791$$

Using the calculated Δ= 0.791 to enter the master table for a p = 0.05, two-tailed test, indicates that 16 decision-makers would yield a statistical power of between 0.95 and 0.99 (Kraemar & Theimann, 1987).

Table 32: Excerpt from Master Table for p = 0.5, Two-Tailed Test

5% level, Two-Tailed Test					
Δ	99	95	90	80	70
0.70	26	19	15	12	10
0.75	21	15	13	10	*
0.80	17	12	10	*	*
0.85	13	10	*	*	*
0.90	10	*	*	*	*

APPENDIX 4-1.1: DETAILS OF THE IMPLEMENTATION OF THE DQEM IN CASE STUDY ONE

This case study describes the use of the DQEM and the associated evaluation structure described in Chapter Three in an evaluation of the GDSS developed for a project at the U.S. Army Engineer Research and Development Center (ERDC). The goal of the evaluation was to assess the impact of the use of the GDSS on decision making. The goal of this case study was to demonstrate the use of the DQEM in the evaluation of the GDSS and decision-making. The specific GDSS evaluated in the assessment were the Battlespace Terrain Reasoning and Awareness – Battle Command (BTRA-BC) tools (U.S. Army, 2003). The BTRA-BC program, which builds upon a commercial GIS tool (ARCINFO), has resulted in mature components that have been integrated into the Army's Digital Topographic Support System (DTSS), a system that provides topographic engineering support to topographic technicians as they assist military planners (Herrmann, 2002). DTSS provides geospatial data generation, collection, management, information processing, and services. The GDSS expands the capabilities of DTSS through the creation of information and knowledge products that enhance soldiers' understanding of terrain and weather as it impacts their functional responsibilities. The BTRA-BC capabilities assessed in this study include the identification of obstacles, the production of a Modified Combined Obstacles Overlay (MCOO), the generation of

Mobility Corridors (MCs), the combining of MCs to form routes, and the identification of Choke Points (CPs). While this assessment provided essential information to evaluate the contribution of the BTRA-BC tools in particular and the GDSS in general to the military decision making process, it also provided data on the usefulness of using the DQEM to evaluate decision quality.

Scope of Case Study One (CS-1)

The primary goal of this case study was to demonstrate the usefulness of the DQEM in evaluating decision quality in complex and ill-structured problems. A secondary goal was to describe the implementation of the DQEM in such a way that this chapter could be used as a guide for others to use in designing evaluations of decision quality. A third goal was to generate the lessons learned from this implementation of the DQEM that could be used to refine the DQEM for use in more complex problems. CS-2 (Chapter Five) incorporates these lessons learned.

In order to test the DQEM, a suitably complex and ill-structured problem had to be devised. A Military planning problem seemed the ideal choice to test the GDSS. There were several distinct advantages to using a military planning problem. First, a military planning problem is a complex and ill-structured planning problem. All but the simplest military planning problems exhibit all the characteristics of complexity and ill-structure noted in Chapter Two. Second, military planning problems vary widely in their complexity and level of ill-structure. This wide variation allows specific planning problems to be tailored in complexity and level of ill-structure such that they support the requirements of individual experiments. Although there is no objective standard to which

the level of complexity and ill-structure could be compared to get a specific measure of either characteristic, SMEs experienced with military planning could determine the relative level of complexity and ill-structure and determine whether specific problems were appropriate to the evaluation. These SMEs could design military planning problems of sufficient complexity and ill-structure to stress the GDSS ability to support decision-making. Third, the military planning process is relatively standardized and well documented. Because of this, the likely use of the GDSS in the planning process could be predicted, and the scenarios tailored to emphasize GDSS use. Fourth, the characteristics of the problem are easily identifiable. Once the GDSS and the scenarios were determined, the hierarchical nature of the military decision-making helped determine the qualifications and experience level of the intended users and thus those of the representative SMEs.

Battlespace Terrain Reasoning and Awareness – Battle Command (BTRA-BC)

The GDSS evaluated in CS-1 were those included in the Battlespace Terrain Reasoning and Analysis – Battle Command (BTRA-BC) suite of geospatial tools. BTRA-BC contains various tools that can be applied to military planning problems of varying levels of complexity and ill-structure. Because of this, the complexity and ill-structure of the experimental problems could be tailored to evaluate the effects of the BTRA-BC GDSS on decision quality in an appropriate context.

The individual components of the BTRA-BC GDSS are referred to as tactical Spatial Objects (TSOs). TSOs are computationally lightweight software engines that transform geospatial data into geospatial information unique to a specific military

planning analysis. BTRA-BC TSOs are the GDSS whose capabilities include analysis engines, data manipulation routines, and other software products in support of terrain reasoning (USACE, 2003). BTRA-BC generates information addressing (1) Observation, Cover and Concealment, Obstacles and Mobility, Key terrain and Avenues of approach (OCOKA); (2) integrated products defining operational Positions of Advantage; (3) high-fidelity weather/terrain effects on mobility and signature physics; (4) advanced mobility analysis; (5) digital ground and air maneuver potential; and (6) tactical structures relating information produced by the other components (USACE, 2010). BTRA-BC's focus is the development of software analytics designed to create information and knowledge products that capture integrated terrain and weather effects and develop predictive decision tools to exploit those products. The ultimate objective is to empower commanders, soldiers, and systems with information that allows them to understand and incorporate the impacts of terrain and weather on their functional responsibilities and processes (USACE, 2009).

BTRA-BC TSOs are categorized as Tier 1, Tier 2, and Tier 3. Tier 1 TSOs are geospatial information products that do not depend on the mission. They are derived only from the terrain and characteristics representative of general ground forces and are independent of specific forces involved or missions. Tier 1 TSOs are generally more straightforward to develop than Tier 2 and 3 TSOs, assist in the intelligence preparation of the battlespace, and provide background information for the more complex Tier 2 and 3 TSOs. Tier 2 TSOs are products designed to assist a specific force in the performance of well-defined military tasks consistent with a mission or objective (Visone, 2008). Tier

2 TSOs build on the information generated by Tier 1 TSOs to generate products that aid military planners in developing Courses of Action (COA) in support of specific missions with friendly and opposing forces. Tier 3 TSOs are products suitable for a specific force to performed well-defined military tasks consistent with a mission or objective and refined by the current situation (Visone, 2008). Tier 3 products are designed to build on Tier 2 information and incorporate data on the current situation in near real time.

Table 33: CS-1 – Case Study Summary

Case Study One	
GDSS	BTRA-BC Tier 1 TSOs
Decision-makers	Terrain Analysts
Number	18
Host System	DTSS
Mission	Tactical Movement
Tasks	Terrain Analysis
	Recommend Avenues of Approach
Output	Digital Plan
	Terrain Understanding Questionnaire
	Comparison Questionnaire
	Post Trial Discussions

This case study evaluated the BTRA-BC GDSS that consisted of Tier 1 TSOs. Because Tier 1 TSOs only require terrain data as inputs and not data on forces and objectives, their outputs are independent of the mission being planned. Because the portion of the military decision-making process known as the Intelligence Preparation of the Battlefield (IPB) is only loosely linked with forces and mission, the tasks in this case study were those associated with IPB tasks. The planning-specific problem was to

develop routes as part of a recommended Course of Action (COA). The supporting tasks dealt primarily with analyzing the terrain and required little experience with combat operations on the part of the decision-makers. Table 33, above, summarizes the design factors incorporated into CS-1. The reasoning behind the selection of the host system, unit size, mission, tasks, and output is discussed in the following sections.

Generation of decision quality characteristics and measures

Much of the general method discussed here was used as the running example in Chapter Three, but significantly more detail was incorporated into the discussion of this case study than was presented in the examples of Chapter Three.

Determine the GDSS to be evaluated

The GDSS evaluated in Case Study I was Tier 1 TSOs related to determining routes for military units of various sizes from a start point (assembly area) to an endpoint (objective). These TSOs use terrain data from digital maps to generate overlays that display graphic representations of the specific geospatial information generated by the respective TSOs. The standard data contained in a military digital map is found in Table 34.

The specific TSOs selected to be evaluated in CS-1 are listed in Table 35. Since these TSOs all relate to using the digital map data to find routes for units of various sizes, the mission and scenario designed for the experiment could be less complicated than if it needed to provide context for TSOs that were intended for more widely divergent uses.

Once the TSOs were identified, their functionality was used to determine the critical aspects of each.

Table 34: Standard Digital Map Data

Digital Map Data	Description
Slope	Identifies the magnitude of the slope from the horizontal and the compass direction of downslope. Arbitrarily steep slopes slow moment (restricted terrain) or prevent most movement (highly restricted terrain)
Stem Size	Identifies the average diameter of the foliage stems (tree trunk diameter). Stem size affects the types of vehicles that can traverse the area
Stem Spacing	Identifies the average distance between foliage stems. Stem spacing affects the sizes and types of vehicles that can traverse the area
Soil Type	Identifies which of the standard soil types the area contains. The vehicle weight a given type of soil type can support and the moisture content affects the types of vehicles that can traverse the area
Hydrology	Identifies the water characteristics of the area that can limit vehicle movement through the area (streams, rivers, swamp)

Table 35: CS-1 – TSOs

CS-1 TSOs	Description
Obstacles	Identifies restrictions to movement due to natural and man-made obstacles
Mobility Corridors	Identified lanes of movement between obstacles categorized by unit size
Movement Projection (MP)	Optimizes routes given digital map data, obstacles, and user input for 12 standard vehicle classes
Choke Points	Identifies areas along a route where unit movement would be restricted due to distance between obstacles

Determine critical aspects of the GDSS

As discussed in Chapter Three, once the list of potential GDSSs has been determined, the aspects of the GDSSs that were critical to the development of the evaluation may be determined. The critical aspects were those that will impact the steps in the DQEM and later steps in the construction of the evaluation. These critical aspects were: (1) The actual functions of the GDSS, (2) The actual information output by the GDSS, (3) The types of decisions the GDSS were designed to support, (4) The target user(s), (5) The geospatial environment(s), and (6) Anticipated host system requirements. These aspects were not necessarily independent, and identifying the specific aspects will require consultation with the GDSS developers and Subject Matter Experts (SMEs), as well as referencing design requirement and doctrinal documents.

Once the TSOs to be evaluated were determined, the first of the critical aspects of the TSO, the actual functions of the TSO that will be included in the experiment can be identified. For initial planning, the intended functionality may be used; but when determining the critical aspects of the TSOs, only the actual functionality should be considered. Initial planning using the intended functionality will likely be altered to some extent when the actual functionality is determined. The functions that were actually included in the first case study were determined by a demonstration that ensured that the functionality worked. Any intended functionality that did not work was not included in the evaluation.

Table 36 presents the critical aspects of the five TSOs selected for evaluation.

Although the Obstacles TSO is presented in Table 35, it is not directly assessed in this evaluation. As discussed in Chapter Three, the effect of a GDSS on decision-making can only be directly evaluated if the decision-maker can both vary the inputs to that function and discern changes in the output. Because the Obstacles TSO uses only the terrain data, the output of the Obstacles TSO is static for any given terrain. Since the decision-makers were not able to vary the input or observe a change in the output, the effect of this TSO on the decision-makers' decision would be constant. This TSO would have no effect on the results in a repeated measures structured evaluation. Therefore, this TSO is not considered in the remainder of this case study.

The demonstration that was used to determine the functions that would actually be evaluated also identified the outputs of the TSOs (Table 35). The outputs of the TSOs are closely linked to their functions, but the way the outputs were displayed would affect the decisions the TSOs could most easily be used to support.

Table 36: CS-1 – Critical Aspects of the GDSS

TSO	Actual TSO Functions	Output of TSO	Decisions Supported	Geospatial Environ-ment
Obstacles	Identify areas of highly restricted and restricted terrain due to terrain factors	Combined Obstacle Overlay (COO) with areas of highly restricted and restricted terrain identified	Identification of areas through which vehicular travel is not possible	Open terrain with natural and man-made obstacles to vehicular movement. The terrain must have MC that vary along the potential routes, terrain that exhibits choke points for variously sized units, and both potential on-road and off-road routes. There should be enough variation in the terrain that multiple routes are possible.
	Identify areas of highly restricted and restricted terrain due to man-made obstacles			
Mobility Corridors (MCs)	Identify Mobility Corridors (MCs)	Overlay with MCs categorized by size	Identification of possible routes in support of Course Of Action (COA) development	
Movement Projection	Identify potential routes by primary vehicle type	Overlay of potential routes for designated vehicle types		
	Optimize routes for time	Overlay of fastest routes by vehicle types		
	Optimize routes for distance	Overlay of shortest routes for designated vehicle types		
	Optimize routes given user input barriers	Overlay with routes that do not cross user input barriers		
	Optimizes routes for both on-road and off-road	Overlay of routes specified by user selection of on- or off-road		
	Calculate travel times for vehicle type	Travel times for generated routes by vehicle type		
Choke Points (CPs)	Identify areas of restricted movement due to unit size	Overlay of areas along routes where movement will be restricted categorized by unit size		

The decisions that the TSOs could support were derived from the actual TSO functions and their outputs. The actual functions were applied in the context of a military planning problem to determine what specific military planning decisions could be affected by the TSOs' functions. The decisions supported by the TSO functions were determined by SMEs who analyzed the actual functions and their outputs with respect to the planning problem. As seen in Table 36, the Choke Point (CP), Mobility Corridor (MC), and Movement Projection (MP) TSOs all supported the same decision i.e. in the identification of possible routes. Since the Military Decision Making Process (MDMP) is organized in phases that require specific decisions, the phase of the MDMP that would be most affected by information generated by the TSOs was determined. In this case, the decisions best supported by the TSOs were associated with products that previously had been manually generated during the IPB process.

There were two additional critical aspects noted in Chapter Three that were identified from the five previously identified GDSS: the target user and host environment. Discussions with experienced SMEs provided input on the level of decision-making at which each TSO would be most valuable to the military decision-maker. Because the functions of all the TSOs were related to terrain analysis, the target user, in this case a terrain analyst, is consistent for all the TSOs. Likewise, all the experiment TSOs have the same host system requirements in order to function; and therefore, the environment is also constant.

The host system used in this case study was the fielded version of the DTSS suite of tools, as implemented using ARC-GIS 9.1. The DTSS tool suite consists of a package

of software tools used to generate tactical decision aids for producing a number of products including (1) off-road and on-road speed products; (2) Combined Obstacle Overlays (COOs); (3) shaded time/distance, maneuver networks, and predictions; (4) masked/visible areas for observation; and (5) fields of fire, cover and concealment, obstacles, key terrain, and avenues of approach. DTSS was selected because it provided the necessary support for the BTRA-BC TSOs (interactive display tools and database management functions) and because the target users were familiar with its use. The compatibility of DTSS with BTRA-BC had been previously demonstrated when less sophisticated BTRA-BC TSOs were fielded with DTSS.

The actual TSO functions, outputs of the TSOs, and decisions affected were used to generate two of the other experimental structure elements (1) general mission and tasks and (2) scenario and planning tasks.

Determine general mission and general tasks and geospatial context

The actual GDSS functionality, the outputs of the GDSS, and the decisions supported by the GDSS were used to develop a general mission for the case study that was similar to that developed as an example in Chapter 3:

> General Mission: Conduct an analysis of the terrain in brigade area of operations and generate potential routes in support of the movement of battalion-sized units from the assembly area to the objective.

The unit size, brigade, is specified here instead of in the scenario because the target user, a terrain analyst, impacts the selection of the unit size. Currently, terrain

analysts are attached to brigades and larger units so the units chosen should be of brigade size or larger. On the other hand, the route planning and movement plans necessarily get more detailed as the unit size gets smaller. These constraints imply that the smallest unit, brigade, would be the most appropriate. Generally, units analyze the movement for subordinate units one and two echelons down, so a brigade would plan routes for battalion- and company-sized units. Units of these sizes would require sufficiently detailed route planning to exercise the functions of the BTRA-BC GDSS. Therefore, the unit size chosen was the brigade.

The three additional TSOs added for the evaluation, beyond the MP TSO discussed in Chapter Three, supported an additional decision: the determination of areas through which vehicular travel is not possible. This supported decision could have led to an additional general task, but this specific supported decision is generally included in the general task as specified in Chapter Three: generate valid routes for course=of-action development. Therefore, no additional general task was required to be identified for these additional functions.

Once the general task was identified, the requirements for a geospatial context that would support the tasks could be derived. In the case of the MP TSO these requirements could be as follows:

> Open terrain with natural and man-made obstacles to vehicular movement. The terrain must have MCs that vary along the potential routes, terrain that exhibits choke points for variously sized units, and both potential on-road

and off-road routes. There should be enough variation in the terrain that multiple routes are possible.

A survey of potential areas meeting these criteria that also had geospatial data structure to support the TSOs yielded two possibilities: the National Training Center (NTC) in Fort Irwin, CA and portions of the Korean peninsula. Trials with versions of the TSO still in development indicated that the terrain at NTC would provide the necessary variety of terrain to provide a challenging problem that would require the decision-makers to use all the functions of the TSO.

Determine general decision quality characteristics and hypotheses

Once the general tasks and general mission were identified, general characteristics of the decisions and the hypotheses to be assessed were determined. In this case study, three general decision quality characteristics and the associated hypotheses discussed in Chapter Three were used as the basis for the hypotheses described below. Because this evaluation was going to use a repeated measure design, an additional two (hypotheses 4 and 6) were added. The six hypotheses for this case study stated that trained, experienced, military personnel who use the GDSS would:

1. *Produce a higher quality plan than personnel not using the GDSS.* Rationale: The automaton in BTRA-BC should reduce errors of omission and calculation. Furthermore, the standardized graphical representation of important terrain features and decision graphics will display information more succinctly and allow decision-makers to evaluate planning options more easily thus improving the quality of their plans.

2. *Produce the designated plans more quickly than personnel not using the GDSS.* Rationale: The automation and analysis functions in a GDSS should allow the decision-makers to complete the repetitive, rote tasks and analyses more quickly. This time saving should translate into time savings for the overall set of tasks. As most military planning is done with significant time constraints, the time saved may be used to improve the quality of the output plan.
3. *Display as good an understanding of the impact of the given terrain on military decision-making than personnel not using the GDSS.* Rationale: The cognitive process required to complete the required tasks when using a GDSS will still require the decision-maker to be as intimately familiar with the terrain and its effects as when performing the tasks manually.
4. *Produce decisions with BTRA-BC that are more uniform, i.e. have less variance in the first two of the three categories above (speed and quality), than output generated without the use of BTRA-BC.* Rationale: The automation incorporated into the TSOs should provide a consistent quality of information to decision-makers which should contribute to less variation in the output decisions.
5. *Consider using a GDSS superior with respect to (1) allowing them to complete the tasks more quickly, (2) allowing them to produce higher quality output, (3) allowing then to have a greater terrain understanding, and (4) overall.* Rationale: A well-designed GDSS should increase the decision-

makers' confidence in their planning and decision making as a result of using automated tools. Without such confidence, a GDSS, no matter how well designed, would not see much use.

The following general hypothesis does not deal with the evaluation of decision quality but will provide control information for the statistical analysis.

6. *There should not be a learning effect due to experimental design.* Rationale: The structure of the experiments requires the repetition of various tasks which results in concern that a learning effect might skew the results of the experiment. As the decision-makers have previous training, extensive experience using C2 planning tools, the tasks the decision-makers are asked to perform are those that they have performed in the normal course of their duties; and they are trained to proficiency with the GDSS, there will not be a learning effect.

The analysis of this hypothesis will indicate if there was a bias in the evaluation due to the decision-makers' decisions improving due to exposure to the host system, the GDSS, and the type of problem. This analysis will allow any variation in decision due to a learning effect to be taken into account when analyzing for decision quality. Also a large learning effect may have called the validity of the results into question if the learning effects were significant compared to the change in decision quality.

Decompose decision quality and define decision quality characteristics

The decision quality sub-characteristics used in this case study were the result of research and discussion by two independent SMEs. The SMEs were experienced brigade

planners having each held staff and command positions. Each of the SMEs used the critical aspects of the GDSS, the general decision quality characteristics, the hypotheses, the general mission, and tasks as the basis for developing the decision quality characteristics. Each SME independently reviewed the guidance, doctrine, and procedures of the MDMP and generated a list of the characteristics of a good plan that met the mission parameters. They compared their individual lists, discussed the relevant portions of the guidance, attributed each sub-characteristic to a general characteristic, and reached consensus on the characteristics that needed to be considered when evaluating the quality of plans generated in support of the mission.

Initially, the list of decision quality characteristics and sub-characteristics was not as extensive as those presented in Table 37. As discussed in Chapter Three, the process of decomposing decision quality characteristics and the generation of measures is an iterative process that is only completed when the measures used to assess the decision quality characteristics address a single concept that an SME can evaluate with a single judgment. In this case study, neither SME's initial list of characteristics was as complete as that in Table 37 for two reasons: (1) Each SME's list omitted some sub-characteristics and also included sub-characteristics that were not included in the final consensus; and (2) The final decomposition resulting from the iterative process also included sub-characteristics neither SME had initially defined.

Interestingly, the two SMEs suggested the inclusion of sub-characteristics related to the overall quality of the plan in addition to sub-characteristics directly related to the decision quality characteristics. The SMEs saw the quality of the presentation of the plan

as an indication of the decision-makers' understanding of the terrain's effect on the mission and the completeness and clarity of the underlying terrain analysis. Table 37 shows the complete list of quality characteristics, the decomposition of the characteristics, and the rationale behind each sub-characteristic.

Table 37: CS-1 – Decision Characteristic Decomposition

Decision Quality Characteristics			Rationale
Quality of Mobility Corridors			Generating Mobility Corridors is a result of detailed analysis of the COO and judging the size unit that can pass through the terrain. This analysis is necessary for follow- on steps and the quality of the MCs will affect follow-on decisions (e.g. CPs and AoAs).
	MC locations clearly indicated		The location and relationship to other mobility corridors is necessary for analyzing routes.
	MCs Categorized by size and clearly indicated		The effect of the MC on routing depends on its size. Clearly indicating the size category of the MCs is necessary for efficiently analyzing routes for units of varying sizes.
	MCs sized correctly		The accuracy with which the MCs are sized will directly affect analyzing routes.
Quality of potential AoAs			A primary goal of terrain analysis is to generate routes that are suitable for the movement of units through the given terrain. The quality of the AoAs generated will affect the selection of the recommended AoAs and the movement plan.
	Quality of routes		The route is the culmination of much terrain analysis and is a primary input to movement plans and COAs.
		Valid start point	The AoA starts from a point in a secure area. No additional routing will be required to get to the start point.
		Valid end point	The AoA get the force where it needs to go.
		No unnecessary turns	Turns may require changes of formation, may cause delay, and may expose units to observation.
		independent routes	The choice between AOAs needs to be between routes that do not use same the MCs.
		Choke Points on Route	Minimum number of choke points is desirable to reduce formation changes and thus the time not accounted for in automated time calculations.
		Transit times calculated	Transit times calculated for correct three vehicles
		Size of limiting Choke Point	The size of the smallest choke point limits the size of unit that can move along the route without changing formation.
		Avoid obstacles	Routes that pass through built up areas or highly restricted terrain will disrupt and slow movement.
	Routes clearly indicated		The usefulness of a terrain routing product is partially in its ability to convey information quickly. Well defined routes

			also typically indicate completeness in the analysis.
		AoAs clearly labeled	Labeling aids analysis and discussion during briefing and indicates the planner identified different routes.
		AoAs uniquely identified	Using some graphic to identify each route avoids confusion during analysis.
		AoAs obvious on digital display	A measure of the overall awareness of the planner of the "big picture" of AoA planning
		Buffered Routes	Battalion routes buffered with a 1000M band improve the analysis of obstacle clearance.
Quality of Choke Points			Generating choke points is a result of analysis of the COO and MCs and consists of identifying points/areas where movement is restricted, i.e. a unit must break into sub units or change formation to pass. CPs are also potential ambush sites and negatively affect the route selection for COAs. This CP analysis is critical for follow-on steps, and the quality of the CPs will affect follow-on decisions (e.g. AoAs).
	CPs on AoAs are clearly indicated		The location and relationship to AoAs is necessary for analyzing routes.
	CPs on AoAs are categorized by size		The effect of the CO on routing depends on its size. Clearly indicating the size category of the MCs is necessary for analyzing whether formation changes or possible ambush points occur along routes.
	CPs on AoAs are sized correctly		The accuracy with which the MCs are sized will directly affect analyzing routes.
	CPs not on AoAs are indicated, categorized and sized correctly		Analyzing choke points off of the AoAs indicates the planner is thinking of how the possible enemy routes may affect the selection or recommended AoA.
Quality of Recommended AoAs			Using the information from the analysis of MCs, CPs, and the potential AoAs generated the choice of AoAs to recommend will require subjective judgment of all factors.
	Recommended AoAs clearly indicated		Like the CPs and MCs, analyzing AoA requires that the recommended AoAs be clearly indicated. Non-clearly indicated AoAs will lead to confusion during analysis.
		Recommended AoAs labeled	Labeling aids in analysis and discussion during brief and indicates the planner identified different routes.
		Recommended AoAs uniquely indicated	Using some graphic to identify each route avoids confusion during analysis.
	Recommended AoAs meet mission specifications		Part of the analysis of AoAs requires determining whether AoAs meet mission requirements. AoAs that don't meet mission requirements are not valid AoAs.
		Appropriate for BN-sized units	The AoAs route must be able to support the proper sized units.
		Avoid built up areas	AoAs must avoid built up areas as these will restrict movement and be possible ambush sites.
		1st and 2nd choices are independent routes	The choice between AoAs needs to be between routes that do not use the same route or portion of route.
		Analyzed for on-road and off-road	Needed for proper analysis of all routing options
	Recommended AoAs are better than non-recommended		Recommended AoA should be the best choices based on doctrine and information generated from analysis.

	AoAs		
Quality of BN Boundaries			BN Boundaries are generated after the analysis of recommended AoAs and require judgment and understanding of the terrain and its impact on unit operations.
	BN boundaries give BNs room to maneuver		The BN boundaries are placed so that BNs have sufficient maneuver room based upon the terrain along the AoAs.
	BN Boundaries follows natural geographic features		Natural features allow geographic features to be used as boundary landmarks.
	AoAs are within BN AOO		AoAs that cross BN boundaries may subject friendly forces to blue-on-blue fire. Both routes must be encompassed by the BN boundaries else alternate BN boundaries need to be specified with alternate AoAs.
Plan Quality			This is an overall judgment of the coherence of all analyses and the ability of the analyses to be briefed to the CO.
	Essential Information included		Is all the essential terrain analysis information included in the plan?
	Information clearly presented		Indication of the clarity and completeness of the terrain analysis. Is the information presented such that the plan can be briefed as is or is further work required?
		Information clearly labeled	Are all critical aspects of the AoAs labeled?
		Information uniquely identified	Are all the aspects of the plan uniquely identified to avoid confusion?
	Plan meets all mission specifications		Plans for which mission critical aspects are not complete and indicate that an incomplete analysis was done.

Determine scenario and planning tasks

Because all the TSO functions support AoA generation, the scenario uses the mission developed as an example in Chapter Three:

> Elements of the 2nd Brigade Combat Team consisting of the 1/6 and 2/6 Mechanized Infantry Battalions and the 1/35 Armored Battalion will advance from their current position in Assembly Area BOSTON northwest of Phase Line MIAMI to assault hostile units (mechanized infantry battalion augmented by a heavy armored company) in

Engagement Area DIAMOND southeast of Phase Line PEARL in order to occupy said position.

Likewise the full evaluation uses Commander's Intent presented in Chapter Three:

> 2BCT will advance in a 2 up / 1 back formation with 1-6th and 2-6th Mechanized Infantry Battalions forward and the 1-35th Armored Battalion as the reserve. 1-6th and 2-6th Mechanized Infantry Battalions will advance along two routes to arrive at their designated firing positions simultaneously. As the hostile force has had time to dig in, is of sufficient strength, and has the advantage of position, we must use surprise and misdirection to succeed in our mission. In that light, G Trp 1st CAV BRT will probe to the south to draw the attention of hostile forces while the main force advances covertly.

But, because the additional TSO functions, specific tasks must be added to ensure that the decision-makers exercise all the available functions including the added functions. The complete list of tasks becomes as follows:

Generate the following products from your analysis:

1. Combined Obstacle Overlay (COO)
2. Identified Mobility Corridors (MC) categorized by type of force
3. Four independent Avenues of Approach
    - Two routes are considered independent if they have no common MCs

4. Transit times required on each Avenue of Approach for three vehicles
    - M1 Abrams, M2 Bradley , and LAV 25
5. Choke Points in each Avenue of Approach
6. Recommended Areas of Operations (AO) boundaries for the two BNs in the BCT

These tasks are implicit in the Mission and Commander's Intent, but stating them explicitly and/or specifying the inputs/outputs required ensured that all the decision-makers exercised all the functions. The bulleted caveats in the task list above add further detail to the tasks that improve the "tightness" of the evaluation by guiding the decision-makers' efforts. These caveats are consistent with the Mission, Commander's Intent, and Task Organization (Taskorg). The entire OPORD used in CS-1, including additional detail included in the Task organization and Enemy Disposition, is attached in Appendix 4-2.

Define decision quality characteristic measures

The measures defined in conjunction with the definition of the decision quality characteristics and sub-characteristics can be evaluated in two ways: (1) objectively or (2) subjectively. Generally, there may be measures of some aspects of decision quality that can be categorized and counted or that consist of binary answers requiring little subject knowledge in order to evaluate. The objective measures for this case study are presented in Table 38. The individual decision quality sub-characteristics that can be assessed with

objective measures are associated with the various decision quality characteristics. Although these aspects are evaluated separately from the subjectively evaluated characteristics, the scores associated with these measures were aggregated with the subjective scores before conducting the statistical analysis.

Table 38: CS-1 – Objective Measures

Decision Quality Characteristics			Measures
Quality of Mobility Corridors			
	MCs Categorized by size		Are MCs categorized by size? (Binary)
	MCs sized correctly		Number of MC incorrectly sized (Integral)
Quality of potential AoAs			
	Quality of Routes		
		Choke Points on Route	Number of choke points on route (Integral)
		Size of limiting Choke Point	The size of the smallest choke point (Categorical)
		Avoid obstacles	Does the route pass through no-go area? (Binary)
	Routes clearly indicated		
		Buffered Routes	Battalion routes have 2000m buffer? (Binary)
Quality of Choke Points			
	CPs on AoAs are categorized by size		Are the MCs categorized by size? (Binary)
	CPs on AoAs are sized correctly		Number of CP identification errors (Integral) Number of AOA CP sizing errors (Integral)
Quality of Recommended AoAs			
	Recommended AoAs meet mission specifications		
		1st and 2nd choices are independent routes	Does each BN AOO have two routes? (Binary)
Quality of BN Boundaries			
	AoAs are within BN AOO		Does any AoA cross BDE boundary? (Binary)

The subjective measures generated by the SMEs are presented in Table 39. The SMEs' goal was to evaluate the overall quality of the decision-makers' terrain analysis.

As can be seen from Tables 54 and 55, some characteristics that had objectively scored measures also had measures that were scored subjectively. The measures that were scored subjectively were those that required the judgment of the SME in order to be properly evaluated. For some of the sub-characteristics, the standard the measures refer to is a "qualitative assessment based on the ability to justify the AOA to the Commanding Officer." This standard was selected by the SMEs because the ultimate goal of terrain analysis is to provide information and recommendations to a decision-maker. They felt that a well laid out plan of movement usually indicates a thorough analysis and understanding of the impact of the terrain on a unit's movement. In general, the sub-characteristics with this type of measure were not directly related to any TSO but reflected the overall quality of the plan of movement. Table 39, describes measures, subjective in this case, that were designed to assess each decision quality sub-characteristic. The actual scoring of each measure was done on a 5-point Likert scale, the construction of which is described.

Table 39: CS-1 – Subjective Measures

Decision Quality Characteristics	Measures
Quality of Mobility Corridors	
MC locations clearly indicated	Qualitative assessment of the impact of any errors. The location of errors is important.
MCs Categorized by size and clearly indicated	Qualitative assessment of the impact of any errors. The location of errors is important
MCs sized correctly	
Quality of potential AoAs	
Quality of routes	
Good route from AA to start point	Good route from AA to start point
Good route from endpoint to objective	Good route from endpoint to objective
Qualitative assessment of the impact of large (>45^{o}) and whether they are unnecessary	Qualitative assessment of whether turns are large (>45^{o}) or unnecessary
Qualitative assessment of the impact of the critical MCs being common to multiple routes	Qualitative assessment of the impact of the critical MCs being common to multiple routes
Transit times are correct and calculated for M1, Bradley , LAV	Transit times are correct and calculated for M1, Bradley , LAV.
Routes clearly indicated	
Qualitative assessment based on the ability to justify potential AoA to CO	Qualitative assessment based on the ability to justify potential AoA to CO
Quality of Choke Points	
CPs on AoAs are clearly indicated	Qualitative assessment based on the impact on the ability to justify potential AoAs to the Commanding Officer.
CPs on AoAs are categorized by size	
CPs on AoAs are sized correctly	
CPs not on AoAs are indicated, categorized and sized correctly	Qualitative assessment based on the impact on the ability to describe potential enemy AoAs to the Commanding Officer.
Quality of Recommended AoAs	
Recommended AoAs clearly indicated	
Qualitative assessment based on impact on the ability to justify recommended AoAs to the Commanding Officer	Qualitative assessment based on ability to justify recommended AoAs to the Commanding Officer.
Recommended AoAs meet mission specifications	
Mostly BN or larger MCs	Mostly BN or larger MCs used unless it is necessary to use

	used unless it is necessary to use smaller MCs	smaller MCs
	Recommended AoA do not enter or skirt built up areas. Qualitative assessment of proximity to built up areas	Recommended AoA do not enter built up areas
	Qualitative assessment of whether enough of the critical MCs are not in common	Qualitative assessment of whether enough of the critical MCs are not in common.
	One AOA in each BN AOO should be analyzed for each on-road and off-road movement	One AOA in each BN AOO should be analyzed for each on-road and off-road movement.
Recommended AoAs are better than non-recommended AoAs		Qualitative assessment of whether the best AoAs were recommended
Quality of BN Boundaries		
BN boundaries give BNs room to maneuver		Qualitative assessment: are the BN areas evenly distributed or does one BN need more room based on terrain?
BN Boundaries follows natural geographic features		Qualitative assessment based on terrain features
Plan Quality		
Essential Information included		No missing information for Recommended AoAs, Bn boundaries
Information clearly presented		
	Information clearly labeled	Qualitative assessment based on ability to justify potential overall movement plan to the Commanding Officer
	Information uniquely identified	
Plan meets all mission specifications		All major points in mission and commander's intent are covered.

Determine decision-maker characteristics

As indicated in Chapter Three, the target user was a trained terrain analyst with experience on a brigade staff. The decision-makers were actually Army and Marine Corps enlisted personnel who had all been trained as Terrain Analysts. All decision-makers had completed the Basic Terrain Analysis Course (BTAC) at the National Geospatial Intelligence Agency University (NGAU) and were currently enrolled in the

Advanced Terrain Analyst Course (ATAC). The 18 decision-makers consisted of 1 Army Chief Warrant Officer, 10 Army Staff Sergeants, 3 Army Sergeants, 3 Marine Corps Sergeants, and a Marine Corps Corporal. Their operational experience with terrain analysis varied from several years of continuous experience to only formal training.

Experimental structure and procedures

Determine evaluation structure

The evaluation structure of this first case study formed the basis of the discussion in Chapter Three. Specifically, the discussions concerning the determination of the decision-makers (15-16) and the determination of independent variables (System, System Order, and Scenario Order) and balancing the groups are completely consistent with the previous discussions. Table 56 and 57 illustrate the basic 3x2 structure of the evaluation and the four balanced groups into which the decision-makers were divided to support the evaluation of the effect of the System Order and Scenario Order independent variables.

Data Collection

Because the experiment in CS-1 had dual purposes (1) to assess the impact of using decision quality as the primary measure of the effectiveness of using the GDSS , and (2) to evaluate the BTRA-BC TSOs, two sets of data were collected. The data used in the evaluation of the first purpose was that specified in the tasks given to the decision-makers:

1. Combined Obstacle Overlay (COO)
2. Identified Mobility Corridors (MC) categorized by type of force

3. Four independent Avenues of Approach
4. Transit times required on each Avenue of Approach for three vehicles (M1 Abrams, M2 Bradley, and LAV 25)
5. Choke Points in each Avenue of Approach
6. Recommended Areas of Operations (AO) boundaries for the two BNs in the BCT

This data was saved in the form of a graphic representation of the decision-makers' plans of movement.

Table 40: CS-1 – Design Elements

Case Study One Design Elements		
Variable	**Manipulation**	**Levels**
System	Within Subject	With System Without System
System Order	Between Subjects	With System then Without System Without System then With System
Scenario Order	Between Subjects	Scenario 1 then Scenario 2 Scenario 2 then Scenario 1

Table 41: CS-1 – Experimental Groups

Case Study One Experimental Groups		
	Scenario Order	
System Order	With System then Without System Scenario 1 then Scenario 2	With System then Without System Scenario 2 then Scenario 1
	Without System then With System Scenario 1 then Scenario 2	Without System then With System Scenario 2 then Scenario 1

In addition to the plans generated by the decision-makers, the results of the objective and subjective evaluation of those plans were required to be able to assess the quality of those plans, i.e. the quality of the decisions made by the decision-makers. Data was gathered on the measures for each sub-characteristic as defined in step VII. This data was gathered by questionnaire, the construction of which is described in the discussion of step XI.

The evaluation of purpose (2) required all the data for purpose (1) as well as:

1. A record of the time each decision-maker required to complete the tasks
2. Information on the decision-makers understanding of the impact of terrain on the tasks
3. A subjective comparison of the relative merits of DTSS and BTRA
4. Biographic information on the decision-makers for possible post hoc analysis

With the exception of the record of completion times which was compiled by the evaluators, the above data was elicited by questionnaire and post trial discussions. The construction of these questionnaires was similar to that used for the SME evaluation and is described in the discussion of Material Preparation.

Design training

In addition to the description of the general training requirements presented in Chapter Three, several steps were taken during training to ensure that the decision-makers were trained to proficiency. Procedures were also used to limit the decision-

makers' knowledge of the capabilities of BTRA-BC prior to using DTSS. The SMEs and evaluators felt that exposure to the automated analysis tools of the BTRA-BC prior to performing the tasks with DTSS might influence the decision-makers' use of DTSS. The goal of the training on both DTSS and BTRA-BC was to train the decision-makers to the point that they were familiar enough with the operation of both tools that the manipulations required to use the system would not interfere with their use of the system

Although all the decision-makers had previous formal training in the use of DTSS, their recent practical experience with the system varied widely. In order to ensure that the decision-makers were proficient in the use of DTSS, refamiliarization training was conducted with all the decision-makers as one group. This training on DTSS and the subsequent training on BTRA-BC followed the same format. Because half the workstations were configured for DTSS and half for BTRA-BC, two decision-makers were assigned to each of the DTSS machines for the refamiliarization training. Training slides were projected on a large screen visible to all the decision-makers at their workstations. The decision-makers had hardcopies of the slides for note-taking and a digital version was available on one of two screens at the workstation. The training was organized as an introduction to the planning problem and training on each DTSS function. The training consisted of the purpose of each function, the method of accessing the function, and descriptions of the possible inputs and outputs. The decision-makers duplicated the instructor's exploration of each function on their own workstations. The trainer verified that both decision-makers at each workstation were comfortable with each function prior to moving on to the next function. Following this hands-on walkthrough,

the decision-makers were guided through a further exploration of the use of the DTSS functions by a scenario that was a simplified version of the scenarios they would see during the evaluation. The decision-makers as a whole were not deemed proficient in using DTSS until each decision-maker had completed all the tasks in the simplified scenario and stated that they were comfortable using DTSS.

Upon completion of the refamiliarization training, the decision-makers who would be using BTRA-BC in the first trial were trained on its functions in a manner identical to the refamiliarization training. As a supplement to this training, the decision-makers were also given a "cheat sheet" that explained the various BTRA function icons. The decision-makers who were to use BTRA-BC in their second trial received identical training with BTRA-BC just before their second trial. Conducting this training immediately before each subgroup performed the tasks with BTRA-BC was intended to lessen any impact of the knowledge of BTRA-BCs' capabilities on the trials performed without BTRA-BC. Because of the repeated, balanced structure of the evaluation, half of the decision-makers would be exposed to BTRA-BC in the first trial and then perform the second trial without BTRA-BC. Using BTRA-BC before DTSS seemed to have an effect on the decision-makers' use of DTSS. The analysis of this effect is discussed in (Powell et al., 2008), and the effect as it pertains to decision-making is discussed the results section below.

Conduct human subjects review

The application to the Human Subjects Review Board (HSRB) was submitted three months prior to the date of the evaluation. Because the evaluation was actually

conducted by the program sponsor, the sponsor provided a letter stating that their review indicated that there was no potential harm to the decision-makers. The HSRB determined that the evaluation was exempt from additional HSRB review.

Material preparation

As discussed in Chapter Three, the preparation of the materials used in the evaluation can significantly impact the conduct of the evaluation as well as the data collected. Obviously, the questionnaires, materials, and procedures used to collect the data can directly affect the data gathered and thus the results of an evaluation. Likewise, any materials such as training slides and scenarios with which the decision-makers interact can influence their perceptions of the GDSS and thus influence the data gathered from their subjective evaluations. Less obvious is the impact of the other supporting materials such as briefings and seating charts that can also affect the success of the evaluation.

Other than the questionnaires, into which much thought was put, the materials that were of particular importance are the training and scenario. Because the decision-makers must be comfortable (trained to proficiency) using the host system and the GDSS, the training slides and script were rigorously vetted to ensure that the information was presented logically and coherently. Ill-prepared training materials can reduce the effectiveness of the training, result in frustration on the part of the decision-makers, and ultimately affect the decision-makers' perceptions of the GDSS. Likewise, special care was taken in the construction of the scenarios. The scenarios must provide sufficient background and justification for the mission and tasks the decision-makers are asked to

undertake. A lack of detail can result in decision-makers not taking the tasks seriously and possibly not providing the detailed evaluation desired.

Overall, the materials used in the evaluation that are well prepared should almost go unnoticed. No errors should be noticed by the decision-makers; the materials should be clear and generate few requests for clarification. The materials generated for CS-1 were as follows:

- Briefings – given by evaluators and sponsors that define the reasons for conducting the evaluation, the importance of the decision-makers' input, how the decision-makers' input will affect the development of the GDSS, and the desire that the decision-makers' give us their frank evaluations of the GDSS Copies of the briefing slides given to the decision-makers will allow them to take notes which will aid in their retention of the information presented in the briefings.
- Informed consent – explanation that participation is voluntary, that there is no risk to the decision-makers, and that their input will be anonymous.
- Decision-maker designation in preaddressed envelopes – giving the decision-makers a designation provides anonymity while providing a means of keeping track of each decision-maker's responses. Pre-addressing the envelopes prevents confusion and enhances anonymity during designation distribution.
- BTRA-BC TSO cheat sheet – cheat sheets which summarize training and procedures provide a reference for the decision-makers and will help reduce frustration on the part of the decision-makers when using a new system.

- Seating chart for each session – to reduce confusion, the decision-makers should be assigned to workstations by decision-maker designation. This is done by labeling each workstation with a decision-maker's designation. This accomplishes two purposes: (1) The workstations can be preloaded with the version of the system that a particular decision-maker will be using during the trial; and (2) Cross contamination among the decision-makers will be reduced by ensuring that decision-makers working on the same scenario are not seated next to each other or not able to seen each other's screens.
- Other supporting documents – documents such as preformatted sheets to record the start time, finish time, and the duration any breaks taken by the decision-makers should be pre-staged.
- Questionnaires for decision-makers – constructed as discussed in Chapter Three, the evaluation questionnaires included a questionnaire that elicited subjective evaluations of each TSO, a questionnaire eliciting subjective evaluations of the usefulness of using each TSO compared to not using it, and a questionnaire eliciting each decision-maker's understanding of the terrain and its impact on their decision-making. These questionnaires were constructed using Likert scale numerical responses and comment blocks. The number of questions was sufficient to elicit data on all the TSO functions but limited in number to reduce fatigue on the part of the decision-makers. Also cross validation questions were not used for this reason and due to SME

perception that the decision-makers would consider "answering the same question twice" a waste of time and color their responses.

- Questionnaires for evaluators – since these questionnaires were not used until the SMEs evaluated the decision-makers' plans, these were the only material not completely constructed prior to the Pilot Test. For an evaluation designed solely to evaluate relative decision quality, most of the generation of the evaluator questionnaires could be completed prior to the Pilot test. Since the secondary purpose of the evaluation was to assess the value of the DQEM, the evaluation of decision quality with successive iterations of these questionnaires was desired; and these evaluation iterations (and thus the construction of the questionnaires) could only be conducted after the decision-makers generated their plans. All the questionnaires used in CS-1, with the exception of the evaluator questionnaires, were evaluated during the Pilot Test described in section XII.

All The materials used in the evaluation were prepared carefully in advance of the Pilot Test and were evaluated by the pilot test decision-makers, SMEs, evaluation designers, and the technical support staff. The questionnaires that were used in CS-1 can be found in Appendices 4-3 and 4-4.

Review and practice procedures

Problems related to the functioning of the TSOs, the TSO interface with DTSS, or the TSOs not responding as anticipated were discovered during the technical walkthrough described in Chapter Three. Discovery and subsequent resolution of several of these

problems during the evaluation would have delayed the evaluation and resulted in the loss of data. The technical walkthrough reduced the likelihood that the decision-makers would not able to evaluate functions that were not responding properly. The technical walkthrough was conducted far enough in advance of the Pilot Test (two weeks) that these problems could be corrected or so that the experimental procedures and training could be updated to reflect the current state of the system. The goal of the technical walkthrough was to resolve all the technical issues prior to assessing the experimental design and procedures with representative users in the Pilot test.

Like the Technical Walkthrough, the Pilot Test was conducted far enough in advance of the actual evaluation (two weeks) so that deficiencies in the paperwork, training, and evaluation procedures could be resolved prior to actually conducting the evaluation. The pilot test was the first time that all aspects of the evaluation were brought together. The pilot test was a full dress rehearsal conducted with two pilot-test decision-makers (ATAC instructors) just as it was planned to be with the actual evaluation decision-makers. The pilot test decision-makers were encouraged to assess every aspect of the evaluation including the operation of the system, the training, administrative procedures, scenarios, tasks, questionnaires, etc. Improvements in these were made on the basis of the decision-makers' recommendations.

As a result of conducting the Technical Walkthrough and Pilot Test, no significant delays were encountered during the conduct of the evaluation.

Conduct Evaluation

The general considerations associated with the conduct of the evaluation were discussed in Chapter Three and modified as described in the preceding sections.

Analysis

The analysis of the collected data was in accordance with the analyses discussed in Chapter Three with specific analyses highlighted below.

Determine scoring criteria

The determination of scoring criteria is the process by which the SMEs reach consensus on the specific attributes of decision quality that are required to attain a specific score on a specific decision quality measure. The determination of scoring criteria is similar for objective and subjective evaluations though typically the determination of scoring criteria for subjective evaluations is more complex.

Objective Scoring

In CS-1, some of the measures that supported the decision characteristics were able to be determined objectively. The objective scoring of these measures still required evaluators to assess each plan, but these elements had binary (yes/No) or integer (1, 2, 3…) answers. Essentially even the integer assessments were a series of binary assessments that could be determined by looking at the digital representation of the plan. The objective scoring measures were determined in consultation with the SMEs, and the objective scoring was conducted by the researchers. Unlike the subjective scoring, there

was little discussion needed to determine elements that would affect the measure. Like the subjective scoring, a consensus was reached concerning the Likert scale to ensure consistent scoring for non-binary measures. The objective scoring criteria can be found in Appendix 4-5. The data generated from the objective scorings can be found in Appendix 4-8.

Subjective scoring

In CS-1, the subjective scoring was conducted in three iterations in order to determine the effect of the DQEM on the SMEs' ability to determine relative decision quality. These three successive scorings were not required for the evaluation of the value of the GDS to decision-making since only the data from the final scoring was used for this purpose. Because successive scorings of plans were used, there was concern that a previous scoring of a given plan might contaminate a later scoring; but, because there were 36 separate similar plans generated by the 18 decision-makers in CS-1, it seemed possible to have the SMEs do preliminary scorings of samples of the 36 plans without fear of previous scoring contaminating the final scorings. Efforts were made to minimize rescoring of a plan by the same SME; but due to the large number of similar plans and the large number of measures, it was considered unlikely that an SME would remember how he previously scored a given plan if he recognized it. The SMEs' inability to remember previous scored plans was confirmed verbally by both SMEs at the conclusion of the scoring.

The initial scoring consisted of five plans randomly selected from those generated using the BTRA-BC GDSS (BTRA) and an equal number from those not using the GDSS

(DTSS). Each SME evaluated the same 10 plans using the agreed upon measures of decomposed decision quality characteristics but without any discussion of what elements of the plans would specifically contribute to the score for each measure. Each SME was using his own experience in terrain analysis and planning to score the plans on each measure. As mentioned in the experimental design section, the SMEs evaluated each measure on a 5-point Likert scale; and Figure 17 illustrates a first scoring SME evaluation question with a generic Likert scale and no criteria on which to evaluate the measure. To reduce cross contamination, each plan was scored on all measures before proceeding to the next plan. SMEs were allowed to rescore plans to the extent that they felt all plans were scored fairly. Scoring using questions such as these yielded poor correlations (<< 0.7) between the SMEs' scores which indicated that further refinement of the scoring method was warranted. Specific results are presented in the discussions of statistical analysis.

Do AoAs take a direct route from phase line to phase line?

Avenues of Approach	Poor		Fair		Good
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

Criteria
None designated

Figure 17: CS-1 − Example of Question without Criteria

The second scoring consisted of 10 randomly selected plans from each of the 15 remaining plans generated by using BTRA and DTSS. Therefore there was no overlap with the plans used in the initial scoring. In the second scoring, a consensus was reached among the SMEs and the researcher concerning the elements of a plan that should be considered in the scoring of each measure. For this scoring, the Likert scales remained the same as for the initial scoring, i.e. they were not tailored for each measure and the relative value of the presence or absence of specific elements present to the measure score was not discussed. In contrast to the initial scoring, the SMEs now had more guidance (criteria) concerning the elements of the plan on which they should concentrate their evaluation and which elements should affect which measure (Figure 18). They did not know how each criterion should affect the measure, i.e. the Likert scales had not been tailored for each measure. Scoring using Likert scales that incorporate criteria showed higher correlations between the SMEs' previous scores but generally did not result in statistically significant agreement.

Do AoAs take a direct route from phase line to phase line?

Avenues of Approach	Poor		Fair		Good
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

Criteria

1. Valid start point (behind PL and valid polygon)
2. Valid end Point (beyond PL and valid polygon)
3. No unnecessary turns
4. Routes are independent
5. Transit times calculated

Figure 18: CS-1 − Example of question with criteria

In the third and final scoring, the SMEs and the researcher reached consensus on how each element discussed in the second scoring would affect the scoring of the applicable measure. The Likert scales for each measure were tailored so that the requirements for each score were defined and understood by the SMEs (Figure 19). In comparison to the previous two scorings, the SMEs now had specific guidance as to what was required to achieve specific Likert scores on each measure. This guidance did not eliminate the subjective component to the scoring. As can be seen in Appendix 4-6, the judgment of the SMEs was still required to evaluate how well each plan fulfilled the requirements.

Do AoAs take a direct route from phase line to phase line?

Avenues of Approach	meets less than 2 criteria	Meets 2 of 5 criteria	Meets criteria 1,2, & 4	Meets criteria 1,2, & 4,5	Direct - meet all 5 criteria
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

Criteria

1. Valid start point (behind PL and valid polygon)
2. Valid end Point (beyond PL and valid polygon)
3. No unnecessary turns
4. Routes are independent
5. Transit times calculated

Figure 19: CS-1 − Example question with Likert scale consensus

It was intended that each SME would score all 36 plans after consensus was achieved on the Likert scale criteria, including rescoring the 20 most recently scored; but due to availability constraints, one SME only rescored the 20 plans from the second scoring while the remaining SME scored all 36 plans. The correlation between the SMEs' scores for the individual sub-characteristics and the overall scores on the 20 plans rescored by both SMEs were high enough (> 0.7) that the scores from the plans that only one SME scored were used directly in the final data analysis of the effectiveness of the GDSS in decision-making. The discussion of the correlation between the evaluation scores of the two SMEs is discussed in the results section below.

Upon completion of the third scoring, the SMEs were asked to provide weights that would indicate the relative importance of the questions to their evaluation of decision quality. As discussed in chapter 3, decision quality sub-characteristics were grouped into

functional areas that were used to generate questions that were included in the questionnaires the SMEs used in their evaluations. The SMEs were asked to provide weights, on a scale of one to ten, which represented the relative importance of the functional area of each question to the overall decision quality of the plan. These weights were elicited after the third scoring so that the SMEs would be the most familiar with the aspects of the functional decomposition. These weights were used to explore whether a weighted average of decision quality characteristic scores was a better estimate of the overall decision quality score than a simple average.

Determine aggregation criteria

In order to explore the effects of aggregating objective and subjective measures of decision quality and the effect of weighting decision quality characteristic scores, several aggregations were developed. Initially, the objective and subjective measures were not aggregated and scored separately. This is the method that was used in the evaluation of the effectiveness of the GDSS. Subsequently, the objective and subjective score were aggregated with both simple and weighted averages; and the agreement in the SMEs' scores were compared. Both the simple and weighted aggregations were used in the assessment of the usefulness of the DQEM.

The objective and subjective evaluation questions were derived directly from the decision quality characteristic and sub-characteristics, but they are not in a one-to-one correspondence. Some questions assess single decision quality sub-characteristic measures, some assess multiple sub-characteristic measures, and others assess all the measures associated with all the sub-characteristics of a decision quality characteristic.

Further complicating the aggregation, the objective measures do not assess all the sub-characteristics addressed by the subjective measures; although there was some overlap. Also, because the measures were grouped for ease of evaluation and not rigidly grouped as in the decision quality decomposition, the scores for the measures need to be aggregated with respect to the sub-characteristics to which they correspond.

Two aggregations of the objective and subjective measure scores were investigated: aggregating scores from the objective and subjective measure scores using a simple average and using a weighted average. For aggregation using a simple average, any objective score for a decision quality measure was averaged with any corresponding subjective score; and all the resulting decision quality measure scores were averaged to obtain the decision quality sub-characteristic score. Characteristic scores were obtained through a simple average of sub-characteristic scores. Likewise the overall decision quality score was the average of the decision quality characteristic scores. Using the simple average, the overall decision quality scores was the simple averages of all the measure scores.

Table 42: CS-1 – Matching of Objective and Subjective Scoring Criteria to Decision Quality Characteristics

Decision Quality Characteristics			Subjective Question	Objective Question
Quality of Mobility Corridors				
	MC locations clearly indicated		S-3	
	MCs categorized by size and clearly indicated			
	MCs sized correctly			O-1
Quality of potential AoAs				
	Quality of routes			
		Valid start point	S-1.1	
		Valid end point	S-1.2	
		No unnecessary turns	S-1.3	
		Independent routes	S-1.4	
		Choke Points on Route		O-2,3
		Transit times calculated	S-1.5	
		Size of limiting Choke Point		O-5
		Avoid obstacles		O-6
	Routes clearly indicated			
		AoAs clearly labeled	S-2	
		AoAs uniquely identified		
		AoAs obvious on digital display		
		Buffered Routes		O-8
Quality of Choke Points				
	CPs on AoAs are clearly indicated		S-4.1	
	CPs on AoAs are categorized by size		S-4.2	
	CPs on AoAs are sized correctly		S-4.3	O-4
	CPs not on AoAs are indicated, categorized and sized correctly		4-4.4	
Quality of Recommended AoAs				
	Recommended AoAs clearly indicated			
		Recommended AoAs labeled	S-5	
		Recommended AoAs uniquely indicated		
	Recommended AoAs meet mission specifications			
		Appropriate for BN-sized units	S-6.1	
		Avoid built up areas	S-6.2	
		1st and 2nd choices are independent routes	S-6.3	O-7
		Analyzed for on-road and off-road	S-6.4	
	Recommended AoAs better than non-recommended AoAs		S-6.5	
Quality of BN Boundaries				
	BN boundaries give BNs room to maneuver		S-7.1	
	BN Boundaries follows natural geographic features		S-7.2	
	AoAs are within BN AOO		S-7.3	O-9
Plan Quality				
	Essential Information included		S-8.1	
	Information clearly presented		S-8.2	
		Information clearly labeled		
		Information uniquely identified		
	Plan meets all mission specifications		S-8.3	

Special care was taken when aggregating the objective and subjective data using a weighted average. For the weighted average, the process was complicated by the fact

that weights were only elicited for the *subjective questions*; weights were not elicited for *objective questions (measures)* nor were they elicited for decision quality *characteristics*. Since the subjective questions were grouped by function and not by characteristic, the weighting process was less straight forward that it could have been. Because the decision quality measures were grouped by planning considerations to ease evaluation by the SMEs, questions can addresses measures related to one or more decision quality characteristic. Therefore the weights for each subjective question may apply to measures of different decision quality sub-characteristics or characteristics. Conversely, not all the measures of a sub-characteristic are always covered by a single subjective question. There may be two questions that, in total, address all the measures associated with a sub-characteristic and therefore the weighting method must account for the different weights of each measure score. In this case the measure scores associated with each question may have different weights. The weights associated with different measure scores were taken into account when determining sub-characteristic and characteristic scores.

In order to apply weights defined by functional area to decision quality characteristics, the following procedure was used: First, if there were individual objective and subjective scores for the same measure, the scores were averaged. If there is a subjective score encompassing several measures, the sub-characteristic score is the weighted average using the number of measures covered by the subjective score. All measure scores are mass-weighted when determine the sub-characteristic scores.

Weights elicited from the SMEs are applied to sub-characteristic scores when calculating characteristic scores. When determining the overall decision quality score,

the average of any weights assigned to the sub-characteristics are used as the characteristic weights. This method seems to preserve the relative importance of the information as weighted by the SMEs.

The relationship of the objective and subjective measures to the sub-characteristic scores are shown in 42, above. The raw and aggregated data can be found in Appendix 4-8: Case Study One Data.

Conduct statistical analysis

Two sets of analyses were conducted on the data generated by the evaluation; (1) one set to assess the effectiveness of using the BTRA-BC GDSS on the quality of military planning decisions, and (2) one set to evaluate the effect of using Table the DQEM to evaluate the quality of the decision-makers' decisions. The first set of analyses is summarized below with more detailed description contained in (Powell et al., 2010).

Summary of GDSS evaluation analyses

Statistical analyses were run on both the unaggregated objective and subjective metrics. The primary analysis used to determine whether the system variable had an effect on the decision-makers' decision-making was a repeated-measures analysis of variance (repeated measures ANOVA) of the decision-makers' responses. Because the small sample size, 18, normal-probability plots were used to determine whether the data is near normal. The normal probability plots suggested that the data are near normal, and a repeated-measures ANOVA was used to determine whether the decision-makers' average response when using a GDSS is significantly different from their average

response when not using the system. The results of the ANOVA for each subjective measure and the averages are presented in the results section. Even though the data appeared to be near normal, the validity of the ANOVA results were verified using a Wilcoxon Signed-Rank Test which supported the results of the ANOVA.

As the System variable is within-subject, the repeated-measures ANOVA should be able to determine statistical significance for a smaller main effect than for an effect due to the between-subjects variables, System Order and Scenario Order. Using an ANOVA also provided evidence of any existing interaction among System, System Order, and Scenario Order variables. Since the average data on which the ANOVA is run is composed of data from individual measures which are of interest in evaluating the decision-making of the decision-makers, repeated-measures ANOVAs and correlations of the decision-makers' responses were used to determine whether the individual characteristics had significant influences on the decision-makers' overall decision-making.

In addition to the ANOVA, other tests were also conducted to determine the validity of the evaluation hypotheses. For instance, some objective data, such as the binary data derived from the objective scoring, could not be treated as normal. For these data, a chi-square test was used to perform the roughly the same function on the binary data as a t-test does on normal data. A chi-squared test was used to determine whether the responses from the With and Without trials came from different distributions. Equal variance tests were used to determine whether the average variation in the data is smaller

and whether the decision-makers' responses were more consistent when using the GDSS under evaluation than when it is not used.

Summary of DQEM assessment analyses

The goal of CS-1 is to determine whether the use of the DQEM improved the ability of independent evaluators to evaluate the quality of the decisions made by decision-makers. In addition to the statistical analyses, e.g. ANOVA, used to determine if differences in decision quality exist, direct analysis of the effect the DQEM centers on two factors, the internal consistency of each SMEs evaluations and the agreement between their evaluations. SMEs' evaluations that are more internally consistent and that are in better agreement should have reduced variation in the individual scores which in turn should manifest itself as enhancing the ability to discriminate differences in decision quality.

In order to assess the ability of the DQEM to improve the evaluation of decision quality; the analysis addressed two hypotheses:

(1) Each SMEs' overall evaluations of decision quality will be more consistent with their evaluation of individual decision quality characteristics as the decomposition of decision quality becomes more detailed.

(2) The SMEs' evaluations of decision quality will agree more closely as the decomposition of decision quality becomes more detailed.

The hypotheses both refer to the decomposition of decision quality becoming more detailed as a basis for the analysis of the effectiveness of decision quality. The underlying premise of the DQEM is that decomposing decision quality will aid SMEs in

making consistent evaluations of decision quality and therefore be able to discriminate between the quality of decisions. This implied the need to assess decision quality in stages in which the decomposing of decision quality varied and directly resulted in the three scorings described previously.

In support of the two hypotheses, the analysis of the scoring of decisions quality characteristics focused on (1) the correlations between the SMEs' subjective overall evaluations and the average of their scores for individual decision quality characteristic measures (within-SME correlations) and (2) the correlations between both the SMEs' subjective overall evaluations and the averaged characteristic scores (between-SME correlations). The analysis also explores the effect of applying subjective weights elicited from the SMEs when aggregating objective and subjective scores.

The within-SME correlations were chosen as a measure of consistency of the SMEs ability to judge the overall decision quality of a series of decisions made in response to a complex and ill-structured problem. *Internally consistent*, in the case of assessing quality, means that the evaluators' overall evaluation of decision quality is based on their evaluations of each of the characteristics that support decision quality. Internal consistency of the evaluators' evaluations is not in itself essential to the assessing the ability of the DQEM to improve the evaluation of decision quality since the ability to determine relative overall decision quality is better reflected in the agreement between the SMEs evaluations (between-SME correlations). But, since the DQEM focuses on decomposing decision quality into characteristics and sub-characteristics, the evaluation of those decision quality sub-characteristics should support the evaluation of the decision

quality characteristics that in turn should support the evaluation of overall decision quality. If the evaluations of decision quality characteristics support the overall evaluation of decision quality, then that would indicate that the functional decomposition is appropriate to the evaluation of the decision quality of that particular problem. Further, an appropriate decomposition should be a useful guide in the evaluation of decision quality and should help this evaluation be more consistent; and if all the SMEs' evaluations are consistent, the agreement between the SMEs' evaluations (between-SME correlations) should be better.

The between-SME correlations were chosen as a measure of the reliability of the SMEs' assessments of decision quality in a complex and ill-structured problem. In order to assess the ability of the DQEM to improve the evaluation of decision quality, data was generated to address the hypothesis that the SMEs' evaluations of decision quality will agree more closely as the decomposition of decision quality becomes more detailed. The correlations between the SMEs' scores (between-SME correlations) were a measure of the level of agreement and were used as an indication of the reliability of the SMEs' evaluations. Successively higher between-SME correlations would indicate that there was more agreement between the SME's independent assessments and indicate that the SMEs' evaluations were converging on an unknown value that estimated the true quality of decisions. Even if the SMEs' evaluations converge on a specific value or are in perfect agreement, this does not indicate that their evaluations are absolutely accurate. Their evaluation would be reliable, i.e. the SMEs would agree on a specific score for the overall decision quality; but this score may differ from the "true" decision quality by

some bias. But, assuming that each SME's bias is relatively constant, more reliable evaluations should permit the discrimination of the relative quality of decisions.

Any bias inherent in the SMEs' scores should have little effect on the usefulness of between-SME correlations in analyzing the effect of using the DQEM. With respect to the between-SME correlations, the SMEs' biases could cause the correlations *not* to converge toward a single estimate of decision quality. SMEs' scores that had different constant biases would mean that even if each SME's estimates of decision quality were perfectly precise, their actual scores quantifying decision quality may not agree. This difference would be evident in the SMEs' actual scores but would not manifest itself in the between-SME correlations, adding a constant bias to each SMEs' score would not change the between-SME correlations. A perfect correlation would indicate that the SMEs' evaluations of the *relative decision quality* were in prefect agreement even if the actual decision quality scores were not the same. However, in a less than perfect world, the SMEs' biases are likely not to be constant across their evaluations of all the decision quality characteristics of multiple decision-makers; and the variation in any bias would serve to introduce variation in their scores and would reduce the between-SME correlations. Yet a substantial increase in between-SME correlations as the detail in the decomposition of decision quality increases would indicate that the SMEs' evaluations are converging, becoming more precise, as the level of detail increases. This increase in the reliability of the SMEs' evaluations and the resultant reduction in the variation of those scores should improve the ability of appropriate statistical analyses ability to discriminate differences in those scores.

Because the analyses of the internal and between-SME correlations were done in three successive stages, the effect of multiple comparisons musts be addressed when analyzing the significance of the correlations. When doing multiple comparisons among statistical tests, including correlations, the increased possibility of Type I errors must be considered; the more samples that are compared the more likely it is that at least one of the correlations will be high due to random variation in the data. Of the numerous methods for correcting significance levels, the Bonferroni correction is the most conservative (Abdi, 2007; Dunn, 1961). Using a Bonferroni calculator (Uitenbroek, 1977) to determine the individual significance level which corresponds to the combined level of significance of $p = 0.05$ for three samples, the significance level should be $p = 0.01695$ for each individual comparison (given an unknown correlation). Using a $p = 0.05$ as the threshold, each comparison would yield an overall probability of a Type 1 error of 0.15. In this case study, significance levels were assessed relative to the Bonferroni correction for the correlations compared between the three scoring, i.e. the correlations for which p-values were generated to compare the changes over multiple scorings. For dependent variables which were not compared over multiple scorings, any p-values generated were assessed as above the $p = 0.05$ level or below the Bonferroni corrected level of 0.01695. P-values that fell between $p = 0.05$ and 0.01695 were reported as such.

APPENDIX 4-1.2: DETAILS OF THE RESULTS FROM CASE STUDY ONE

As discussed in Chapter Four, CS-1 provided data for both the evaluation of the value–added when using BTRA-BC for military decision-making and for the assessment of the usefulness of using the DQEM to evaluate decision quality. A summary of the results of the evaluation of BTRA-BC and the results of the assessment of DQEM are presented below.

Summary of results of BTRA-BC evaluation

The evaluation of the value of the BTRA-BC GDSS centered on five aspects of plan quality: (1) time to completion, (2) objective plan quality, (3) subjective plan quality, (4) understanding of the terrain, and (5) the decision-maker perception of the GDSS. The first two were evaluated objectively, the second two were evaluated subjectively by SMEs, and the last was a subjective evaluation by the decision-makers. This section summarizes these results. A detailed discussion of these results stressing the improvement in decision quality due to using the BTRA-BC GDSS can be found in (Powell et al., 2010). The data gathered during the assessment and the results of the assessment are detailed in Appendices Appendix 4-8. Although both weighted and simple averages of the SMEs' evaluation scores are investigated in the section discussing the DQEM results, simple averages of the third scoring decision characteristic scores are

used in the evaluation of the impact of the GDSS. The results of the evaluation were as follows:

Time to Completion – A repeated-measures analysis of variance (ANOVA) indicated that decision-makers' average time to completion when they used DTSS with BTRA-BC ($\bar{x}$ = 1.140, s = 0.231) was significantly faster ($p < 0.001$) than when they used DTSS without BTRA-BC ($\bar{x}$ = 3.120, s = 0.890). On average, decision-makers completed the tasks using DTSS with BTRA-BC 64% faster than without BTRA-BC. An F-test ($p < 0.0001$) for unequal variances indicated that the variance in time to completion when the decision-makers used DTSS with BTRA-BC ($s^2 = 0.053$) was significantly lower than when they used DTSS without BTRA-BC ($s^2 = 0.793$).

Objective Quality – A repeated-measures ANOVA, confirmed by a Wilcoxon Signed Ranks test, indicated strong statistical evidence ($p < 0.001$) that decision-makers' average objective quality score when they used DTSS with BTRA-BC ($\bar{x}$ = 3.850, s = 0.626) was significantly higher than when they used DTSS without BTRA-BC ($\bar{x}$ = 2.920, s = 0.609). An equal variance F-test ($p = 0.440$) did not indicate a significant difference in variance between the objective quality scores for DTSS with BTRA-BC ($s^2 = 0.392$) and DTSS without BTRA-BC ($s^2 = 0.371$).

Subjective Quality – A repeated-measures ANOVA provided strong statistical evidence ($p = 0.003$) that decision-makers' average subjective quality score when they used DTSS with BTRA-BC ($\bar{x}$ = 3.400, s = 0.425) was significantly higher than when they used DTSS without BTRA-BC ($\bar{x}$ = 2.720, s = 0.749. An equal variance F-test (p =

0.012) indicated a significant difference in variance between the subjective quality scores for DTSS with BTRA-BC ($s^2 = 0.180$) and DTSS without BTRA-BC ($s^2 = 0.561$).

Terrain Understanding – A repeated-measures ANOVA provided some support ($p = 0.059$) that knowledge and understanding of terrain was greater when decision-makers used DTSS with BTRA-BC ($\bar{x} = 3.185$, $s = 0.861$) than without BTRA-BC ($\bar{x} = 2.565$, $s = 0.950$). An F-test ($p = 0.345$) did not indicate a significant difference in variance between the terrain understanding scores for DTSS with BTRA-BC ($s^2 = 0.741$) than without BTRA-BC ($s^2 = 0.902$).

The conclusions that can be drawn from these results are discussed with the results of the assessment of using the DEQM in the conclusion section.

Results of the assessment of using the DQEM

The three separate scorings of the subjective data permitted an in depth investigation of the effects of decomposing scoring criteria on SME scores and the effects of using weights to aggregate objective and subjective data. The three scorings correspond to the three levels of the decomposition of decision quality: the initial decomposition into decision quality characteristics and sub-characteristics, the definition of the measures and a decomposition iteration to support these measures, and tailoring Likert scales with scoring criteria specific to each measure. These three scorings provide data on the SMEs' ability to consistently evaluate decision quality with respect to a progressively more detailed consensus on the concepts that constitute decision quality.

The data in this case study show interesting relationships between the SMEs subjective evaluations and the level of detail in the scoring criteria.

First Scoring – with decision sub-characteristics

The initial scoring was conducted with ten plans: five DTSS plans and five BTRA-BC plans. The SMEs were provided with the decision characteristics and sub-characteristics, but a discussion as to what the factors would affect any measures had not yet taken place. Therefore, the SMEs were using only their experience to evaluate the initial 10 plans based on the decision quality sub-characteristics. The subjective questionnaire used to elicit scores from the SMEs can be found in Appendices 4-3; and the scores are presented in Appendix 4-4. Table 43 presents the within-SME correlations between each SME's subjective overall scores and both the simple and weighted averages of their sub-characteristic scores. These within-SME correlations are indicated by the row heading identifying the SME (SME1 & SME 2). Also included in Table 43 are the between-SME correlations between the SMEs' overall subjective, weighted average, and simple average scores. These between-SME correlations are identified by the External row heading. Table 44 also includes the internal and between-SME correlations for the subjective data alone (w/o objective data) and with the objective data aggregated with the subjective data (w/ objective data).

Table 43: CS-1 – Score correlations (1st scoring)

		Subjective Overall Score	Simple Average	Weighted Average	Change due to weighting
w/o objective data	SME 1	--	0.8958	0.9351	**0.0393**
	SME 2	--	0.8374	0.8402	**0.0028**
	External	-0.0976	0.4995	0.3459	-0.1537
w/ objective data	SME 1	--	0.5727	0.6328	**0.0600**
	SME 2	--	0.6811	0.6819	**0.0008**
	External	N/A	0.8605	0.7761	-0.0844
Change due to aggregating objective data	SME 1	--	-0.3231	-0.3023	
	SME 2	--	-0.1562	-0.1583	
	External	N/A	**0.3610**	**0.4302**	

Table 44: CS-1 – Correlation Significance (1st scoring)

Correlation Significance (1st scoring)

		Subjective Overall Score	Simple Average	Weighted Average	Significance of weighting
w/o objective data	SME 1	--	0.0799	0.4803	0.4803
	SME 2	--	0.1127	0.3115	0.3115
	External	0.4610	0.2916	0.4794	0.4794
w/ objective data	SME 1	--	0.2573	0.4965	0.4965
	SME 2	--	0.2029	0.4052	0.4052
	External	N/A	0.0976	0.4102	0.4102
Significance of aggregation	SME 1	--	0.0675	***0.0375***	
	SME 2	--	0.2378	0.2333	
	External	N/A	0.0812	0.1034	

Table 44 presents the statistical significance of (1) the individual correlations in the same relative positions in both tables, (2) the changes in the correlations due to using weighted averages in the right-most column, and (3) and the changes in the correlations due to aggregating the objective data are presented in the three bottom-most rows. The

formats of Tables 43 and 44 are used consistently throughout this chapter and Chapter Five.

Overall, these correlations, for the least decomposed scoring criteria, seem to be consistently affected by aggregating objective data and by using weighted averaging. Beginning with the within-SME correlations for each SME, the correlations between the aggregated individual scores and the SMEs' subjective overall scores vary from 0.5727 (SME 1 w/ objective data) to 0.9351 (SME 1 w/o objective data) with all but two approaching or above 0.7 which is the general threshold indicating some level or linkage between the distributions. The within-SME correlations for the subjective scores alone (w/o objective data) were consistently higher than for the aggregated subjective and objective scores. Also, all the within-SME correlations for the subjective scores alone were above 0.7, and none of the correlations for the aggregated data were above this threshold. Individually, each within-SME correlation decreased when the objective scores were aggregated. Unlike this negative change in the within-SME correlations due to aggregating the objective data, the changes due weighting of the averages was positive but minor. The weighting of the average scores increased the within-SME correlations for SME 1 but had a negligible (but positive) effect on the correlations for SME 2.

Significance levels for individual correlations indicate the probability that the individual samples come from a population in which the population correlation is 0. Because the significance of the individual correlations did not test multiple correlations, a $p = 0.05$ threshold was used. None of the individual correlations approaches this threshold; and therefore, the hypothesis that the data comes from a population with a

correlation other than 0 could not be supported. For the comparison of two correlations, e.g. analyzing the change in correlations, the significance level indicates the probability that the samples come from populations with different correlations. Since the tests for change in the correlations compare two samples from the same scoring iteration, the significant level of $p = 0.05$ was used; and with the exception of aggregating the objective data into SME 1's weighted average score, none of the changes due to using a weighted average of aggregating the objective data were significant and indicated that the hypothesis that the correlations come from populations with different correlations could not be supported.

Unexpectedly, the behavior of between-SME correlations was the opposite of that of the within-SME correlations. First, there appears to be little correlation between the SMEs' subjective overall scores (-0.0976) whereas the within-SME correlations were strongly positive. Second, the correlations between the SMEs' averaged characteristic scores increased when the subjective scores were aggregated with the objective scores instead of decreasing. Third, the between-SME correlations decreased, not increased, when the averages of the characteristic scores were weighted. The analysis of these seemingly conflicting results will be discussed in the conclusions section below.

Second Scoring – with measure consensus

For this scoring, the SMEs had reached consensus on the scoring measures and the factors affecting the subjective scoring of those measures. The SMEs have not yet reached consensus on the association of performance levels associated with each of those factors to Likert scale values. In this scoring, essentially, the decision-quality

characteristics had been decomposed into sub-characteristics, the detailed measures had been developed, but the SMEs' had not reached consensus on how the to evaluate the measures.

Even though the specifics of the evaluation of each measure had not been determined, the following overall changes in the correlations were expected due to the increased decomposition of the decision quality characteristics:

- Within-SME correlations would improve because of the following:
    a. The decomposition of decision quality would better defined the sub-characteristics and focus the SMEs' assessments of the decision characteristics. The guidance provided by the more detailed decomposition and better assessments of each characteristic should increase the agreement between each SME's overall subjective score and the aggregated average scores resulting in increased within-SME correlations.
    b. Since the objective data should reflect the actual decision quality, aggregating this data into the simple and weighted averages should cause these averages to be better estimators of the actual decision quality. The change in the within-SME correlations will provide evidence as to the relative accuracy of the SMEs' subjective overall evaluations and the averaged characteristic scores.
- Between-SME correlations between the SMEs' scores would improve for reasons similar to those given for the within-SME correlations:

- The decomposition of decision quality and the resultant focusing of the SMEs' assessments would reduce the variation in their assessments and result in higher correlations between the SMEs' scores.
- Increasing the detail in the decomposition of decision quality should further reduce the variation in their assessments resulting in the SMEs' scores converging and manifesting as higher between-SME correlations.

The internal and between-SME correlations generated from the second scoring are presented in Table 45, and the significance of the internal and between-SME correlations can be found in Table 48. The changes in these correlations between first scoring and second scorings can be found in Table 49, and the significances of these changes are presented in Table 50.

Table 45: CS-1 – Score Correlations (2nd scoring)

		Subjective Overall Score	Simple Average	Weighted Average	Change due to Weighting
W/O objective data	SME 1	--	0.8225	0.8279	**0.0054**
	SME 2	--	0.8510	0.7973	-0.0537
	External	0.5068	0.6674	0.6771	**0.0097**
W/ objective data	SME 1	--	0.6573	0.6556	-0.0018
	SME 2	--	0.6489	0.5987	-0.0502
	External	N/A	0.9068	0.9197	**0.1300**
Change due to aggregating objective data	SME 1	--	-0.1652	-0.1724	
	SME 2	--	-0.2020	-0.1986	
	External	N/A	**0.2393**	**0.2427**	

From the data in Table 45, it is evident that the behavior of the within-SME correlations in the second scoring was both similar to and different from those in the first

scoring. The correlations for the simple and weighted average scores both with and without the aggregation of objective data were strongly positive. The correlations for the subjective scores alone were again higher than the aggregated scores, ranging from 0.7973 to 0.8510, while the aggregated scores all remained below 0.7 (0.5987 to 0.6573). In the second scoring however, the change in SME 2's averaged scores was greater than SME 1's. Similar to the first scoring, the change in within-SME correlations due to weighting the SMEs' scores was negligible although in this scoring generally negative. Like the first scoring, weighting the averages had a greater effect on SME 2's within-SME correlations than on SME 1's; but in this scoring, the change in SME 2's within-SME correlations were negative (-0.0537 and -0.0502) when the averages were weighted. The changes in SME 1's within-SME correlations were again negligible (0.0054 and -0.0018).

Table 46: CS-1 – Correlation Significance (2nd scoring)

		Subjective Overall Score	Simple Average	Weighted Average	Significance of weighting
W/O objective data	SME 1	--	0.1221	0.1187	0.4803
	SME 2	--	0.1039	0.1376	0.3115
	External	0.2883	0.2101	0.2051	0.4794
W/ objective data	SME 1	--	0.3127	0.2162	0.4965
	SME 2	--	0.2196	0.2448	0.4052
	External	N/A	0.0656	0.0562	0.4102
Significance of aggregation	SME 1	--	0.1362	0.1239	
	SME 2	--	0.0781	0.1217	
	External	N/A	***0.0202***	***0.0130***	

The significance of the individual within-SME correlations and the changes in the within-SME correlations due to aggregating in the objective data and using weighted averages were similar to the first scoring. None of the individual correlations reached the $p = 0.05$ threshold; and therefore, the hypothesis that the data comes from a population with a correlation other than 0 again could not be supported. When testing for the significance of the changes in correlations when weighted averages were used or the objective data aggregated in, none of the changes in correlation were significant at the $p = 0.05$ level.

Unlike the within-SME correlations, the between-SME correlations showed consistently positive changes. The between-SME correlation between the SMEs' overall subjective scores was 0.5506. This correlation is still below but approaching the generally accepted threshold of 0.7. In contrast to the within-SME correlations, the between-SME correlations increased when the average characteristic scores were weighted. Weighting the averages increased the correlations for the With and Without objective data cases by 0.1633 and 0.0974 respectively. In stark contrast to the within-SME correlations, the between-SME correlations for both the simple and weighted averages were higher when the objective scores were aggregated (changes of 0.2393 and 0.2427 respectively) than when the subjective scores alone were used.

Table 47: CS-1 – Change in Score Correlations (1st to 2nd scoring)

		Subjective Overall Score	Simple Average	Weighted Average
W/O objective data	SME 1	--	-0.0733	-0.1072
	SME 2	--	**0.0136**	-0.0430
	External	**0.6044**	**0.1679**	**0.3312**
W/ objective data	SME 1	--	**0.0846**	**0.0228**
	SME 2	--	-0.0322	-0.0832
	External	N/A	**0.0463**	**0.1437**

The significance (Table 50) of the individual between-SME correlations and the changes in the within-SME correlations due to aggregating in the objective data and using weighted averages were similar to the first scoring. Although the significance of the between-SME correlations when the objective data was aggregated approached 0.05, none of the individual between-SME correlations reached the $p = 0.05$ threshold; and therefore, the hypothesis that the data comes from a population with a correlation other than 0 again could not be supported. Like the changes in within-SME correlations, the significance of the changes in the between-SME correlations when using a weighted average was not significant (0.0.4794 and 0.4102). But, unlike the within-SME correlations, the changes in the between-SME correlations due to aggregating the objective data were significant when using both simple and weighted averages (0.0202 and 0.0130) indicating that aggregating in the objective data improved the agreement between the SMEs' averaged decision quality characteristic scores.

Table 48: CS-1 – Correlation Significance (2nd scoring)

		Subjective Overall Score	Simple Average	Weighted Average	Significance of weighting
w/o objective data	SME 1	--	0.1221	0.1187	0.4803
	SME 2	--	0.1039	0.1376	0.3115
	External	0.2883	0.2101	0.2051	0.4794
w/ objective data	SME 1	--	0.3127	0.2162	0.4965
	SME 2	--	0.2196	0.2448	0.4052
	External	N/A	0.0656	0.0562	0.4102
Significance of aggregation	SME 1	--	0.1362	0.1239	
	SME 2	--	0.0781	0.1217	
	External	N/A	***0.0202***	***0.0130***	

The significance of the individual within-SME correlations and the changes in the within-SME correlations due to aggregating in the objective data and using weighted averages were similar to the first scoring. None of the individual correlations came near the $p = 0.05$ threshold, and therefore the hypothesis that the data comes from a population with a correlation other than 0 again could not be supported. When testing for the significance of the changes in correlations when weight averages were used or the objective data aggregated in, none of the changes in correlation were significant at the $p = 0.05$ level.

Unlike the within-SME correlations, the between-SME correlations showed consistently positive changes. The between-SME correlation between the SMEs' overall subjective scores was 0.5506. This correlation is still below but approaching the generally accepted threshold of 0.7. In contrast to the within-SME correlations, the between-SME correlations increased when the average characteristic scores were weighted. Weighting the averages increased the correlations for the With and Without

objective data cases by 0.1633 and 0.0974 respectively. In stark contrast to the within-SME correlations, the between-SME correlations for both the simple and weighted averages were higher when the objective scores were aggregated (changes of 0.2393 and 0.2427 respectively) than when the subjective scores alone were used.

Table 49: CS-1 – Change in Score Correlations (1st to 2nd scoring)

		Subjective Overall Score	Simple Average	Weighted Average
W/O objective data	SME 1	--	-0.0733	-0.1072
	SME 2	--	**0.0136**	-0.0430
	External	**0.6044**	**0.1679**	**0.3312**
W/ objective data	SME 1	--	**0.0846**	**0.0228**
	SME 2	--	-0.0322	-0.0832
	External	N/A	**0.0463**	**0.1437**

The significance (Table 50) of the individual between-SME correlations and the changes in the within-SME correlations due to aggregating in the objective data and using weighted averages were similar to the first scoring. Although the significance of the between-SME correlations when the objective data was aggregated approached 0.05, none of the individual between-SME correlations reached the $p = 0.05$ threshold; and therefore, the hypothesis that the data comes from a population with a correlation other than 0 again could not be supported. Like the changes in within-SME correlations, the significance of the changes in the between-SME correlations when using a weighted average was not significant (0.0.4794 and 0.4102). But, unlike the within-SME correlations, the changes in the between-SME correlations due to aggregating the objective data were significant when using both simple and weighted averages (0.0202

and 0.0130) indicating that aggregating in the objective data improved the agreement between the SMEs' averaged decision quality characteristic scores.

Table 50: CS-1 – Significance of Changes in Correlations (1st to 2nd scoring)

		Subjective Overall Score	Simple Average	Weighted Average
W/O objective data	SME 1	--	0.2621	0.1253
	SME 2	--	0.4580	0.3855
	External	***0.0126***	0.2833	0.1513
W/ objective data	SME 1	--	0.3806	0.4654
	SME 2	--	0.4488	0.3764
	External	N/A	0.3169	0.1097

The changes in the correlations between the first and second scorings are shown in Table 49, and the significances of those changes are presented in Table 50. The changes in the within-SME correlations were inconsistent and generally small ranging from 0.1072 to 0.846. Since the changes in these correlations considered here are between scoring, the Bonferonni correction indicates that a significance of 0.1695 should be used. The changes in the within-SME correlations were not significant ranging from 0.1253 to 0.4645. Likewise the changes in the between-SME correlations, with the exception of the SMEs' subjective overall scores, were also not significant ranging from 0.1097 to 0.2833. In contrast to the changes in the other correlations, the between-SME correlation between the SMEs' overall scores increased by 0.6044 resulting in a correlation of 0.5506 and was significant at 0.0102. The conclusions that can be drawn

from the changes in the correlations and their significance, or lack of significance, will be discussed in the conclusion section.

Third Scoring – with criteria consensus

This third and final scoring was conducted with decision quality decomposed down to the decision quality measure level, and the SMEs had reached consensus on the level of accomplishment that was required in each factor affecting each decision quality measure to achieve each value on the Likert scale. The expected changes in the correlations from the second to third scorings were the same as those discussed for previous scoring. Table 51 presents the correlation data from the final scoring, and Table 31 presents the significances of the individual internal and between-SME correlations and the changes in those correlations due to using weighted averages and aggregating the objective data.

Table 51: CS-1 – Score Correlations (3rd Scoring)

		Subjective Overall Score	Simple Average	Weighted Average	Change due to weighting
W/O objective data	SME 1	--	**0.8593**	**0.8755**	**0.0162**
	SME 2	--	**0.8587**	**0.8471**	-0.0116
	External	**0.7198**	**0.9202**	**0.9188**	-0.0014
W/ objective data	SME 1	--	**0.7812**	**0.7922**	**0.0111**
	SME 2	--	**0.8175**	**0.8085**	-0.0090
	External	N/A	**0.9677**	**0.9672**	**0.0085**
Change due to aggregating objective data	SME 1	--	-0.0781	-0.0832	
	SME 2	--	-0.0412	-0.0386	
	External	N/A	**0.0475**	**0.0574**	

The within-SME correlations in the third scoring were consistent with those in the previous two scorings. The correlations between the aggregated individual scores and the SME's subjective overall scores were still high, varying from 0.7812 to 0.8755; but in this scoring, all the within-SME correlations were greater than 0.7. Like the previous scorings, the correlations generated when the objective scores were aggregated (ranging from 0.7812 to0.8175) were all lower than the corresponding non-aggregated correlations (ranging from 0.8471 to 0.8755). In contrast to the previous scorings, the weighting of the averages had little effect on the within-SME correlations with the overall subjective scores. The changes within-SME correlations due to weighting the characteristic scores ranged from -0.0116 to 0.0162 with SME 1 generating positive changes and SME 2 generating negative changes in the within-SME correlations. Like the previous scoring, the change in the within-SME correlations due to aggregation of the objective scores with the subjective scores caused all these correlations to decrease (ranging from -0.0386 to -0.0832).

Like the within-SME correlations, the between-SME correlations in the third scoring, including the correlation of the SMEs' subjective overall scores, were all greater than 0.7. Noteworthy, the change in the between-SME correlation between the SMEs' overall subjective scores continued to increase, in this scoring by 0.2130, resulting in a correlation of 0.7198. In this scoring, the between-SME correlations (ranging from 0.9188 to 0.9677) were all greater than the within-SME correlations which ranged from 0.7812 to 0.8755. Like the within-SME correlations the changes in the between-SME correlations due to using weighted averages and aggregating the objective data were

small, ranging from -0.0014 to .0574, but generally positive. Unlike the within-SME correlations, the between-SME correlation exhibited modest but positive changes (0.0475, 0.0574) when the objective data was aggregated.

Table 52: CS-1 – Correlation Significance (3rd scoring)

		Subjective Overall Score	Simple Average	Weighted Average	Significance of weighting
W/O objective data	SME 1	--	0.0984	0.0875	0.4243
	SME 2	--	0.0997	0.1064	0.4506
	External	0.1822	0.0559	0.0569	0.4895
W/ objective data	SME 1	--	0.1472	0.1407	0.4665
	SME 2	--	0.1252	0.1308	0.4691
	External	N/A	***0.0200***	***0.0136***	0.3258
Significance of aggregation	SME 1	--	0.2400	0.2081	
	SME 2	--	0.3425	0.3598	
	External	N/A	0.0879	***0.0335***	

As seen in Table 52, in the third scoring, the significances of all the individual correlations decreased between 0.0042 and 0.1542 resulting in p-values ranging from 0.0549 to 0.1822. Although the p-values for all the individual correlations were lower than in the second scoring, only the two between-SME correlations associated with aggregating the objective data were significant at the 0.5 level (bolded Table 48) indicating the likelihood that the individual correlations came from populations where the correlation was not 0. Like the previous scorings, the changes in the individual correlations due to using weighted averages and the changes in within-SME correlations due to aggregating the objective data were not significant. The change in the between-SME correlations when aggregating the objective data was only significant when using a

weighted average (bolded in Table 52) and not when using a simple average of the decision quality characteristic scores.

Table 53: CS-1 – Change in Score Correlations (2nd to 3rd scoring)

		Subjective Overall Score	Simple Average	Weighted Average
W/O objective data	SME 1	--	0.0368	0.0475
	SME 2	--	0.0077	0.0499
	External	0.2130	0.2528	0.2417
W/ objective data	SME 1	--	0.1239	0.1367
	SME 2	--	0.1686	0.2098
	External	N/A	0.0609	0.0565

Table 54: CS-1 – Significance of Changes in Correlations (2nd to 3rd scoring)

		Subjective Overall Score	Simple Average	Weighted Average
W/O objective data	SME 1	--	0.3565	0.3052
	SME 2	--	0.4668	0.3261
	External	0.1562	**0.0222**	**0.0271**
W/ objective data	SME 1	--	0.2239	0.1971
	SME 2	--	0.1366	0.1042
	External	N/A	0.0559	**0.0348**

The changes in the internal and between-SME correlations over all three scorings are presented in Table 54 and the associated significances in Table 55. Two conclusions can be drawn from the changes in the within-SME correlations from the first to the third scoring (Table 51). First, all the within-SME correlations increased except for the correlations for SME 1 when only subjective scores were considered. The within-SME correlations for the simple and weighted averages of these scores from SME1, both exhibited small negative changes, -0.0365 and -0.0597 respectively. These negative scores seem to be the result of the high correlations in the first scoring and the subsequent negative change between the first and second scoring overshadowing the positive change between the second and third scoring. Second, the changes in the within-SME correlations for the simple averages were more positive than those for the weighted averages. This is consistent with the changes in these correlations between the first sand second scoring, but not with the nearly equal changes between these correlation between the second and third scorings. Even given the disparity in the changes in the within-SME correlations between the first and second scorings and second and third scorings, the p-

values associated with the overall changes in the within-SME correlations do not approach significance at either the p = 0.05 or p = 0.01695 levels.

Table 55: CS-1 – Significance of changes in correlations (1st to 3rd scoring)

		Subjective Overall Score	Simple Average	Weighted Average
w/o objective data	SME 1	--	0.3609	0.2236
	SME 2	--	0.4328	0.4787
	External	< 0.0001	***0.0102***	***0.0033***
w/ objective data	SME 1	--	0.1884	0.2304
	SME 2	--	0.2394	0.2592
	External	N/A	**0.0454**	***0.0045***

Overall, the between-SME correlations exhibit a consistently positive and generally larger change than the within-SME correlations between the first and third scorings. Most notably, the change in the between-SME correlation of the SMEs' subjective overall scores was 0.8174, of which 0.6044 was generated between the first and second scoring and 0.2130 was generated between the second and third scoring. This large change resulted in a p-value of <0.0001 indicating that the correlations in the third scoring were almost certainly higher than those from the first scoring. Unlike the within-SME correlations, the averaged subjective scores exhibited larger changes (0.4207, 0.5729) than the aggregated subjective and objective scores (0.1072, 0.2001). These large overall changes resulted in p-values of 0.0102 and .0033, both of which are significant at the p = 0.01695 level. Even though the between-SME correlations associated with aggregating the objective data did not exhibited the same magnitude of

change, the obtained p-values (0.0454, 0.0045) were significant at the p = 0.05 level; and the change in the weighted average correlations was significant at the 0.0165 level. Also unlike the within-SME correlations, the changes in the between-SME correlations of the weighted averages and simple averages were not consistently larger.

Analysis of the data from the three scorings

When analyzing the within-SME correlation data from the three scorings, several conclusions can be drawn:

- Individual within-SME correlations do not support the hypothesis that the population correlation is not 0.
- The within-SME correlations are high in all three scorings ranging from 0.5727 to 0.9851 with all but two of these correlations greater than 0.6 and only 16 of the remaining 22 were greater than 0.7.
- Even with such high correlations, the within-SME correlations for the subjective data alone are always greater than for the aggregated subjective and objective data in each scoring, but this change is generally not significant (Table 56).
- The within-SME correlations seem to exhibit be an inconsistent change between the first and second scorings and a positive change between the second and third scorings resulting in a positive overall change in the within-SME correlations.
- The overall magnitude of the change in the within-SME correlations is smaller, ranging from -0.0597 to 0.0213, for the correlation of the subjective scores alone than for the aggregated scores ranging from 0.1364 to 2.085.

- The correlations for the aggregated data consistently exhibit changes that are larger in magnitude, whether negative or positive, than the correlations for the subjective data alone.

Table 56: CS-1 –Significance of changes in correlations due to aggregating objective data

		Subjective Overall Score	Simple Average	Weighted Average
1st Scoring	SME 1	--	0.0675	***0.0375***
	SME 2	--	0.2378	0.2333
	External	N/A	0.0812	0.1034
2nd Scoring	SME 1	--	0.1362	0.1239
	SME 2	--	0.0781	0.1217
	External	N/A	***0.0202***	***0.0130***
3rd Scoring	SME 1	--	0.2400	0.2081
	SME 2	--	0.3425	0.3598
	External	N/A	0.0879	***0.0335***

- The only exception to the noted changes in the within-SME correlations are SME 1's simple and weighted averages for the subjective scores only; the exceptionally high within-SME correlations for SME 1's averaged scores (0.8958, 0.9351) in the first scoring seem to overwhelm the positive change between the second and third scoring an resulted in the only overall negative changes in the within-SME correlations (-0.0365, -0.0597).
- The changes in the within-SME correlations due to weighting the characteristic scores were small and inconsistent ranging from -0.0537 to 0.0393, and overall none of these changes are significant for any scoring (Table 57).

- Overall, the changes in within-SME correlations do not seem to be statistically significant (Table 54).

Table 57: CS-1 –Significance of changes in correlations due to weighting characteristic scores

		1st Scoring	2nd Scoring	3rd Scoring
w/o objective data	SME 1	0.4803	0.4803	0.4243
	SME 2	0.3115	0.3115	0.4506
	External	0.4794	0.4794	0.4895
w/ objective data	SME 1	0.4965	0.4965	0.4665
	SME 2	0.4052	0.4052	0.4691
	External	0.4102	0.4102	0.3258

Like the analysis of the within-SME correlations several conclusions can be drawn from the between-SME correlations:

- The p-values for the individual between-SME correlations for both the weighted and simply averaged scores in the third scoring support the hypothesis that the population correlation is not 0 (Table 58). Specifically, the individual correlations for the aggregated scores in the third scoring are significant at the p = 0.05 level (bolded and italicized in Table 58) and the p-values for individual correlations for the subjective scores alone approach significance (bolded in Table 58).
- The p-values for the between-SME correlations decrease through successive scorings (Table 54).

Table 58: CS-1 – Summary of significance of between-SME correlations

		Subjective Overall Score	Simple Average	Weighted Average
1st Scoring	Subjective data	0.4610	0.2916	0.3591
	Aggregated data		0.0976	0.1502
2nd Scoring	Subjective data	0.2883	0.2101	0.2051
	Aggregated data		**0.0656**	**0.0562**
3rd Scoring	Subjective data	0.1822	**0.0559**	**0.0569**
	Aggregated data		***0.0200***	***0.0136***

- The p-values of the correlations for the aggregated scores area consistently lower than for the subjective scores alone.
- The changes in the between-SME correlations are constantly positive through all three scorings.
- Overall (first to third scorings), the changes in the between-SME correlations are significant at the p = 0.01695 level for three of the four correlations (bolded and italicized in Table 59). Also the changes in three of the four between-SME correlations between the second and third scorings are significant the p = 0.05 level (bolded in Table 59).
- The p-values associated with all the changes in the between-SME correlations of averaged scores decrease with each successive scoring. (Table 56).
- The exceptionally low initial correlation of the SMEs' subjective overall scores (-0.0976) probably resulted in the significance of the change in this correlation between the first and second scorings.

Table 59: CS-1 – Summary of significance of change in between-SME correlations

Summary of Significance of Change in Between-SME correlations

		Subjective Overall Score	Simple Average	Weighted Average
1st to 2nd Scoring	Subjective data	***0.0126***	0.2833	0.1513
	Aggregated data		0.3169	0.1097
2nd to 3rd Scoring	Subjective data	0.1562	**0.0222**	**0.0271**
	Aggregated data		0.0559	**0.0348**
1st to 3rd Scoring	Subjective data	***< 0.0001***	***0.0102***	***0.0033***
	Aggregated data		**0.0454**	***0.0045***

- The significance of the changes in the between-SME correlations due aggregating the objective data were inconsistent (Table 59). Both the p-values from the second scoring supported significance at the p = 0.5 level as did the p-value for the correlation of the weighted average scores in the third scoring, but the other three p-values did not. There was consistent change in these p-values.
- The changes in the between-SME correlations due to weighting the characteristic scores were inconsistent ranging from -0.187944 to 0.1552, and overall none of these changes are significant for any scoring (Table 57).
- Overall, the 4 of 5 of the changes in the between-SME correlations are significant at the p = 0.01695 level and all the changes are significant at the 0.05 level.

The following section discusses the results presented here for both the evaluation of the BTRA-BC The GDSS and the assessment of the DQEM and conclusions are presented about the value of the BTRA-BC The GDSS to military decision-making and the usefulness of the DQEM in evaluating relative decision quality a part of a structured evaluation.

APPENDIX 4-2: CASE STUDY ONE OPERATION ORDER

SCENARIO #2BCT

MISSION:

Elements of the 2nd Brigade Combat Team consisting of the 1-6th and 2-6th Mechanized Infantry Battalions and the 1-35th Armored Battalion will advance from its current position in Assembly Area BOSTON northwest of Phase Line MIAMI to assault hostile units (mechanized infantry battalion augmented by a heavy armored company) in Engagement Area DIAMOND southeast of Phase Line PEARL in order to occupy said position.

ENEMY DISPOSITION:

Two to Three BN size enemy units are concentrated in Engagement Area DIAMOND. A number of light enemy militia forces are present between Phase Lines MIAMI and PEARL, and they will most probably be concentrated in the urban areas. These light militia forces are expected to defend their prepared positions, but they are not expected to leave the urban areas to attack our mechanized forces.

COMMANDER'S INTENT:

2BCT will advance in a 2 up / 1 back formation with 1-6th and 2-6th Mechanized Infantry Battalions forward and the 1-35th Armored Battalion as the reserve. 1-6th and 2-6th Mechanized Infantry Battalions will advance along two routes to arrive at their designated firing positions simultaneously. As the hostile force has had time to dig in, is of sufficient strength, and has the advantage of position, we must use surprise and misdirection to succeed in our mission. In that light, G Trp 1st CAV BRT will probe to the south to draw the attention of hostile forces while the main force advances covertly.

TASK ORGANIZATION:

2d BCT 1AD
HHC, 2d BCT 1AD

G Trp, 1st CAV BRT
1-6th Mech Bn
2-6th Mech Bn
1-35th Armd Bn
4-27th FA Bn
47th FSB

INSTRUCTIONS:

You are on the staff of the 2nd BCT of the 1AD. You are to conduct an analysis of the BCT area of operations for off-road as well as on-road movement from **Phase Line MIAMI to Phase Line PEARL**. The BCT CDR wants you to find **four independent avenues** of approach for BN sized units. You are to avoid built-up (urban) areas when generating these avenues of approach. The pipelines in our AOR are not to be considered obstacles as they are underground. The following products are required from your analysis:

1. **Combined Obstacle Overlay (COO**)
    a. Save with filename COO
2. **Identified Mobility Corridors (MC) categorized by type of force**
    a. Save as an annotation file with filename MC
3. **Four independent Avenues of Approach**
    a. Two routes are considered independent if they have no common MCs
    b. Save as an annotation file with filename AA1, AA2, AA3, or AA4
4. **Transit times required on each Avenue of Approach for three vehicles (M1 Abrams, M2 Bradley , and LAV 25)**
    a. Save in excel file AA#_TIME
5. **Choke Points in each Avenue of Approach**
    a. Save as an annotation file with filename AA#_CHKPT
6. **Recommended Areas of Operations (AO) boundaries for the two BNs in the BCT**
    a. Save as a new annotation file with filename BN

After you have completed the tasks above, you will be asked to evaluate the advantages and disadvantages of each avenue of approach.

APPENDIX 4-3: CASE STUDY ONE SAMPLE SME SUBJECTIVE EVALUATION QUESTIONNAIRE

1. Do AoAs take a direct route from phase line to phase line?

Avenues of Approach	meets less than 2 criteria	Meets 2 of 5 criteria	Meets criteria 1,2, & 4	Meets criteria 1,2, & 4,5	Direct - meet all 5 criteria
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

Criteria

1. Valid start point (behind PL and valid polygon)
2. Valid end Point (beyond PL and valid polygon)
3. No unnecessary turns
4. Routes are independent
5. Transit Times calculated

2. Are AoAs clearly indicated?

	Unable to distinguish AA route	AA route is either not obvious or not uniquely indicated and not labeled	AA route is obvious and uniquely indicated, but not labeled	AA route is labeled, but either not obvious or not uniquely indicated	AA route is labeled, obvious, and uniquely indicated
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

3. Are MCs sized and clearly indicated?

MC are neither sized nor clearly indicated	MCs are indicated, but not clearly	Some MCs are clearly indicated but not sized	Some MCs are sized and clearly indicated	All MCs are sized and clearly indicated
1	2	3	4	5

4. Are choke points clearly indicated and categorized by size?

	Does not meet any criteria	Meets criteria 1	Meets Criteria 1-2	Meets Criteria 1-3	Meet all four criteria
	1	2	3	4	5
AoA1					
AoA2					
AoA3					
AoA4					

Criteria

1. CPs on AoAs are clearly indicated
2 CPs on AoAs are categorized by size
3. CPs on AoAs are sized correctly
4. CPs not on AoAs are indicated and sized

5. Are the recommended AoAs clearly labeled?

	Recommended AoA is neither labeled nor uniquely indicated		Recommended AoA is Labeled, but not uniquely indicated		Recommended AoA is labeled and uniquely indicated
	1	2	3	4	5
1st AoA					
2nd AoA					

6. Did the recommended AoA meet mission specifications?

	Meets none of the criteria	Meets 2 of 5 criteria	Meets 3 of 5 criteria	Meets 4 of 5 criteria	Meets all 5 criteria
	1	2	3	4	5
1st					
2nd					

Criteria

1. Appropriate for BN-sized units
2. Avoid built up areas
3. 1st and 2nd are independent routes
4. Analyzed for on-road and off-road

7. Is the BN boundary appropriate?

Meets none of the criteria	Meets 1 of 4 criteria	Meets 2 of 4 criteria	Meets 3 of 4 criteria	Meets all 4 criteria
1	2	3	4	5

Criteria

1. BN boundaries give BNs room to maneuver
2. BN Boundary follows natural geographic features
3. AoAs are within BN AOO

8. Overall clarity and presentation of information:

Does not meet criterion 1		Meets criterion 1 and 1 of remaining 2		Breifable as is, meets all criteria
1	2	3	4	5

1. All essential information present for recommended AoAs
2. Information clearly labeled and uniquely identified
3. Plan meet all mission specifications

APPENDIX 4-4: CASE STUDY ONE DATA

Table 60: CS-1 – Objective data

Subject		Question									Average
		1	2	3	4	5	6	7	8	9	
BTRA	1	3.5	2.25	4.5	5	1	3	5	5	5	3.806
	2	2	1.5	4	4.5	1	4	5	4	1	3
	3	2.5	2	4.5	5	2	2	5	5	1	3.222
	4	4	1.75	5	5	1	2	1	5	5	3.306
	5	4	1.75	5	5	1	1	5	5	5	3.639
	6	2	1.5	3	4	1.5	2	5	3	1	2.556
	7	4.5	3.5	3	5	2	5	5	5	5	4.222
	8	2.5	2.25	5	5	2	3	5	5	5	3.861
	9	5	3.25	5	5	3	4	5	5	5	4.472
BTRA	11	5	5	5	5	5	5	5	5	1	4.556
	12	5	2.75	4.5	5	3	5	5	5	5	4.472
	13	4.5	3.75	4.5	5	3.5	4	5	5	5	4.472
	14	5	4.5	5	5	4.5	5	5	5	5	4.889
	16	4	2.25	4	4.5	1	3	5	5	5	3.75
	17	5	1.75	4.5	5	1	3	5	5	5	3.917
	18	4	2.5	4	5	1.5	3	5	5	5	3.889
	19	4	2	3	5	1	2	1	5	5	3.111
	20	4	2.25	4.5	5	1.5	5	5	5	5	4.139
DTSS	1	1	5	3	3	5	5	5	5	5	4.111
	2	1	1	1	1	1	1	5	1	5	1.889
	3	1	3.25	1	1	3.5	4	5	1	5	2.75
	4	1	5	1	1	5	5	1	1	5	2.778
	5	1	2	1	1	1	1	5	1	5	2
	6	1	2.25	1	1	2	5	5	4	5	2.917
	7	1	3.5	2.5	1.5	2	5	5	1	5	2.944
	8	1	2	1	2	2	2	5	1	5	2.333
	9	1	3.25	1.5	1	2.5	5	5	1	5	2.806
DTSS	11	1	3	2	3.5	2	3	1	5	5	2.833
	12	1	3	1.5	2	2	4	5	1	5	2.722
	13	1	2.75	1	2	2	5	5	5	5	3.194
	14	1	4	4	5	3	5	5	1	5	3.667
	16	1	2.75	1.5	1.5	2.5	3	5	1	5	2.583
	17	1	5	5	5	5	5	1	5	5	4.111
	18	1	5	2	2	5	5	5	1	5	3.444
	19	1	2	1.5	5	1.5	2	5	1	5	2.667
	20	1	2.75	3	3.5	2	2	5	1	5	2.806

Table 61: CS-1 – SME subjective weightings

SME Criteria Relative Weights								
SME	**Measure**							**Correlation**
	1	**2**	**3**	**4**	**5**	**6**	**7**	
1	8	6	7	7	6	10	5	0.8550
2	10	5	8	7	5	9	4	

Table 62: CS-1 – 1st Scoring (w/ Quality Sub-Characteristics)

		Sub-Characteristic							**Overall**	**Weighted**	
	#	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Average**	**Average**
SME 1											
BTRA	3	3.5	3	5	5	1	5	3	4	3.816	3.643
	9	2.75	5	5	5	3	5	5	5	4.388	4.393
	12	3.5	3	5	5	1	1	3	3	3.000	3.071
	16	2.25	3	5	5	1	4	4	5	3.510	3.464
	19	2.75	3	5	5	1	1	3	3	2.878	2.964
DTSS	4	3	5	1	5	1	1	3	3	2.592	2.714
	11	4	1	1	5	1	4	5	4	3.082	3.000
	12	4.75	1	1	5	1	5	5	4	3.408	3.250
	14	1.75	1	1	5	1	3	1	1	2.102	1.964
	19	5	5	1	3	1	3	3	3	3.041	3.000
SME 2											
BTRA	3	4	3	5	4	1	1	3	3	3.104	3.000
	9	4.25	2	5	4	1	2	2	3	3.156	2.893
	12	2	3	5	3	1	3	4	3	3.000	3.000
	16	4	3	5	3.75	3	4	1	3	3.672	3.393
	19	3.25	3	5	4.25	1	1	4	4	3.068	3.071
DTSS	4	3.75	2	1	1.75	1	2.5	2	3	2.151	2.000
	11	4.5	2	1	2.5	1	1	3	2	2.219	2.143
	12	3	1	1	2	1	1	2	1	1.646	1.571
	14	4	5	1	2	1	3	3	3	2.729	2.714
	19	3	2.25	1	1	1	1	1	1	1.547	1.464

Table 63: CS-1 – 2nd Scoring (w/ quality measure consensus)

	#	Sub-Characteristic							Overall	Weighted	
		1	2	3	4	5	6	7	8	Average	Average
SME 1											
BTRA	2	2.75	3	5	5	1	1	3	3	2.878	2.964
	4	2.5	3	5	5	1	1	1	4	2.633	2.643
	5	3.5	3	5	5	1	1	3	3	3.000	3.071
	6	2.75	5	5	5	3	5	5	5	4.388	4.393
	7	2.75	3	5	5	1	1	5	3	3.082	3.250
	11	3	3	5	5	1	5	5	4	3.939	3.857
	13	3.5	3	5	5	1	5	3	4	3.816	3.643
	14	2.5	5	5	5	1	1	3	5	3.082	3.214
	17	2.25	3	5	5	1	4	4	5	3.510	3.464
	20	3.25	5	5	5	1	5	3	4	4.020	3.893
DTSS	1	4	1	1	5	1	4	5	4	3.082	3.000
	3	2.75	3	1	5	1	5	3	3	3.122	2.964
	5	3	5	1	5	1	5	5	4	3.612	3.571
	6	4	5	5	5	1	1	5	4	3.531	3.714
	8	3	5	1	5	1	1	3	3	2.592	2.714
	9	4.75	1	1	5	1	5	5	4	3.408	3.250
	13	5	5	1	3	1	3	3	3	3.041	3.000
	16	3.75	3	1	3	1	1	5	2	2.388	2.536
	17	1.75	1	1	5	1	3	1	1	2.102	1.964
	20	1.75	3	1	1	1	3	3	1	1.980	1.964
SME 2											
BTRA	2	4	3	5	4	1	1	3	3	3.104	3.000
	4	3.5	3	5	3	2	3	2	3	3.250	3.071
	5	4.25	2	5	4	1	2	2	3	3.156	2.893
	6	2	3	5	3	1	3	4	3	3.000	3.000
	7	3.75	3	5	3.5	1	1	3	3	2.979	2.893
	11	4	3	5	3.75	1	3.5	2	3	3.453	3.179
	13	4	3	5	3.75	3	4	1	3	3.672	3.393
	14	4.5	5	5	2	1	3.5	3	3	3.594	3.429
	17	3.25	3	5	4.25	1	1	4	4	3.068	3.071
	20	4	5	5	5	1	1	3	4	3.458	3.429
DTSS	1	3.75	2	1	1.75	1	2.5	2	3	2.151	2.000
	3	4.5	3	1	1	1	1	1	1	1.938	1.786
	5	4	4	1	3	1	1	3	3	2.396	2.429
	6	4.5	5	1	4	1	3	4	4	3.208	3.214
	8	4.5	2	1	2.5	1	1	3	2	2.219	2.143
	9	3	1	1	2	1	1	2	1	1.646	1.571
	13	4	5	1	2	1	3	3	3	2.729	2.714
	16	4.75	2	1	3	1	4	5	4	3.073	2.964
	17	3	2.25	1	1	1	1	1	1	1.547	1.464
	20	2.5	3	1	1.5	1	1	2	1	1.677	1.714

Table 64: CS-1 – 3rd Scoring (with criteria consensus)

	#	Sub-Characteristic							Overall	Weighted Average	Average
		1	2	3	4	5	6	7	8		
SME 1											
BTRA	2	5	4	5	5	2	2	4	4	4.167	3.857
	4	4.5	4	5	4	4	3	2	4	3.944	3.786
	5	4.25	2	5	5	3	2	3	4	3.556	3.464
	6	2	3	5	4	3	4	4	4	3.333	3.571
	7	4.75	4	5	4.5	1	1	4	4	3.833	3.464
	11	3	2	5	3.75	2	2.5	2	3	2.889	2.893
	13	5	4	5	3.75	3	5	3	4	4.222	4.107
	14	5	5	5	2	1	3.5	2	4	3.722	3.357
	17	2.25	2	5	3.25	1	1	5	3	2.556	2.786
	20	5	5	5	5	2	2	2	4	4.278	3.714
DTSS	1	4.75	3	1	2.75	1	2.5	3	3	3.056	2.571
	3	3.5	2	1	2	1	1	1	2	2.056	1.643
	5	4	4	1	5	2	1	2	3	3.444	2.714
	6	4.5	5	1	5	1	2	5	3	4.000	3.357
	8	5	3	2	3.5	1	3	3	3	3.389	2.929
	9	2	1	2	2	1	2	2	3	1.778	1.714
	13	5	5	2	2	2	4	4	3	3.722	3.429
	16	3.75	1	2	3	1	5	5	3	2.889	2.964
	17	3	2.25	3	1	1	2	1	2	2.000	1.893
	20	2.5	3	1	2.5	3	2	3	3	2.556	2.429
SME 2											
BTRA	2	5	5	5	2	2	4.5	3	5	4.000	3.786
	4	5	5	5	5	2	1	3	4	4.222	3.714
	5	4.75	2	1	4	2	5	5	4	3.611	3.393
	6	3	4	4	4	2	4	4	3	3.611	3.571
	7	4.25	4	5	4.25	1	1	4	4	3.667	3.357
	11	4	5	1	3	1	4	3	3	3.556	3.000
	13	5	4	5	4.75	4	5	1	5	4.444	4.107
	14	4	3	5	3.75	2	3.5	2	4	3.500	3.321
	17	4	4	2	4	1	1	4	4	3.389	2.857
	20	3	4	4	4	2	4	4	3	3.611	3.571
DTSS	1	3.75	2	2	2.75	1	3.5	3	3	2.778	2.571
	3	3	1	1	2	2	1	3	2	1.889	1.857
	5	5	3	1	2.5	2	1	4	3	3.000	2.643
	6	4	3	5	5	2	1	3	4	3.556	3.286
	8	5	3	5	4	1	2	2	5	3.611	3.143
	9	3.5	2	1	2	1	2	2	2	2.222	1.929
	13	4.75	4	5	3.5	1	1	4	5	3.667	3.321
	16	3.5	3	5	4	2	4	1	4	3.444	3.214
	17	3.5	2	1	2	1	2	2	2	2.222	1.929
	20	3.5	4	2	1.5	2	2	3	2	2.722	2.571

APPENDIX 5-1: CASE STUDY TWO DECISION QUALITY DECOMPOSITION

Decision Quality Characteristics		Rationale
Quality of Choke Points		
	CP locations clearly sized and indicated	Clarity of GCMs
		Size accurately indicated
Quality of Mobility Corridors		
	MC locations clearly sized and indicated	Clarity of GCMs
		Size accurately indicated
Quality of potential Routes		
	Valid start point	Good route from AA to start point
	Valid end point	Good route from endpoint to objective
	Avoid Obstacles	Number of obstacles traversed (Integer)
	Avoid choke points	Number of choke point on route (Integer)
	Maintain Formation	Instances of suboptimal formation (Integer)
	Result in synchronized movement	Arrival time of platoons
	Stay within Boundaries	Planned routes are within operational boundaries (binary)
	Maintain Combat Power Forward	Analysis of generated transit timing, timing in the written plan, and unit missions in written plan
	Secure Flanks	Analysis of generated transit timing, timing in the written plan, and unit missions in written plan
	Accomplish commanders guidance	2 up 1 back
		Proper unit to proper objective
		Route to hide position
		Route to Battle position
	Use multiple movement corridors	Subjects assessment of maximum use of non-common MCs
	Use of concealment	Portion of route during which units concealed
Quality of Choke Points		
	CPs on AoAs are clearly	Clarity of GCMs

	indicated and sized		Size accurately indicated
Quality of NAIs			
	Written analysis		Placement near enemy positions
			Terrain analysis of probably enemy MCs
	Graphic analysis		Clarity of GCMs
			GCMs support analysis
Quality of Engagement Areas			
	Written analysis		Considered enemy weapon range
			Considered LOS
			Considered friendly weapon range
			Considered coverage of objective
	Graphic analysis		Clarity of GCMs
			GCMs support analysis
Quality of Battle Positions			
	Written analysis		Analysis considers LOS
			Analysis considers concealment
			Analysis considers Cover
			Analysis considers weapon range
	Graphic analysis		Clarity of GCMs
			GCMs support all analyses
	Quality of Hide Positions		
		Written analysis	Analysis considers Cover
			Analysis considers distance to BP
			Concealed approach route
		Graphic analysis	Clarity of GCMs
			GCMs support analysis
Quality of Ambush Positions			
	Quality of Egress route		Appropriateness of egress route
	Quality of Cover/Concealment		Analysis considers Cover/Concealment
	Quality of unit frontage		Analysis of friendly unit frontage compared to enemy unit size
	Quality of enemy concentration		Analyses of terrain's ability to concentrate enemy forces
	Quality of Rally Point		Appropriateness of recommended rally points
Quality of CO Boundaries			
	Quality of maneuver room		Sufficient area available for units to maneuver tactically

	Quality of C2 consideration		Lines of communication considered
Quality of Plan			
	Quality of graphic control measures		Ease of determining main effort
			Ease of determining support effort
	Plan Executabilty		Minimal additional guidance needed
	Evaluation of enemy operations		
		Enemy Timing	Possible enemy AoA considered
			Possible locations of enemy considered relative to friendly AoAs identified
		Impact on friendly movement	AoAs avoid most likely enemy positions
			Alternate routes identified
	Direction of mission essential tasks		
		Written analysis	All tasks directed
			Specific of tasks given
		Graphic analysis	Clarity of GCMs
			GCMs support tasks
	Graphics support of concept of Ops w/o being over prescriptive		Clarity of GCMs
			Plan easy to understand from GCMs
			GCMs Appropriate to tasks
Overall quality			
	Written plan		Subjective judgment
	Graphic plan		
	Quality of overall analysis		
	Feasibility of plan		
	Overall		

Figure 20: CS-2 − Decomposition

APPENDIX 5-2: CASE STUDY TWO DATA

The tables below contain the raw data of the SMEs for both scorings and the consensus scoring.

Table 65: CS-2 – 1st Scoring SME 1

		SME 1																																					
System	**ID**	1	2												3	4		5	6		7		8		9		10		11	12		13	14 - Ovearll					**DQCS**	**GDSP**
			a	b	c	d	e	f	g	h	i	j	k	l		a	b		a	b	a	b	a	b	a	b	a	b		a	b		a	b	c	d	**SOE**		
BTRA	AC	5	5	5	4	4	3	5	3	4	5	5	1	5	5	2	1	3	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	1	2	2	2.677	2.913
	AD	5	5	5	5	5	0	5	4	4	5	4	0	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4.581	4.435
	AE	5	5	3	5	3	5	5	3	3	5	1	1	5	1	5	5	1	5	4	5	5	3	5	1	1	5	5	4	4	4	1	4	5	3	5	5	3.645	3.696
	AH	5	5	3	5	3	4	5	0	2	5	1	2	5	0	3	3	5	4	4	1	1	5	5	5	5	5	4	4	5	5	4	3	3	3	3	3	3.645	3.739
	AM	5	5	5	5	5	3	5	5	5	5	5	5	0	2	3	4	4	2	2	0	0	5	5	1	0	5	5	4	4	4	3	4	4	4	5	4	3.581	3.913
	AS	5	5	3	5	4	5	5	5	5	5	4	3	5	5	5	5	5	4	4	5	5	3	3	1	1	5	4	5	5	5	4	5	5	5	5	5	4.290	4.174
	BF	5	1	5	1	4	4	5	4	4	2	5	3	5	4	3	4	4	4	4	0	1	5	5	1	1	5	3	5	5	5	4	5	5	4	5	5	3.581	3.478
	BP	5	5	5	5	5	1	5	3	3	5	1	5	5	1	0	0	5	5	4	0	0	4	5	0	0	5	0	3	4	4	3	3	3	4	4	4	3.097	3.130
CSE	AF	1	5	5	5	5	2	5	2	4	5	3	1	1	1	1	2	5	2	2	1	1	4	4	5	5	5	3	3	3	3	3	3	3	3	3	3	3.129	3.435
	AJ	5	5	4	5	5	5	5	3	3	5	5	2	1	3	5	5	5	4	4	1	1	3	3	4	5	1	1	4	4	4	3	4	4	4	4	4	3.645	3.739
	AK	3	5	3	5	5	4	5	5	5	5	3	1	1	5	5	5	5	3	3	1	5	3	3	1	1	5	4	5	5	5	4	5	5	5	5	5	3.806	3.783
	AL	1	5	3	5	5	5	5	4	4	5	1	4	1	2	2	4	5	3	3	1	1	5	5	5	5	5	5	5	5	5	3	5	5	5	5	5	3.774	3.913
	AV	1	1	1	3	3	1	5	3	3	3	5	1	1	1	1	3	3	1	1	1	1	1	1	1	1	1	1	0	1	3	1	4	1	1	4	2	1.742	1.957
	BB	3	4	1	5	4	4	5	5	5	5	5	1	1	5	3	3	5	4	3	1	1	4	4	4	4	1	1	4	3	3	4	4	4	4	4	4	3.387	3.391
	BE	3	5	1	5	2	4	5	4	4	5	4	3	2	5	3	3	4	4	4	1	1	3	3	1	1	5	5	4	3	3	3	3	3	3	3	3	3.323	3.304
	BR	3	5	1	5	4	1	1	4	4	5	5	4	1	5	5	5	3	4	5	1	0	3	0	1	0	5	5	4	3	4	3	4	3	3	4	3	3.194	3.043

Table 66: CS-2 – 1st Scoring SME 2

SME 2																																							
System	ID	1	2												3	4		5	6		7		8		9		10		11	12		13	14 - Ovearll					DQCS	GDSP
			a	b	c	d	e	f	g	h	i	j	k	l		a	b		a	b	a	b	a	b	a	b	a	b		a	b		a	b	c	d	SOE		
BTRA	AC	5	5	5	5	5	0	5	3	3	4	5	2	5	4	1	1	1	0	0	1	0	1	1	1	1	1	0	2	1	1	1	1	2	1	1	1	2.258	2.609
	AD	5	5	5	5	5	0	5	4	4	5	5	0	5	5	5	5	3	2	2	5	5	5	5	5	5	5	4	5	5	4	3	4	5	5	5	5	4.226	4.348
	AE	5	5	5	5	5	0	5	3	3	5	1	3	5	5	0	0	5	5	5	2	0	3	5	4	4	5	3	5	5	5	4	4	4	4	4	4	3.710	3.435
	AH	5	5	3	5	3	1	5	2	3	5	1	3	5	1	1	2	5	3	2	1	1	5	5	1	1	5	3	4	3	5	5	4	4	2	4	3	3.194	3.261
	AM	5	5	5	5	5	0	5	3	3	5	1	3	5	5	5	5	5	4	4	3	3	0	5	3	0	5	4	4	4	4	3	4	4	4	4	4	3.742	3.696
	AS	5	5	5	5	5	3	5	3	3	5	4	3	5	5	4	5	3	5	5	5	4	4	5	1	5	5	5	5	5	4	3	5	4	5	5	5	4.323	4.217
	BF	5	5	3	5	4	1	5	0	2	5	1	0	1	0	3	3	3	3	3	1	0	5	5	5	5	5	3	4	4	4	3	4	4	4	4	4	3.097	3.217
	BP	5	5	5	5	5	1	5	3	3	5	1	4	5	1	1	1	1	2	1	1	2	3	4	1	0	1	3	2	2	1	2	2	1	2	2	2	2.613	3.000
CSE	AF	1	5	5	5	4	0	5	4	3	5	3	3	1	1	3	2	5	3	2	5	2	3	1	1	0	4	3	3	4	3	4	4	4	4	4	4	3.000	2.957
	AJ	1	3	5	4	4	1	5	3	3	4	4	3	1	2	3	3	4	4	4	1	0	5	5	1	0	5	5	4	4	4	4	3	4	4	4	4	3.194	3.130
	AK	1	5	5	5	5	5	5	5	5	5	4	1	1	4	3	3	5	3	3	5	5	5	5	1	3	5	4	5	5	5	4	5	5	5	5	5	4.032	3.957
	AL	1	5	0	5	5	0	5	5	5	5	3	5	1	5	4	5	5	5	4	1	0	4	5	4	5	4	5	5	5	5	4	5	5	5	5	5	3.871	3.739
	AV	1	0	0	0	1	0	5	3	3	3	0	1	1	1	1	2	2	1	1	1	0	0	0	1	0	0	0	0	2	3	1	2	1	1	2	2	1.097	1.043
	BB	1	5	5	5	5	1	5	2	3	5	1	1	1	1	1	1	3	2	2	1	0	4	5	5	5	1	2	3	3	2	2	4	3	3	3	3	2.677	2.913
	BE	1	4	5	5	3	1	5	3	4	5	5	5	1	5	2	2	2	1	1	1	0	5	5	0	0	1	0	0	3	3	2	3	2	3	3	3	2.581	2.783
	BR	1	5	5	5	5	1	5	1	3	5	1	2	1	1	1	1	3	3	3	1	2	0	0	0	0	0	0	3	2	2	3	1	2	2	2	2	2.097	2.043

Table 67: CS-2 – 2nd Scoring SME 1

SME1																																							
System	ID	1	2												3	4		5	6		7		8		9		10		11	12		13	14 - Ovearll					DQCS	GDSP
			a	b	c	d	e	f	g	h	i	j	k	l		a	b		a	b	a	b	a	b	a	b	a	b		a	b		a	b	c	d	SOE		
BTRA	AC	5	5	5	4	4	3	5	3	4	5	5	1	5	5	2	1	3	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	1	2	2	2.677	2.913
	AD	5	5	5	5	5	0	5	4	4	5	4	0	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4.581	4.435
	AE	5	5	3	5	3	5	5	3	3	5	1	1	5	1	5	5	1	5	4	5	5	3	5	1	1	5	5	4	4	4	1	4	5	3	5	5	3.645	3.696
	AH	5	5	3	5	3	4	5	0	2	5	1	2	5	0	3	3	5	4	4	1	1	5	5	5	5	5	4	4	5	5	4	3	3	3	3	3	3.645	3.739
	AM	5	5	5	5	5	3	5	5	5	5	5	5	0	2	3	4	4	2	2	0	0	5	5	1	0	5	5	4	4	4	3	4	4	4	5	4	3.581	3.913
	AS	5	5	3	5	4	5	5	5	5	5	4	3	5	5	5	5	5	4	4	5	5	3	3	1	1	5	4	5	5	5	4	5	5	5	5	5	4.290	4.174
	BF	5	1	5	1	4	4	5	4	4	2	5	3	5	4	3	4	4	4	4	0	1	5	5	1	1	5	3	5	5	5	4	5	5	4	5	5	3.581	3.478
	BP	5	5	5	5	5	1	5	3	3	5	1	5	5	1	0	0	5	5	4	0	0	4	5	0	0	5	0	3	4	4	3	3	3	4	4	4	3.097	3.130
CSE	AF	1	5	5	5	5	2	5	2	4	5	3	1	1	1	1	2	5	2	2	1	1	4	4	5	5	5	3	3	3	3	3	3	3	3	3	3	3.129	3.435
	AJ	5	5	4	5	5	5	5	3	3	5	5	2	1	3	5	5	5	4	4	1	1	3	3	4	5	1	1	4	4	4	3	4	4	4	4	4	3.645	3.739
	AK	3	5	3	5	5	4	5	5	5	5	3	1	1	5	5	5	5	3	3	1	5	3	3	1	1	5	4	5	5	5	4	5	5	5	5	5	3.806	3.783
	AL	1	5	3	5	5	5	5	4	4	5	1	4	1	2	2	4	5	3	3	1	1	5	5	5	5	5	5	5	5	5	3	5	5	5	5	5	3.774	3.913
	AV	1	1	1	3	3	1	5	3	3	3	5	1	1	1	1	3	3	1	1	1	1	1	1	1	1	1	1	0	1	3	1	4	1	1	4	2	1.742	1.957
	BB	3	4	1	5	4	4	5	5	5	5	5	1	1	5	3	3	5	4	3	1	1	4	4	4	4	1	1	4	3	3	4	4	4	4	4	4	3.387	3.391
	BE	3	5	1	5	2	4	5	4	4	5	4	3	2	5	3	3	4	4	4	1	1	3	3	1	1	5	5	4	3	3	3	3	3	3	3	3	3.323	3.304
	BR	3	5	1	5	4	1	1	4	4	5	5	4	1	5	5	5	3	4	5	1	0	3	0	1	0	5	5	4	3	4	3	4	3	3	4	3	3.194	3.043

Table 68: CS-2 – 2nd Scoring SME 2

SME 2																																							
System	**ID**	1	2												3	4		5	6		7		8		9		10		11	12		13	14 - Ovearll					**DQCS**	**GDSP**
			a	b	c	d	e	f	g	h	i	j	k	l		a	b		a	b	a	b	a	b	a	b	a	b		a	b		a	b	c	d	**SOE**		
BTRA	AC	5	5	5	5	5	0	5	3	3	4	5	2	5	4	1	1	1	0	0	1	0	1	1	1	1	1	0	2	1	1	1	1	2	1	1	1	2.258	2.609
	AD	5	5	5	5	5	0	5	4	4	5	5	0	5	5	5	5	3	2	2	5	5	5	5	5	5	5	4	5	5	4	3	4	5	5	5	5	4.226	4.348
	AE	5	5	5	5	5	0	5	3	3	5	1	3	5	5	0	0	5	5	5	2	0	3	5	4	4	5	3	5	5	5	4	4	4	4	4	4	3.710	3.435
	AH	5	5	3	5	3	1	5	2	3	5	1	3	5	1	1	2	5	3	2	1	1	5	5	1	1	5	3	4	3	5	5	4	4	2	4	3	3.194	3.261
	AM	5	5	5	5	5	0	5	3	3	5	1	3	5	5	5	5	5	4	4	3	3	0	5	3	0	5	4	4	4	4	3	4	4	4	4	4	3.742	3.696
	AS	5	5	5	5	5	3	5	3	3	5	4	3	5	5	4	5	3	5	5	5	4	4	5	1	5	5	5	5	5	4	3	5	4	5	5	5	4.323	4.217
	BF	5	5	3	5	4	1	5	0	2	5	1	0	1	0	3	3	3	3	3	1	0	5	5	5	5	5	3	4	4	4	3	4	4	4	4	4	3.097	3.217
	BP	5	5	5	5	5	1	5	3	3	5	1	4	5	1	1	1	1	2	1	1	2	3	4	1	0	1	3	2	2	1	2	2	1	2	2	2	2.613	3.000
CSE	AF	1	5	5	5	4	0	5	4	3	5	3	3	1	1	3	2	5	3	2	5	2	3	1	1	0	4	3	3	4	3	4	4	4	4	4	4	3.000	2.957
	AJ	1	3	5	4	4	1	5	3	3	4	4	3	1	2	3	3	4	4	4	1	0	5	5	1	0	5	5	4	4	4	4	3	4	4	4	4	3.194	3.130
	AK	1	5	5	5	5	5	5	5	5	5	4	1	1	4	3	3	5	3	3	5	5	5	5	1	3	5	4	5	5	5	4	5	5	5	5	5	4.032	3.957
	AL	1	5	0	5	5	0	5	5	5	5	3	5	1	5	4	5	5	5	4	1	0	4	5	4	5	4	5	5	5	5	4	5	5	5	5	5	3.871	3.739
	AV	1	0	0	0	1	0	5	3	3	3	0	1	1	1	1	2	2	1	1	1	0	0	0	1	0	0	0	0	2	3	1	2	1	1	2	2	1.097	1.043
	BB	1	5	5	5	5	1	5	2	3	5	1	1	1	1	1	1	3	2	2	1	0	4	5	5	5	1	2	3	3	2	2	4	3	3	3	3	2.677	2.913
	BE	1	4	5	5	3	1	5	3	4	5	5	5	1	5	2	2	2	1	1	1	0	5	5	0	0	1	0	0	3	3	2	3	2	3	3	3	2.581	2.783
	BR	1	5	5	5	5	1	5	1	3	5	1	2	1	1	1	1	3	3	3	1	2	0	0	0	0	0	0	3	2	2	3	1	2	2	2	2	2.097	2.043

REFERENCES

Adelman, L. (1992). Evaluating Decision Support and Expert Systems. New York: John Wiley & Sons, Inc.

Aldag, R. J., & Power, D. J. (1986). An Empirical Assessment of Computer-Assisted Decision Analysis. Decision Science, 17, 572–588.

Amason, A. C. (1996). Distinguishing the Effects of Functional and Dysfunctional Conflict on Strategic Decision Making: Resolving a Paradox for Top Management Teams. The Academy of Management Journal, 39(1), 123–148.

Armstrong, M. P., & Densham, P. J. (1990). Database Organization Strategies for Spatial Decision Support Systems. International Journal of Geographical Information Systems, 4(1), 3–20.

Axelrod, R., & Cohen, M. (1999). Harnessing Complexity: Organizational Implications of a Scientific Frontier. New York: Free Press.

Baron, J. (1988). Thinking and Deciding. New York: Cambridge University Press.

Belton, V., & Stewart, T. (2002). Multiple Criteria Decision Analysis: An Integrated Approach. Springer.

Bennet, A., & Bennet, D. (2008). The Decision-Making Process in a Complex Situation. In F. Burstein & C. W. Holsapple (Eds.), Handbook on Decision Support Systems 1: Basic Themes (pp. 3–20). Berlin: Springer.

Berry, D. C., & Broadbent, D. E. (1995). Implicit Learning in the Control of Complex Problems. In P. A. Frensch & J. Funke (Eds.), Complex Problem Solving: The European Perspective (pp. 131–150). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Bich, W., Cox, M., & Harris, P. (2006). Evolution of the Guide to the Expression of Uncertainty in Measurement. Metrologia, 43(4), S161.

Blech, C., & Funke, J. (2010). You Cannot Have Your Cake and Eat It, too: How Induced Goal Conflicts Affect Complex Problem Solving. The Open Psychology Journal, 3, 42–53.

Bolia, R. S., Nelson, W.T., Vidulilch, M. A., & Taylor, R. T. (2004). From Chess to Chancellorsville: Measuring Decision Quality in Military Commanders. In P. A. Hancock (Ed.), Human Performance, Situation Awareness, and Automation: Current Research and Trends (pp. 269–282). Mahway, NJ: Lawrence Erlbaum Associates, Publishers.

Borchers, J. G. (2005). Accepting Uncertainty, Assessing Risk: Decision Quality in Managing Wildfire, Forest Resource Values, and New Technology. Forest Ecology and Management, 211, 36–46.

Braddock, C. H., Edwards, K. A., Hasenberg, N. M., Laidley, T. L., & Levinson, W. (1999). Informed Decision Making in Outpatient Practice. Journal of the American Medical Association, 282(24), 2313–2320.

Brehmer, D. (1995). Feedback Delays in Complex Dynamic Decision Tasks. In P. A. Frensch & J. Funke (Eds.), Complex Problem Solving: the European Perspective (pp. 1–25). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Brooke, J. (1996). SUS-A Quick and Dirty Usability Scale. In Usability Evaluation in Industry (pp. 189–194).

Buchner, A. (1995). Basic Topics and Approaches to the Study of Complex Problem Solving. In P. A. Frensch & J. Funke (Eds.), Complex Problem Solving: the European Perspective (pp. 27–64). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Casey, M. J., & Austin, M. A. (2002). Semantic Web Methodologies for Spatial Decision Support. Presented at the DSIage 2002, University College Cork. Ireland.

Cats-Baril, W. L., & Huber, G. P. (1987). Decision Support Systems for Ill-Structured Problems: An Empirical Study. Decision Sciences, 18(3), 350–372.

Chen, T. ., Wang, H. ., & Lu, Y. (2011). A Multicriteria Group Decision-Making Approach Based on Interval-Valued Intuitionistic Fuzzy Sets: A Comparative Perspective. Expert Systems with Applications, 38(6), 7647–7658.

Clark, M. J., & Richards, K. J. (2002). Supporting Complex Decisions for Sustainable River Management in England and Wales - Clark - 2002 - s - Wiley Online Library. Aquatic Conservation: Marine and Freshwater Ecosystem, 12, 471–483.

Clemen, R. T., & Reilly, T. (2001). Making Hard Decisions. Pacific Grove, CA: Duxbury.

Courtney, J. F. (2001). Decision Making and Knowledge Management in Inquiring Organizations: Toward a New Decision-Making Paradigm for DSS. Decision Support Systems, 31(1), 17.38.

Crossland, M. D., Wynne, B. E., & Perkins, W. C. (1995). Spatial Decision Support Systems: An Overview of Technology and a Test of Efficacy. Decision Support Systems, 14(3), 219–235. doi:doi: DOI: 10.1016/0167-9236(94)00018-N

Cummins, R. A., & Gullone, E. (2000). Why We Should Not Use 5-Point Likert Scales: The Case for Subjective Quality of Life Measurement. Proceedings, Second International Conference on Quality of Life in Cities, 74–93.

Davern, M. J., Mantean, R., & Stohr, E. A. (2008). Diagnosing Decision Quality. Decision Support Systems, 45(1), 123–139.

Dawes, R. M. (1988). Rational Choice in an Uncertain World. San Diego: Harcourt Brace Jovanovich.

De Silva, F. N., Eflese, R. W., & Pidd, M. (2003). Evacuation Planning and Spatial Decision-Making: Designing Effective Spatial Decision Support Systems Through Integration of Technologies. In G. Mora, G. Forgionne, & J. N. D. Gupta (Eds.), Decision Making Support Systems: Achievement and Challenges for the New Decade (pp. 358–373). Hershey, PA: Idea Group Publishing.

Densham, P. J. (1991). Spatial Decision Support Systems. In D. J. Maguire, M. F. Goodschild, & D. W. Rhind (Eds.), Geographical Information Systems: Principles and Applications (pp. 403–412). New York: John Wiley & Sons, Inc.

Dickinson, H. J., & Calkins, H. W. (1988). The Economic Evaluation of Implementing a GIS. International Journal of Geographical Information Systems, 2(4), 307–327.

Dodge, Y. (2006). The Oxford Dictionary of Statistical Terms. Oxford University Press.

Edwards, W., Kiss, I., Majone, G., & Toda, M. (1984). What Constitutes "A Good Decision?" Acta Psychologica, 56, 5–27.

Fernandes, R., & Simon, H. (1999). A Study of How Individuals Solve Complex and Ill-Structured Problems. Policy Sciences, 32(3), 225–245.

Fischhoff, B. (1975). Hindsight Is Not Equal to Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1(3), 288–299. doi:doi:10.1037/0096-1523.1.3.288

Forgionne, G. A., Gupta, J. N. D., & Mora, M. (2003). Decision Making Support Systems: Achievements, Challenges and Opportunities. In M. Mora, G. Forgionne, & J. N. D. Gupta (Eds.), Decision Making Support Systems: Achievement and Challenges for the New Decade (pp. 358–373). Hershey, PA: Idea Group Publishing.

Frank, A. (2008). Analysis of Dependence of Decision Quality on Data Quality. Journal of Geographical Systems, 10(1), 71–88.

Frensch, P. A., & Funke, J. (1995). Definitions, Traditions, and a General Framework for Understanding Complex Problem Solving. In P. A. Frensch & J. Funke (Eds.), Complex Problem Solving: the European Perspective (pp. 1–25). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Frisch, D., & Clemen, R. T. (1994). Beyond Expected Utility: Rethinking Behavioral Decision Research. Psychological Bulletin, 116, 46–54.

Frisch, D., & Jones, S. K. (1993). Assessing the Accuracy of Decisions. Theory & Psychology, 3, 115–135.

Funke, J. (1991). Solving Complex Problems: Exploration and Control of Complex Systems. In R. J. Sternberg & P. A. Frensch (Eds.), Complex Problem Solving: Principles and Mechanisms (pp. 185–222). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Funke, J. (2010). Complex Problem Solving: A Case for Complex Cognition? Cognitive Process, 11, 133–142.

Gorry, G. A., & Morton, S. (1971). A Framework for Management Information Systems. Sloan Management Review, 13(1).

Hagmayer, Y., & Meder, B. (2013). Repeated Causal Decision Making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(1), 33–45.

Hayes, R. E., & Wheatley, G. (2001). The Evolution of the Headquarters Effectiveness Tool (HEAT) and Its Applications to Joint Experimentation. Presented at the 6th International Command and Control Research and Technology Symposium, Annapolis, MD.

Hough, J. R., & Ogilvie, G. T. (2005). An Empirical Test of Cognitive Style and Strategic Decision Outcomes. Journal of Management Studies, 42(2), 417–448.

Howard, R. A. (1988). Decision Analysis: Practice and Promise. Management Science, 34(6), 679–695.

Huber, O. (1995). Complex Problem Solving as Multistage Decision Making. In P. A. Frensch & J. Funke (Eds.), Complex Problem Solving: the European Perspective (pp. 151–176). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Jacoby, J. (1977). Information Load and Decision Quality: Some Contested Issues. Journal of Marketing Research, 14(4), 569–573.

Janis, I. L., & Mann, L. (1977). Decision Making. New York: Free Press.

Jonassen, D. (2000). Toward A Design Theory of Problem Solving. Educational Technology Research and Development, 48(4), 63–85.

Jonassen, D. H., & Hung, W. (2008). All Problems Are Not Equal: Implications for Problem-Based Learning. The Interdisciplinary Journal of Problem-Based Learning, 2(2), 6–28.

Joyner, R., & Tunstall, K. (1970). Computer Augmented Organizational Problem Solving. Management Science, 17, B212–B225.

Kadish, R., Abbot, G., Cappuccio, F., Hawley, R., Kern, P., & Kozlowski, D. (2006). Defense Acquisition Performance Assessment. Department of Defense.

Kanungo, S., Sharma, S., & Jain, P. K. (2001). Evaluation of A Decision Support System for Credit Management Decisions. Decision Support Systems, 30(4), 419–436. doi:doi: DOI: 10.1016/S0167-9236(00)00126-3

Keen, P. G. W., & Morton, M. S. S. (1978). DSS: An Organizational Perspective. Reading, MA: Addison-Wesley.

Keren, G., & Bruine de Bruin, W. (2003). On the Assessment of Decision Quality: Considerations Regarding Utility, Conflict, and Accountability. In D. Hardma & L. Macchi (Eds.), Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making (pp. 347–363). New York: Wiley.

Kerns, P. A. (1995). Cognitive Flexibility and Complex Problem Solving. In P. A. Frensch & J. Funke (Eds.), Complex Problem Solving: the European Perspective (pp. 201–218). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Keys, D. J., & Schwartz, B. (2007). "Leaky" Rationality: How Research on Behavioral Decision Making Challenges Normative Standards of Rationality. Perspectives on Psychological Science, 2, 162–180.

King, W. R. (1983). Planning for Strategic Decision Support Systems. Long Range Planning, 16(5), 73–78. doi:doi: DOI: 10.1016/0024-6301(83)90080-8

King, W. R., & Rodriguez, J. I. (1978). Evaluating Management Information Systems. Management Information Science Quarterly, 2, 43–51.

Kleinmuntz, D. N. (1990). Decomposition and the Control of Error in Decision-Analytic Models. In R. Hogarth (Ed.), Insights in Decision Making: A Tribute to Hillel J. Einhorn (pp. 107–126). Chicago: Chicago University Press.

Kluwe, P. A. (1995). Single Case Studies and Models of Complex Problem Solving. In P. A. Frensch & J. Funke (Eds.), Complex Problem Solving: the European Perspective (pp. 269–294). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Kraemar, H. C., & Theimann, S. (1987). How Many Subjects?: Statistical - Google Books. Newbury Park, NJ: Sage Publications.

Lilien, G. L., Rangaswamy, A., Van Bruggen, G. H., & Stark, K. (2004). DSS Effectiveness in Marketing Resource Allocation Decisions: Reality vs. Perception. Information Systems Research, 15(3), 216–235.

Lipshitz, R. (1989). "Either a Medal or a Corporal": The Effects of Success and Failure on the Evaluation of Decision Making and Decision Makers. Organizational Behavior and Human Decision Processes, 44, 380–395.

Lowery, R. (2012). T to P Calculator. Calculators for Statistical Table Entries. Web Utility. Retrieved from http://www.vassarstats.net/tabs.html#r

March, S. T., & Smith, G. F. (1995). Design and Natural Science Research on Information Technology. Decision Support Systems, 15, 251–266.

Murphy, E. (2004). Identifying and Measuring Ill-Structured Problem Formulation and Resolution in Online Asynchronous Discussions. Canadian Journal of Learning and Technology, 30(1), 5–20.

Myers, J. L., Well, A. D., & Lorch Jr., R. F. (2010). Research Design and Statistical Analysis (3rd ed.). Routledge.

Mysiak, J., Giupponi, C., & Rosato, P. (2005). Towards the Development of A Decision Support System for Water Resource Management. Environmental Modelling & Software, 20, 203–214.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive Strategy Selection in Decision Making. Journal of Experimental Psychology: Learning, Memory, & Cognition, 14, 534–552.

Peters, D., Jackson, L., Philips, J., & Ross, K. (2008). The Time to Decide: How Awareness and Collaboration Affect the Command Decision Making. In Battle of Cognition (pp. 193–211).

Philips-Wren, G., Hahn, E., & Forgionne, G. (2004). A Multiple-Criteria Framework for Evaluation of Decision Support Systems. OMEGA, 32, 323–332.

Pick, R. A. (2008). Benefits of Decisions Support Systems. In F. Burstein & C. W. Holsapple (Eds.), Handbook on Decision Support Systems 1: Basic Themes (pp. 719–727). Berlin: Springer.

Pomerol, J.-C., & Adam, F. (2003). From Human Decision Making to DMSS Architecture. In M. Mora, G. Forgionne, & J. N. D. Gupta (Eds.), Decision Making Support Systems: Achievement and Challenges for the New Decade (pp. 40–70). Hershey, PA: Idea Group Publishing.

Pool, M., Russ, T., Schneider, D., Murray, K., Fitzgerald, J., Mehrota, M., Miraglia, P. (2003). Evaluating Expert-Authored Rules for Military Reasoning. New York: Association for Computing Machinery.

Powell, W., Laskey, K., Adelman, L., Dorgan, S., Johnson, R., Klementowski, C., Braswell, K. (2008). Evaluation of Advanced Automated Geospatial Tools: Agility in Complex Planning. Presented at the 13th International Command & Control Research and Technology Symposium: Adapting C2 for Complex Endeavors, Seattle, WA. Retrieved from http://ite.gmu.edu/~klaskey/papers/Powell_etal_ICCRTS08.pdf

Powell, W., Laskey, K., Adelman, L., Johnson, R., Dorgan, S., Hieb, M., Powers, M. W. (2010). Evaluation of Geospatial Digital Support Products. Fairfax, VA.

Preacher, K. J. (2002, May). Calculation for the Test of the Difference Between Two Independent Correlation Coefficients. Calculation for the Test of the Difference Between Two Independent Correlation Coefficients. Web Utility. Retrieved from http://quantpsy.org

Quesada, J., Kintsch, W., & Gomez, E. (2005). Complex Problem-Solving: A Field in Search of a Definition? Theoretical Issues in Ergonomics Science, 6(1), 5–33.

Robbins, R., & Hall, D. (2007). Decision Support for Individuals, Groups, and Organizations: Ethics and Values in the Context of Complex Problem Solving. AMCIS Proceedings, Paper 329. Retrieved from ttp://aisel.aisnet.org/amcis2007/329

Saaty, T. L. (1990). How to Make a Decision: The Analytic Hierarchy Process. European Journal of Operational Research, 48(1), 9–56.

Schneider, S. L., & Shanteau, J. (Eds.). (2003). Introduction: Where to Decision Making? In Emerging Perspectives on Judgment and Decision Research (pp. 1–10). New York: Cambridge University Press.

Schweiger, D. M., & Sandberg, W. R. (1989). The Utilization of Individual Capabilities in Group Approaches to Strategic Decision. Strategic Management Journal, 10, 31–43.

Schweiger, D. M., Sandberg, W. R., & Rechner, P. L. (1989). Experiential Effects of Dialectical Inquiry, Devil's Advocacy, and Consensus Approaches to Strategic Decision Making. The Academy of Management Journal, 32(4), 745–772.

Serfaty, D., MacMillan, J., Entin, E. E., & Entin, E. B. (1997). The Decision-Making Expertise of Battle Commanders. In C. E. Zsambok & G. Klein (Eds.), Naturalistic Decision Making (pp. 233–246). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Sharda, R., Barr, S. H., & McDonnell, J. C. (1988). Decison Support System: A Review and an Empirical Test. Management Science, 34(2), 139–159.

Shin, N., Jonassen, D. H., & McGee, S. (2003). Predictors of Well-Structured and Ill-Structured Problem Solving in an Astronomy Simulation. Journal of Research in Science Teaching, 40(1), 6–33.

Simon, H. A. (1973). The Structure of Ill-Structured Problems. Artificial Intelligence, 4(3-4), 181–201. doi:doi: DOI: 10.1016/0004-3702(73)90011-8

Simon, H. A. (1976). From Substantive to Procedural Rationality. In S. J. Latis (Ed.), Method and Appraisal in Economics (pp. 129–148). New York: Cambridge University Press.

Stabell, C. (1994). Towards a Theory of Decision Support. In P. Gray (Ed.), Decision Support and Executive Information Systems (pp. 45–57). Englewood Cliffs, NJ: Prentice-Hall.

Stanners, M., & French, H. T. (2005). An Empirical Study of the Relationship Between Situation Awareness and Decision Making (Technical Report No. DSTRO-TR-1687). Defense Science and Technology Organization. Retrieved from http://dspace.dsto.defence.gov.au/dspace/bitstream/1947/4318/1/DSTO-TR-1687.pdf

Tarantilis, C. D., & Kiranoudis, C. T. (2002). Using a Spatial Decision Support System for Solving the Vehicle Routing Problem. Information & Management, 39(5), 359–375. doi:doi: DOI: 10.1016/S0378-7206(01)00103-3

Taylor, J. (1997). An Introduction to Error Analysis: the Study of Uncertainties in Physical Measurements. University Science Books.

Tyler, L. E. (1983). Thinking Creatively: A New Approach to Psychology and Individual Lives. San Francisco: Jossey-Bass.

USACE. (2003). Battlespace Terrain and Reasoning Awareness-Battle Command (BTRA-BC) Fact Sheet.

USACE. (2009). Battlespace Terrain and Reasoning Awareness-Battle Command (BTRA-BC) Ongoing Research. Retrieved September 15, 2010, from http://www.erdc.usace.army.mil/pls/erdcpub/images/ERDC_FS_Research_BTRA.pdf

USACE. (2010, June). Battlespace Terrain and Reasoning Awareness-Battle Command (BTRA-BC) Fact Sheet. Retrieved September 15, 2010, from http://www.agc.army.mil/fact_sheet/BTRA.pdf

Vahidov, R., & Fazlollahi, B. (2004). Pluralistic Multi-Agent Decision Support System: A Framework and An Empirical Test. Information & Management, 41, 883–898.

Visone, D. (2008). AGC - BTRA-BC CJMTK Extension. Retrieved September 15, 2010, from http://www.agc.army.mil/btra/bc_extension.html

Von Winterfeldt, D. (1980). Structuring Decision Problems for Decision Analysis. Acta Psychologica, 45(1-3), 71–93. doi:doi: DOI: 10.1016/0001-6918(80)90022-0

Xu, J., Wang, G. A., Li, J., & Chau, M. (2007). Complex Problem Solving: Identity Matching Based on Social Contextual Information. Journal of the Association for Information Systems, 8(10), 525–545.

Yates, J. F., Veinott, E. S., & Patalano, L. A. (2003). Hard Decisions, Bad Decisions: on Decision Quality and Decision Aiding. In S. L. Schneider & J. Shanteau (Eds.), Emerging Perspectives on Judgment and Decision Research (pp. 13–63). New York: Cambridge University Press.

Zakay, D. (1984). The Evaluation of Managerial Decisions' Quality by Managers. Acta Psychologica, 56(1-3), 49–57. doi:doi: DOI: 10.1016/0001-6918(84)90006-4

Zsambok, C. E. (1997). Natualistic Decision-Making: Where Are We Now? In C. E. Zsambok & G. Klein (Eds.), Naturalistic Decision Making (pp. 233–246). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

BIOGRAPHY

Walter (Andy) Powell was born in Dayton, Ohio, but spent most of his youth in Mount Solon, Virginia. He attended the United States Naval Academy where he received his Bachelor in Science (Electrical Engineering) and his commission in the United States Navy in 1983. He instructed at Nuclear Power Training Unit, Ballston Spa and completed his division officer tour as Damage Control Assistant on board USS Michigan (SSBN-727)(Gold). While serving as a Visiting Lecturer attached to the ROTC unit, he received his Master of Engineering from Cornell University in 1991. After serving as Combat Systems Officer on board USS Asheville (SSN-758) and staff tours with the Defense Nuclear Agency, he served as the Submarine Liaison Officer to the Canadian Navy, and as Information Technology Officer for Commander U.S. Naval Forces, Central Command and U.S. Fifth Fleet. He graduated from the U.S. Army Command and General Staff College in 2001. He retired from the U.S. Navy as a Lieutenant Commander in 2003. He received his Doctorate in Information Technology from George Mason University in 2014. He will continue as a member of the research faculty at George Mason University affiliated with both the Center of Excellence in Command, Control, Communication, Computing, and Intelligence and the Center for Assurance Research and Engineering.