Winter 2003

Development of an Aggregation Methodology for Risk Analysis in Aerospace Conceptual Vehicle Design

Trina Marsh Chytka
Old Dominion University

DEVELOPMENT OF AN AGGREGATION METHODOLOGY FOR RISK ANALYSIS IN AEROSPACE CONCEPTUAL VEHICLE DESIGN

By

Trina Marsh Chytka
B.S. in Civil Engineering., 1997, Old Dominion University
M.S. in Engineering Management, 2001, Old Dominion University

A Dissertation submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

ENGINEERING MANAGEMENT

OLD DOMINION UNIVERSITY
December 2003

Approved by:

Resit Unal (Director)

Charles B. Keating (Member)

Andres Sousa-Poza (Member)

Bruce A. Conway (Member)

ABSTRACT

DEVELOPMENT OF AN AGGREGATION METHODOLOGY FOR RISK ANALYSIS IN AEROSPACE CONCEPTUAL VEHICLE DESIGN

Trina Marsh Chytka
Old Dominion University
Director: Dr. Resit Unal

The growing complexity of technical systems has emphasized a need to gather as much information as possible regarding specific systems of interest in order to make robust, sound decisions about their design and deployment. Acquiring as much data as possible requires the use of empirical statistics, historical information and expert opinion. In much of the aerospace conceptual design environment, the lack of historical information and infeasibility of gathering empirical data relegates the data collection to expert opinion.

The conceptual design of a space vehicle requires input from several disciplines (weights and sizing, operations, trajectory, etc.). In this multidisciplinary environment, the design variables are often not easily quantified and have a high degree of uncertainty associated with their values. Decision-makers must rely on expert assessments of the uncertainty associated with the design variables to evaluate the risk level of a conceptual design. Since multiple experts are often queried for their evaluation of uncertainty, a means to combine/aggregate multiple expert assessments must be developed. Providing decision-makers with a solitary assessment that captures the consensus of the multiple experts would greatly enhance the ability to evaluate risk associated with a conceptual design.

The objective of this research has been to develop an aggregation methodology that efficiently combines the uncertainty assessments of multiple experts in multiple disciplines involved in aerospace conceptual design. Bayesian probability augmented by uncertainty modeling and expert calibration was employed in the methodology construction. Appropriate questionnaire techniques were used to acquire expert opinion; the responses served as input distributions to the aggregation algorithm. Application of the derived techniques were applied as part of a larger expert assessment elicitation and calibration study.

Results of this research demonstrate that aggregation of uncertainty assessments in environments where likelihood functions and empirically assessed expert credibility factors are deficient is possible. Validation of the methodology provides evidence that decision-makers find the aggregated responses useful in formulating decision strategies.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER I

INTRODUCTION

Background

When challenged with understanding complex, technological systems, managers often use analysis to characterize risk. Managers use this information to guide the design of projects, develop policy, increase reliability and ultimately reduce costs. The complexity of modern systems has required a level of rigor in risk analysis that necessitates the gathering of as much information as possible for a more robust view of risk and uncertainty. Traditional data gathering techniques have been historical information and empirical statistics from which trends, likelihood functions and regression algorithms could be employed in risk analysis. When the traditional data methods have been deficient, the use of expert judgments to provide uncertainty assessments has been used. The subjective nature of expert assessments has prompted decision-makers to elicit uncertainty judgments from multiple experts in an attempt to have a more comprehensive understanding of the uncertainty involved in the analysis. However, multiple subjective judgments does not guarantee a consensus assessment of uncertainty – often the multiple experts have divergent opinions. In an attempt to resolve non-consensus and provide useful information to decision-makers, aggregation methods have been developed to combine divergent opinions. Numerous methods for combining information across experts have been proposed in the literature– from group interaction techniques and brainstorming to opinion pools and Bayesian techniques. There appears to be no consensus regarding the best method to use or relevant attributes for making that determination. An extensive literature review reveals that there has been little to

---

The journal model for the references herein is *The American Psychologist*, the journal of the American Psychological Association

over another. This research focuses on the evaluation and selection of an expert opinion aggregation methodology and its viability in satisfactorily combining multiple opinions in the conceptual design domain.

Problem Domain

The National Aeronautics and Space Administration (NASA) is chartered with the task of developing versatile space transports with improved reliability, lower cost, and with more payload-to-orbit capability. These future vehicles are of revolutionary designs employing innovative and often yet-to-be-developed state of the art systems. To determine the attributes associated with these conceptual design vehicles, NASA utilizes various resources. These resources involve current space transport designs and technology, which must be extrapolated to address the requirements and anticipated technology of the future vehicle. The process of extrapolating from the current technology requires engineering judgment; the application of which is based upon the availability of data and the level of expertise of the analyst performing the extrapolation.

To develop a comprehensive framework to examine the feasibility of some future system, analysts and decision-makers often rely on output from models of behavior, simulation models or other mechanisms of prediction and couple those outputs with expert judgment to assess the credibility of the predictions. The instrument of interest for this research is a methodology that gathers assessments from recognized experts in fields of aerospace engineering and technology and utilizes those estimations to determine the risk and uncertainty associated with some future embodiment of space transportation.

When the value of an uncertain quantity is needed in an analysis and limits in data or understanding preclude the use of traditional statistical techniques, the only remaining option

is to query experts for their best professional judgment (Morgan & Henrion, 1990). Issues with significant variability, issues that are controversial and issues that are highly complex and/or have a high consequence associated with them are well suited for expert opinion elicitation (Ayyub, 2001; Morgan & Henrion, 1990; Beach, 1975). Expert judgments have been used informally for many years; more formal approaches to expert assessment have become prevalent since World War II (Ayyub, 2001). Application areas for expert assessment have been diverse, including nuclear engineering, various types of forecasting (economic, meteorological, technical, etc), military intelligence, seismic risk and environmental risk from toxic chemicals (Clemen & Winkler, 1997; Winkler & Poses, 1993; Chen, Fine & Hubermann, 2003; Schuenemeyer, 2002; Cornell, 1996). Additionally, experts have been tasked with providing estimates of parameters of systems that have yet to be developed. In such cases, experts have relied on values from similar parameters and extrapolated the unknown parameter value (Monroe, 1997; Hampton, 2001).

In some situations, it may be sufficient to gather information from a single subject matter expert. The solitary assessment may then be accepted as the estimate for the quantity or parameter of interest (Ayyub, 2001). However, in many of today's more complex systems, the need to gather as much information as possible for a robust, comprehensive assessment has become vital. The use of multiple experts is often employed when very little empirical data is available and/or when multiple disciplines are involved in the analysis. A consequence of querying opinions from multiple experts may be a lack of consensus or elicited estimations may be quite divergent (Clemen & Winkler, 1997; Rantilla & Budescu, 1999). Decision-makers may not find this type of non-consensus information useful when

evaluating conceptual designs with high levels of uncertainty. Therefore, a means to adjudicate the disparity is necessary.

The manner in which multiple opinions are combined into a solitary assessment is known as aggregation. Aggregation is a term which has been widely used in the literature on problems of group interaction, consensus belief formation and decision-making. The combination of information can occur via group interaction processes (behavioral methods), mathematical algorithms (mathematical methods) or a combination thereof.

The behavioral methods are rooted in the assumption that group interaction will stimulate reduced variability among the expert assessments, thus reducing the uncertainty in the aggregated response (Genest & Zidek, 1986). The objective of a group interaction process may be to reach agreement or consensus about phenomena of interest or simply to share information and have experts learn from each other. Traditional group interaction techniques are the Delphi Method, Nominal Group Technique and Brainstorming (Genest & Zidek, 1986; Lawrence, Edmundson & O'Connor, 1986). Interaction processes can also be altered to suit a particular situation or assessment objective.

The mathematical methods seek to remove the highly subjective group interaction attributes from the combination process and integrate opinions from a more objective principle. Initial mathematical paradigms were postulated on axiom-based formulas. The strategy was to assume certain properties that the combined distribution should follow and then derive the functional form of the combined distribution. The opinion pool algorithms, which use weighted combinations of expert reliability, are the predominant methods utilized in the axiomatic approach (Clemen & Winkler, 1997; Winkler & Poses, 1993; Staël von Holstein, 1970). As complexities of phenomena were revealed, Bayesian approaches

emerged. The Bayesian paradigm asserts that as new information is obtained concerning the variable of interest, prior distributions and likelihood functions can be updated to improve the credibility of the aggregated response. This particular method accounts for expert independence, information redundancy and variable dependencies, which makes it a much more complex technique. The Bayesian approach appears more comprehensive than the axiomatic methods; however, experts agree that it is discouragingly difficult to apply and invalidated if likelihood functions and expert credibility are unavailable (Clemen & Winkler, 1997; Morris, 1977).

Aggregation methods have found wide acceptance in decision science domains where likelihood functions and expert reliability can be easily observed or calculated, such as stock market predictions, weather forecasting, and medical diagnosis (Chen, Fine & Hubermann, 2003; Morris, 1977; Engemann, Miller & Yager, 1995; Staël von Holstein, 1970). More profoundly, combination methods have proven quite useful in environments fraught with high uncertainty and/or high consequence and in areas with limited empirical data (nuclear engineering, petroleum resources, seismic analysis) (Schuenemeyer, 2002; Cornell, 1996). The versatility in combining multiple forms of information (qualitative, discrete, probabilistic) and the flexibility of being applicable to many decision domains has increased the use of combination paradigms to aid analyst and decision-makers in risk assessments.

Ensuring the aggregated result provides useful information to the decision-maker requires careful construction of the data acquisition instrument (Genest & Zidek, 1986). There are several possible impediments that must be identified and mitigated when creating and deploying an expert elicitation mechanism. The identification and selection of appropriate expert(s) whose assessments will be queried is paramount. The experts should,

at a minimum, satisfy predetermined qualification criteria tailored for the specific analysis and show consistency in providing reliable judgments (Ayyub, 2001; Morgan & Henrion, 1990).

Once an appropriate expert base is identified, the form of elicitation must be established. An information gathering mechanism (such as a questionnaire) must be both efficient and effective, ensuring that the time commitment of the participants is considered. Great care must be taken to ensure the elicitation instrument is concise, follows a logical pattern and abides by the highest standards of professional ethics (Ayyub, 2001; Genest & Zidek, 1986).

A third impediment with potential ramifications is the decision about the number of experts to be queried. The number of available experts in a specific discipline may negate this concern if only one expert exists in a particular field. Realistically, however, multiple experts ($n>1$) should be part of the population from which expert selection may occur. If more than one expert is used for the same discipline, a means to aggregate the responses in a meaningful way must be implemented. Otherwise, individual differences in experts' experience, confidence in judgment, and innate baseline from which judgments are based will likely render inconclusive (or at least less precise) results (Conway, 2003).

Significance of Problem

Traditionally, uncertainty has been quantified using historical data or empirical statistics. In domains where data is negligible and experimentation infeasible, the use of expert opinion has become a viable uncertainty quantification method. In order to gain as much knowledge about a system of interest, multiple experts are often queried for their opinion. However, expert assessments may not always agree. Additionally, the individual

assessments, if contradictory or non-uniform, may cause decision-makers more confusion rather than clarity when using these expert judgments as inputs to their decision strategy. To strengthen decision-making strategies, a means to adjudicate the disparity in expert opinions and provide decision-makers with a comprehensive solitary assessment of uncertainty is needed (Rantilla & Budescu, 1999; Morris, 1986). Aggregation is the mechanism to capture multiple opinions and combine them into a single assessment.

Most aggregation paradigms are premised on the availability of likelihood functions and expert credibility factors derived from historical data or empirical statistics. Aggregation methods, conventionally, have been applied to environments laden with data and to states of interest with known outcomes or with outcomes that could be verified in a reasonable time afterwards (knowable outcomes). On the other hand, aggregating multiple assessments in the conceptual domain is problematic since no prior distributions are available to determine likelihood functions of the variable(s) of interest. Without prior distributions or empirical data sets from which to extrapolate variable values/outcomes, the credibility of experts in the prediction of conceptual variables/outcomes is not possible. Additionally, in the present research case (advanced launch vehicle concept development) the wait for validation of an aggregated response could take more than twenty years. The characteristics of no likelihood functions, lack of empirical data sets and extended time horizons gives rise to a class of outcomes labeled "unknown". The term "unknown" for this research case may need to be clarified at this point. In the aerospace conceptual domain, the state of a system at the time decisions need to be made is unknown not necessarily unknowable. Given twenty to thirty years the state of knowledge may progress to such a point that the outcome becomes known. The relevance of the term (known versus unknown) must be coupled with a time reference

and for this research the time reference are decision milestones within a project or times in close proximity to the decision milestones. Currently, there does not exist a holistic methodology to mathematically aggregate multiple opinions of conceptual design experts in the absence of likelihood functions and expert credibility factors and in a domain assessing the "unknown future."

CHAPTER II

LITERATURE REVIEW

Decisions & Uncertainty

The use of expert elicitation and aggregation methods in a conceptual design environment involves the domain of decision sciences. Decision science involves not only the use of models to implement classical decision-making and strategic initiatives but also the use of decision processes as a predictive tool for risk assessment and evaluation of unknown outcomes. A key premise of decision science is that, ultimately, humans are responsible for making and implementing decisions, either directly or through the use of surrogate algorithms and simulations. There are many methods and models for analyzing decisions and designing strategies for implementing them. Each seek to augment or supplement human abilities in some manner.

Sound risk decision strategies cannot be formulated without prior identification and quantification of uncertainties (Morgan & Henrion, 1990). Uncertainty is the inability to determine the true state of a system and is caused by incomplete knowledge or stochastic variability (Haimes, 1998). There have been several attempts to create taxonomies of different kinds of uncertainty (Morgan & Henrion, 1990). Ayyub (2001) outlines a variety of uncertainty categories encountered in engineering design problems. Du and Chen (1999) further classify Ayyub's uncertainty structure into internal and external uncertainties. Internal uncertainty has two primary sources (Apostolakis, 1994; Lasky, 1996). (1) Limited information in estimating the characteristics of model parameters for a given, fixed model structure, often called model parameter uncertainty. (2) Limited information regarding the model structure itself, including uncertainty in the validity of the assumptions underlying the

model. External uncertainty comes from the variability in model prediction arising from plausible alternatives for input values (including both design parameters and decision variables)(Du & Chen, 1999). External uncertainty is also referred to as input parameter uncertainty. Gu, Renaud and Batill (1998) summarize and provide illustrations of the various categories of approximation error (uncertainty) associated with modeling and simulation. Computational error is also identified as an uncertainty that may be quantified but iterative refinement in a model usually negates quantification of this type of uncertainty (Sargent, 1999). An uncertainty category known as linguistic imprecision introduces uncertainty because the problem domain often involves imprecise terms such as "maybe false" or "sort of true" (Ayyub, 2001). The discipline of fuzzy logic was borne out of the need to quantify and mitigate implications of linguistic imprecision (Zadeh, 1992). Failure to account for uncertainties in decision strategies can produce results, which are suspect, misleading and erroneous (Ayyub, 2001).

The task of incorporating uncertainty into decision strategies does not stop at the identification of the uncertainty. Once the uncertainty has been identified, it must be quantified (Hampton, 2001). When assessing a high consequence course of action, most decision-makers prefer to evaluate pertinent information within a range (distribution) of possible values instead of a single discrete value (Ayyub, 2001). A probability distribution allows for the most comprehensive evaluation of uncertainty over a decision domain. Traditionally, probability has been classified as either objective or subjective. The objective view, often called Classical or Frequentist, is associated with such environments as the physical sciences, statistical formulations and modeling. The Subjective view or Bayesian

probabilities are not as easily quantified and are based on the lack of empirical data and reliance on expert based assessment of engineering variables.

The Classical / Frequentist view defines probability as the probability of an event occurring in a particular trial, given by the frequency with which it occurs in a long sequence of similar trials (Morgan & Henrion, 1990). While this view may be applicable to laboratory experiments or academic postulation, in practical application, decision-making is not quite so succinct and the relevant population of trials or even outcomes in unclear. The Bayesian view emphasizes that the probability of an event is the degree of belief that a person has that the event will occur, given all relevant information known to the person. Therefore, the probability now becomes not only the function of the probability of the event but also of the state of information involved. The Bayesian view of events is more aligned with a complex systems perspective and thus more applicable to the ever-increasing complexity of modern decision-making environments (Morris, 1977). While the Bayesian view is indeed based on subjectivity, the Bayesian approach must be consistent with the axioms of probability. For example, if one assigns probability p that an event will occur, then probability $1\text{-}p$ is its complement that the event does not occur (Morgan & Henrion, 1990).

Once the framing of the decision domain has been accomplished and the identification of uncertainty and uncertainty quantification method chosen, the next phase of the risk analysis is to query experts for their "best judgments" for the uncertainty of interest (Ayyub, 2001).

Expert Judgment Elicitation

Decision and policy makers are routinely interested in speculative knowledge and often query experts for their opinions (Ayyub, 2001). Decision-making using expert opinion

is not a new domain; however using a structured mechanism for their acquisition and elicitation did not emerge until after World War II. Expert opinion reached its peak in terms of public confidence during the Vietnam War but waned during the Nixon administration (Ayyub, 2001). Renewed interest in structured mechanisms for expert judgment elicitation can be attributed in part to the challenges of increased technological innovations and the complexity of modern day problems.

Simplistically, expert judgment is an expert's informed opinion, based on knowledge and experience, given in response to a technical problem (Meyer, et al., 2000). It can be viewed as a snapshot of the expert's state of mind and knowledge at the time his or her response to a technical problem is elicited. The use of expert opinion can be utilized to predict future events, provide estimates on new, complex or poorly understood phenomena, or integrate or interpret existing information (Meyer, et al., 2000).

The definition of expert and expert performance is vital to any analysis using expertise. If external standards exist for assessing the accuracy of expert judgment, the evaluation is straightforward. However, external standards rarely exist in domains requiring expertise, which is why experts exist in the first place. Ayyub (2001) asserts that an expert is someone who has had much training and has knowledge in some special field. This is a rather vague and generalized definition and often rather insufficient in determining appropriate individuals to elicit for expert opinion. A more detailed definition is necessary to fully comprehend the degree and type of knowledge required to qualify an individual as an expert. Prior conceptions of an expert used the number of years on the job (relevant experience) as a surrogate to expertise. Unfortunately, while many experts do indeed have significant length in service, time on the job does not necessarily equate to expertise. Some

individuals may work along side experts but never acquire the skills and knowledge to reach true expertise (Shanteau, et al., 2002). Although there are certain instances of positive correlation between experience and expertise, there is no evidence to support applying this standard universally (Jackson, 1999).

In many professions, accreditation has been used as a benchmark of expertise. For example, doctors may become "board certified" and engineers may achieve "Professional Engineer" status, these convocations imply a more skilled individual than someone not certified. The problem with accreditation is that it is often tied to time on the job or passing a one-time exam. Accreditation does not reflect sustained performance, even if a person's performance declines, the title and rank of accreditation remains (Shanteau, et al., 2002).

Peer identification seems to be the most widely utilized method of expert identification. Professionals are asked whom they would consider to be an expert. When there is some agreement on the identity of such individuals, then they are labeled as having expertise (Shanteau, et al., 2002). A drawback to this approach may be in the inherent "popularity effect" - someone who is better known or more popular with their peers is likely to be identified as an expert. (Shanteau, et al., 2002). Someone on the outside will unlikely be viewed as an expert although that person may be on the cutting edge of new insights.

In many fields, the identification of one or more "super experts" becomes a way to establish expert criteria. The answers of the Subject Matter Experts (SMEs) become the de facto standard from which all other experts are measured (Shanteau, et al., 2002). This approach is commonly used when no credible correct answers exist and when outcomes of events are unknown. SMEs offer valuable expertise in the domain of uncertainty because, in addition to knowing what the facts are, they understand what to do with those facts.

Expertise isn't just possessing knowledge or having qualifications; it is a highly specialized set of skills that have been honed in a particular situation of a specific purpose (Morgan & Henrion, 1990; Shanteau, et al., 2002; Jackson, 1999). As such, being an expert is quite distinct from having an education. Experts need to know more than just mere facts or principles of a domain in order to solve problems. Experts need to know which kinds of information are relevant to which kinds of judgments, how reliable different information sources are and how to make hard problems easier by decomposing them into smaller, more manageable units. Querying this type of knowledge, which is normally based on personal experience rather than formal training, is difficult to elicit (Jackson, 1999).

Several studies in the domain of expert elicitation have been performed- primarily in five fields of study: weather forecasting, medical diagnosis, psychology, business applications and military intelligence. The most notable work in the field of meteorological expert elicitation has been done by Murphy and Winkler (1974) with comparable studies performed by Fryback and Erdman (1979), and Daan and Murphy (1985). Murphy and Winkler (1974) demonstrated via the use of a calibration curve that although there was a slight tendency of meteorologists to over-predict the probability of precipitation, calibration was nearly perfect and the mean square error between the forecast probabilities and actual observed frequencies of the event was 0.028. An interesting finding of this study is that not only were the forecasts well calibrated, but they represented a 31% improvement over the climatological probabilities (Wallensten & Budescu, 1983). Fryback and Erdman (1979) suggest that forecasters are so erudite because they have available the climatological data on which to anchor their judgments. Murphy and Winkler (1974) counter this argument,

asserting that forecasters consult but do not necessarily rely upon the forecasting models when formulating their probabilities.

Psychology and expert opinion studies have primarily revolved around the issue of subjective probabilities. Beenen (1970) has done some interesting research using subjective probabilities to improve diagnoses in psychiatric clinical cases. Beenen asserts that while the clinical psychologist showed substantial individual differences in their probability assessments, they all display reasonable stability in their diagnosis processes. Tversky and Kahneman (1971) build upon the work of Beenen but concluded that most psychologist overestimate the power of "hypothetical research designs and underestimate the width of confidence intervals".

Subjective probabilities play an important role in different business applications (Druzdzel, 1989; Beach, 1975; Wallensten & Budescu, 1983). Subjective probability studies using expert elicitation in the business domain have included predicting the future interest rate of certificates of deposits and predicting the price of international stocks. Staël von Holstein (1970) had 72 experts predict prices of shares on the Stockholm Stock Exchange. Only three outperformed their predictions based upon a uniform distribution. The experts were able to observe their performance and received extensive feedback from the investigator. Future prediction iterations resulted in improved average variance indicating the experts were partially able to overcome their overconfidence bias (Wallensten & Budescu, 1983).

In earlier work related to the current research problem, Monroe (1997) developed a methodology for eliciting expert judgment to reduce uncertainty in decision analysis. The elicitation method developed by Monroe incorporated both qualitative and quantitative

elements to the elicitation process. The decision domain from which this methodology was applied was a launch vehicle conceptual design, specifically an application to a single-stage-to-orbit advance aerospace vehicle. Monroe investigated qualitatively assessing uncertainty of weight estimates for the various major components of an advanced orbital craft and then anchored those qualitative assessments quantitatively. The methodology consisted of a series of questions to determine the expert assessed qualitative rating of uncertainty for the parameter under investigation, followed by the anchoring of a most likely value. The qualitative rating of uncertainty was then coupled with a quantitative assessment of what that uncertainty rating meant. The questionnaire results were subsequently used in a Monte Carlo simulation to converge to a "summated" weight estimate for the vehicle under study. While the Monroe expert elicitation methodology is applicable to the current research problem, it does not address the use of multiple experts and multiple disciplines in the decision analysis.

Hampton (2001) adapted Monroe's expert elicitation methodology to quantify risk in a multidisciplinary design environment. The expert elicitation methodology was then combined with a Latin Hypercube Sampling simulation to propagate uncertainties across a multidisciplinary environment for the overall system (Hampton, 2001). Hampton's work addresses expert elicitation and uncertainty assessments but in the context of uncertainty propagation utilizing two disciplines (weights and sizing and aerodynamics). The element not found in these two pertinent research cases is aggregating the multiple opinions used in aerospace vehicle conceptual design. Neither the work of Monroe or Hampton addresses a holistic expert elicitation with aggregation methodology for uncertainty assessment.

Aggregation

The role experts play in a conceptual design environment is critical - their judgments provide valuable information and insight into areas where limited empirical data is available. The motivation to use multiple experts is a desire to obtain as much information as possible. Furthermore, conceptual design of complex systems is multidisciplinary, involving several experts in each discipline. Having multiple, independent assessments of an event or entity causes great concern for decision-makers – whose assessment is most viable? Should one assessment be weighted more strongly than another? To counter this dilemma, scientists and researchers propose an aggregation strategy to combine assessments that can "ideally be viewed as representing a summary of the current state of the expert opinions regarding the uncertainty of interest" (Clemen & Winkler, 1997).

Aggregation procedures are often divided into two approaches: mathematical and behavioral (Hampton, 2001; Clemen & Winkler, 1997; Rantilla & Budescu, 1999). Mathematical aggregation involves the integration of various independent assessments into one singular judgment. The mathematical methods range from simple summary measures (arithmetic or weighted averaging) to complex analysis involving the characteristics of the expert opinions such as the quality and dependence among the expert's assessments (Clemen & Winkler, 1997). Generally, this method is best suited for predictive and prescriptive models along with Bayesian probabilities.

The behavior aggregation methods attempt to generate agreement into a solitary assessment by having the experts interact in some way. This interaction can be either face-to-face group settings or information exchange without direct contact. The focus of the behavior aggregation method is on the quality of the individual expert judgments and the

dependence among such judgments implicitly rather than explicitly (Clemen & Winkler, 1997). Many researchers feel the behavior method reduces the amount of redundant information that must be aggregated and the bias of experts is more easily smoothed (Rantilla & Budescu, 1999; Clemen & Winkler, 1997; Genest & Zidek, 1986). However, no strategy has been discussed on how to eliminate the dominance factor and group polarization so often associated with group techniques.

Regardless of the aggregation method chosen, research indicates that combining the assessments of three experts yields the most advantage to aggregation (Rantilla & Budescu, 1999). There is little to no empirical evidence that adding additional experts improves the effectiveness or efficiency of the model outputs. In fact, a recent study by Rantilla and Budescu (1999) found that utilizing more than three experts led to less confidence in the estimate, a rather counter-intuitive result. The work of Clemen and Winkler (1997), Hogarth (1990), and Rantilla and Budescu (1999) support the three-expert postulate.

Many researchers on the subject of multiple expert aggregation methods agree that modeling is the most appropriate method to facilitate combining of assessments (Hampton, 2001; Vose, 2000; Molak, 1997). Much research has been done on aggregating point estimates from multiple experts but limited research has been done on modeling of aggregated opinions in probability distribution form (Hogarth, 1990; Vose, 2000; Rantilla & Budescu, 1999). The results of research have been mixed; however one theme seems to pervade each study – simpler aggregation methods perform better than complex methods. Clemen and Winkler (1997) assert the advantages of more simplistic approaches are the ease of use, robust performance and defensibility in public-policy settings. In addition, there is a lack of consistent empirical data to support the notion that more complex mathematical

models result in any greater accuracy of performance. Consequently, caution should be taken when evaluating combination methods. More simplistic methods do not allow for explicit consideration of such factors as over confidence and dependence among experts (Clemen & Winkler, 1997). Many researchers feel that further work with Bayesian models for multiple expert aggregation with careful attention to ease of use and sensitivity to intrinsic variability would improve performance. Clemen and Winkler (1997, p.24) quote from a study performed in 1994: " The Bayesian aggregation tool is demonstrably powerful but it is not well understood. Further studies to understand its behavior...need to be undertaken to realize its full potential."

Aggregation Methods

Consulting multiple experts may be viewed as a subjective version of increasing the sample size in an experiment. Because subjective information is often viewed as being "softer" than "hard scientific data", it seems particularly appropriate to consult multiple experts in an attempt to beef up the information base (Clemen & Winkler, 1997). The principle that underlies the use of multiple experts is that a set of experts can provide more information and/or clarity than a single expert. The form of the elicited response from the experts can be varied - qualitative, discrete, or probabilistic; and the elicitation methods themselves can be numerous (group interaction, independent assessment, questionnaires, etc.). To accommodate the multiple forms of elicited data and numerous elicitation approaches available, various forms of aggregation methods have emerged to address different combination protocols.

Mathematical Approaches

There are primarily two classes of algorithms for mathematically combining distributions of elicited experts: weighted averages (opinion pools) and Bayesian combinations. The opinion pool approaches are simple, intuitively appealing, and can generate a wide range of combination rules with ease. Bayesian approaches are motivated by treating each expert's judgment as data to be used in updating a prior distribution (Hammitt & Shlyakhter, 1999). A large variety of aggregation models for these algorithms have been developed, most of which have been reviewed descriptively by Clemen and Winkler (1997). Rantilla and Budescu (1999) assert that the simple mean models tend to work quite well in most applications and that in certain cases simple averaging is the optimal model to utilize. Hammitt and Shlyakhter (1999) affirm that the limited available evidence on relative performance of alternative combination methods suggests that simple averages often perform nearly as well as the theoretically superior Bayesian methods. A breakdown of the most common mathematical approaches follows.

The linear opinion pool (also known as weighted average) is the most simplistic of the mathematical approaches and merely a weighted linear combination of each expert's probability assessment as shown in Equation 1:

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta)$$

Equation 1: Linear Opinion Pool

where n is the number of experts, θ is the unknown variable of interest, $p_i(\theta)$ represents expert i's probability distribution, $p(\theta)$ represents the combined probability distribution, and the weights w_i sum to one. The weights (w_i) assigned to each probability represent the

relative quality of assessment assigned to each expert. The linear opinion pool method can be generalized to provide a broad set of combination rules, however, it does not allow for convenient representation of dependence among experts' judgments (Genest & Zidek, 1986).

A common argument regarding the use of the linear opinion pool is the selection of the weighting values, which are assigned to the experts' probability distribution. The weighting assignments may be based purely on the subjectivity of the decision-maker and his/her assessment of the reliability of the expert in estimating values or it may be based upon proven correlation among past performance of prediction tasks. Additionally, Genest and Zidek (1986) warn that the linear opinion pool is not well suited for the aggregation of density functions because the combined results typically become multimodal and do not provide a concise domain for decision-making.

A considerable advantage of the linear opinion pool is the ease of use and robustness in assigning reliability to the experts' assessments. The linear opinion pool is also less complex by not assessing the non-independence of experts. Non-independence is extremely difficult to quantify and can be attributed to three fundamental sources (Genest & Zidek, 1986, p.141):

- Overlapping data – in most cases, experts assess the same fundamental data from the same basic body of knowledge.
- Overlapping methodology – experts in particular fields have the same academic backgrounds and professional training. This is particularly prevalent in specialized fields or where expertise is scare.
- Direct observation and exchange of viewpoints – the presentation of reports and papers to the discipline community will result in commonalities due to the shared viewpoints.

By not addressing non-independence, the linear opinion pool reduces to a methodology of combining probabilities based upon the credibility assigned to each expert's assessment. The

simplicity of this technique is appealing for use in conceptual environments if a comprehensive framework for assigning weighting factors to the assessments can be developed.

The primary applications of the linear opinion pool method have been in forecasting and trending in such fields as meteorology, banking, medical diagnosis, and marketing (Clemen & Winkler, 1990; Staël von Holstein, 1970; Winkler & Poses, 1993; Engemann, Miller & Yager, 1995; Hurley & Lior, 2002). Hurley and Lior (2002) used linear opinion pool aggregation to forecast the selection of all-stars in a football conference, Staël von Holstein (1970) incorporated this method to forecast the interest rate on short term certificates of deposit while Engemann, Miller and Yager (1995) applied linear opinion pool aggregation to a decision regarding alternative configurations of power generators for money center banking. These applications employed empirical techniques to assign expert weighting factors and validated the aggregation outcomes with either historical data or validation from observance because the occurrence of the event was very near term. The use of subjective assessments for expert credibility and the application of the linear opinion pool method to an analysis where the outcome is in the very distant future (unknown) have not been supported in the literature.

The logarithmic opinion pool is a second form of the opinion pool approach and uses multiplicative averaging to aggregate probabilities. It is mathematically represented in Equation 2:

$$p(\theta) = \prod_{i=1}^{n} p_i(\theta)^{w_i}$$

Equation 2: Logarithmic Opinion Pool

where $p_i(\theta)$ represents expert i's probability distribution, $p(\theta)$ represents the combined probability distribution, θ is the unknown variable of interest, and the weights w_i sum to one. If the individual weights are equal to 1/n, then the combined distribution is proportional to the geometric mean of the individual distributions.

The logarithmic opinion method also utilizes a weighting element and thus suffers from the dilemma of assigning a reliability factor to the expert assessments. However, this approach is typically unimodal and less dispersed (Genest & Zidek, 1986) and, therefore, more likely to represent consensual values for the decision domain. Genest and Zidek (1986) also contend the most compelling reason for using a logarithmic opinion pool is that it is externally Bayesian. The property of external Bayesianity requires that when new data is added to the analysis, the posterior distribution result must be consistent whether the updating of distributions occurs prior to the combination of judgments or to the combined distribution itself. Winkler (Genest & Zidek, 1986) argues, however, that he would expect the weights assigned to the experts to change as new data are seen and therefore questions the advantage of the external Bayesian property.

Like the linear opinion pool, the logarithmic opinion pool does not address expert non-independence and is therefore unencumbered by the task of quantifying the level of expert dependence. Research does not present a clear-cut advantage of using one opinion pool method over the other. The form of the acquired elicited assessments, the depth of pre-existing data and the preference of the analyst are the primary discriminators in choosing one of the two opinion pool methods. Either method (linear or logarithmic) shows promise in combining expert assessments of uncertainty in a conceptual design environment.

The applications of the logarithmic opinion pool aggregation method have been very similar to that of the linear opinion pool method. Weather forecasting, stock market analysis, and medical diagnosis have been the primary domains of application (Clemen & Winkler, 1990; Staël von Holstein, 1970; Winkler & Poses, 1993; Chen, Fine & Hubermann, 2003). Much of the work has been to investigate the advantage of using one opinion pool method over another with the vast majority of research concluding there is no universal advantage to either system – both perform equally well (Rantilla & Budescu, 1999). Rantilla and Budescu continue however, that in light of their finding, simple weighted averaging is the optimal model due to ease of use and mathematical simplicity.

The conjugate aggregation method is more complex than the linear opinion pool or logarithmic opinion pool paradigms. This aggregation approach demands that all individual probability distributions belong to the family of beta distributions. The probability assignment that expert *i* attributes to hypothesis θ is generated by a beta distribution with parameters (α_{ij}, β_i). The consensus of *n* experts' opinions is obtained by applying Baye's Rule in order to obtain *n* beta distributions on hypothesis θ. The conjugate aggregation algorithms is represented by Equation 3:

$$\alpha_j = \sum w_{ij} * \alpha_{ij}, \qquad \beta = \sum \alpha_j, \qquad p(\theta) = \frac{\alpha_j}{\beta}$$

Equation 3: Conjugate Algorithm

For the current research, there is neither guarantee nor assumption that a beta distribution will be elicited from the experts. Beta distributions are plausible distribution alternatives for the final form of aggregated opinion but the initial assessments of uncertainty from the experts is expected to take the triangular distribution form. The triangular

distribution form is a viable distribution when the range of a and b and most likely value c can be approximated. The triangular distribution offers considerable flexibility in its shape and coupled with the intuitive nature of its defining parameters and speed of use makes it a compatible distribution for this research.

Additionally, the conjugate method also fails to solve the problem of how to assign applicable weighting factors to each expert's opinion in the combination process. While this approach presents limited promise for the current case, other methods discussed provide a more robust method for the aggregation of opinion in the current research domain.

An alternative approach to combining expert judgments is based on Bayes' Rule. Beginning with a prior distribution, the analyst treats each expert's distribution as new information and updates the prior distribution using Bayes Rule. The updating technique is dependent upon the likelihood function defined by the analyst. The likelihood function represents the probability that each expert will give the assessment as a function of the underlying state of nature and consequently incorporates the relative quality of experts' judgments (biases and confidence level). Genest and Zidek (1986) conclude that for typical risk analysis situations, in which a group of experts must provide information to a decision-maker, a Bayesian updating scheme is the most appropriate method. Equation 4 represents this approach:

$$p^* = p(\theta \mid C) \propto \frac{p(\theta) L(g_1, \ldots, g_n \mid \theta)}{p(g_1, \ldots, g_n)}$$

Equation 4: Bayesian Updating Scheme

where L represents the likelihood function associated with the experts' information. The notion of this paradigm is relatively straightforward. If n experts provide information $g_1, \ldots,$

g_n to a decision-maker regarding a variable of interest θ, then the decision-maker should use Bayes theorem to update a prior distribution. This formulation can be applied to any type of information from discrete forecasts or estimates to the combination of individual probabilities.

While this approach seems compelling, Clemen and Winkler (1997) caution that applying the method is frustratingly difficult. The primary problem associated with this method is the development of the likelihood function $L(g_1,..., g_n \mid \theta)$. The likelihood function amounts to the probabilistic model for the information $g_1,..., g_n$ and thus must also capture the interrelations among θ and $g_1,..., g_n$. If the interrelations are unknown or vague, trying to capture interrelations becomes problematic and introduces a measure of uncertainty that is difficult to quantify. In addition, if Genest and Zidek's (1986) assertion that the Bayesian techniques are most appropriate for typical risk analysis is accepted, the clarity in defining what constitutes a "typical risk analysis" is deficient.

The difficulty in assessing the likelihood function has given rise to the creation of "off-the-shelf" models (Genest & Schervish, 1985; French, 1981) for aggregating probabilities. Although these models apply various combination algorithms, common elements emerge from their function. Most of these models assume that prior distributions can be obtained for the variable(s) of interest and each expert's precision at forecasting can be observed. Often in the conceptual design environment, prior distributions for the variable of interest do not exist nor is it possible to assess the experts' precision in forecasting since the variables of interest are either state-of-the-art and/or yet to be developed systems. The difficulty in applying the Bayesian approaches, coupled with the inability to assess prior

distributions, makes this combination paradigm ill-suited for the conceptual design environment.

Behavioral Approaches

The focus of the behavior aggregation methods is on the quality of the individual expert judgments and the dependence among such judgments implicitly rather than explicitly (Clemen & Winkler, 1997). Behavioral combination approaches seek to gain consensus among the participants through various forms of interaction. Common behavioral techniques include the Delphi Method, Nominal Group Technique and Brainstorming. Many researchers feel the behavior methods reduce the amount of redundant information that must be aggregated and the bias of experts is more easily smoothed.

Brainstorming is the simplest behavioral combination technique (Morris, 1977). The objective is to assemble participants together and assign the task of generating a "group" consensus on a variable, event or phenomena of interest. Discussion and debate is the chosen forum with consensus reached through iterative sharing of information. Dialogue with unrestricted feedback is natural and easy if the individuals are able to communicate with each other (Genest & Zidek, 1986). The chief merit to brainstorming is the free exchange of information which may result in a reduction in the range of views presented. In practice however, this same interaction may induce conformity, "a degree of agreement beyond that which would be commensurate with the amount of information that is exchanged" (Genest & Zidek, 1986, p.125). Furthermore, experts are often unable to reach group consensus or agreement and thus group interaction breaks down into a process of negotiation and compromise. This may not reflect the "true" combination of opinions of the group but a

consensus of acceptable trade-offs often facilitated by a dominant group member (Clemen & Winkler, 1997).

Cornell (1996) applied the brainstorming technique to a problem domain with a comparable characteristic to the current research domain. Cornell developed a Probabilistic Seismic Hazard Analysis method designed to estimate the likelihood that various levels of earthquake caused ground-motions will be exceeded at a given location in a given future time period. He used the brainstorming technique to query a panel of experts on the validity of his methodology. Of particular significance and relevance to the current research case is that the validity of the brainstorming outcome cannot be empirically validated and the occurrence of the event is not knowable in a specific timeline.

The Delphi Method is perhaps the most widely known method for eliciting and synthesizing expert opinions (Morgan & Henrion, 1990). Although different variations exist, experts typically make individual judgments from a distance – no interaction is permitted. These judgments are shared anonymously with the participants. Each expert may then revise his or her assessment and the process is reiterated until the different opinions converge toward a common distribution. The purpose and specific steps of the Delphi method depend on the nature and purpose of use. Primarily, the uses can be categorized into technological forecasting and policy analysis (Ayyub, 2001). Technological forecasting relies on a group of subject matter experts for the problem being investigated and relies heavily on study facilitators to implement the method. Policy analysis seeks to incorporate opinion and views from the entire spectrum of stakeholders and seeks to communicate the spread of opinions to decision-makers (Ayyub, 2001). The Delphi Method can be inexpensive compared to other

group interaction techniques since experts need not communicate and the anonymity of the technique reduces social pressure often associated with group interactions (Hampton, 2001).

The Nominal Group Technique (NGT) is a related behavioral method. Experts first assess their probabilities individually and then present their distributions to the other group members. No discussion is to occur during the first round until all judgments have been presented (Morgan & Henrion, 1990). Each opinion is subsequently discussed in a structured format designed to prevent anyone from dominating the proceedings (Gustafson et al., 1973) followed by each expert ranking the list of opinions silently. Following silent assessment, each member calls out their ranking profile, rankings are then tallied and a consensus opinion emerges.

Delbecq et al. (1975) offer the following guidelines and cautions when using NGT. (1) NGT is best used for small group meetings called for the purpose of fact-finding, idea generation, or the search of problems or solutions. Once this technique becomes familiar, some steps will seem more important than others in different situations. For instance, clarification is more important when people in the group do not know one another or are from different backgrounds. (2) Formal balloting may not be necessary for relatively simple issues or for agenda setting when only a small number of topics emerge. (3) It is often difficult to convince people to use NGT for the first time. The usual question is, "why is all this structure necessary"? Explanations help to overcome this resistance, but a successful experience helps much more. It is a good idea to try out the process on an issue that can be covered completely in one meeting so that the group can sense the value of the entire process. (4) During early experiences using NGT, it is most difficult for people to keep from

discussing issues before all points are listed, clarified, and prioritized. So, extra care must be taken by the facilitator to prevent discussion from starting too soon.

Although group interaction methods seem relatively straightforward, they can suffer from many complications. Hogarth (1978) notes that some individuals tend to dominate the discussions thus discouraging the emergence of new ideas. Groupthink, or the propensity to adopt a more extreme opinion than would each individual member, is another possible complication of the behavioral methods. Hogarth (1978) counters these arguments by implying that interaction techniques need not be dysfunctional if experienced facilitator's assist in the process. Facilitators would serve to promote open dialogue, direct the discussions to maintain focus and guide participants to a consensus solution (Phillips & Phillips, 1990). However, no strategy has been developed to eliminate the dominance factor and group polarization so often associated with group techniques.

Synopsis of Literature

Aggregation is a methodology deployed for the combination of multiple assessments of uncertainty. The majority of research involves decision domains where data is either empirically or historically available thus likelihood functions and expert credibility factors are grounded in "hard" data. Additionally, the ability to validate an aggregated outcome has been feasible due to the preponderance of the event occurring in the near term. In the current research domain, "hard" data is negligible or non-existent therefore relying on subjective assessments for both uncertainty quantification and expert credibility is necessary. Behavioral aggregation techniques are well suited for subjective combination of expert opinions but suffer from many drawbacks including participant dominance, polarization, groupthink and compromise instead of optimization. Mathematical aggregation methods

remove subjectiveness from the combination process and seek to combine assessments from a more objective principle. Decision-makers and technologians are more comfortable with information based on "hard science" (objective) rather than soft science (subjective) (Sousa-Poza, 2003). Therefore, mathematical aggregation is appropriate for technical domains such as aerospace conceptual design environments. An extensive literature review helps to identify the research relevant to the current study and frames the context of the research case under investigation. Identification of a deficiency in the research domain has resulted in the development of a methodology to enhance the mathematical aggregation body of knowledge.

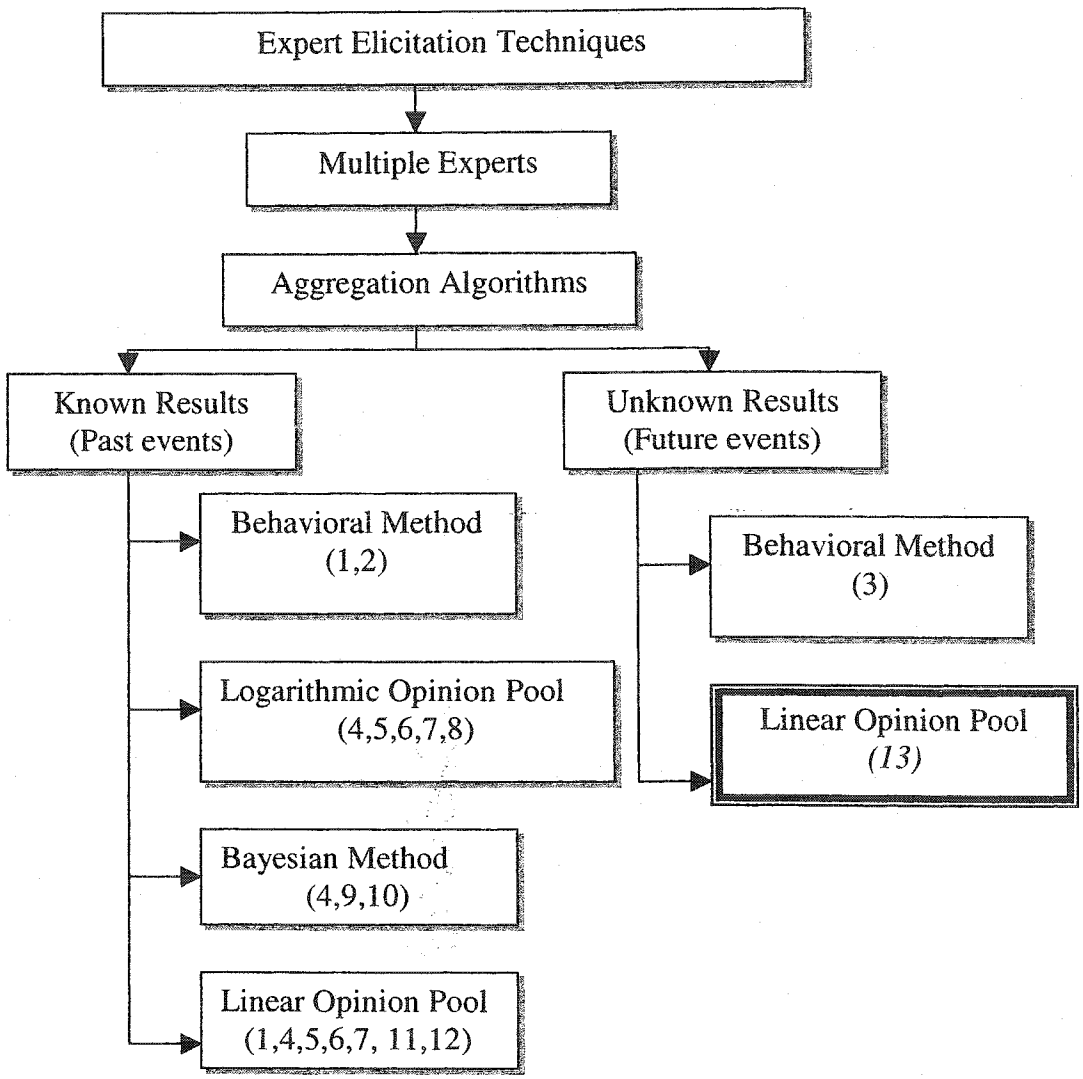
The literature review shows that much research has been accomplished in the domain of expert elicitation (Beenen, 1970; Murphy & Winkler, 1974; Beach, 1975; Wallensten & Budescu, 1983). The literature also indicates moderate research has been performed in the domain of expert elicitation with risk and uncertainty (Druzdzel, 1989; Hampton, 2001; Monroe, 1997). Aggregation of multiple expert judgments has also been moderately investigated but not sufficiently examined in the risk and uncertainty environment (Schuenemeyer, 2002; Cornell, 1996). Furthermore, the literature concerning investigation of risk and uncertainty as it applies to multiple expert elicitation and/or multiple disciplines is scarce (Hampton, 2001). Of complementary importance is the lack of research in a domain where the accuracy of elicited results cannot be empirically verified (Cornell, 1996). Concentrating research in domains with precise and quantifiable data and in environments where validation of results is observable or empirically capable does not enhance decision strategies to domains where data is negligible and outcomes unknown (Morris, 2003). Many of today's more complex engineering systems find themselves working in the latter environment and, thus, methodologies to assist in the reduction of risk and uncertainties for

more robust and efficient engineering designs would be an improvement to conceptual decision strategies.

The results of the literature study indicate the development of a methodology for multi-expert judgment elicitation employing an aggregation function in a multidisciplinary and multi-expert environment has yet to be investigated. Formulation of this methodology and its proven effectiveness by application to a conceptual launch vehicle design concept is a unique contribution to the aggregation literature. Table 1 summarizes the more prominent research related to the current problem and Figure 1 places the current research into context with past and current work in this field.

Author	Application	Aggregation Method				Results verified
		Linear Opinion Pool	Logarithmic Opinion Pool	Bayesian Method	Behavioral Method	
Lindley, D.V. (1983)	Unknown			√		√
Morris, P.A. (1977)	Unknown			√		√
Clemen & Winkler (1990)	Weather forecasting	√	√	√		√
Staël von Holstein (1970)	Stock Market prices	√	√			√
Winkler & Poses (1993)	Medical Diagnosis	√	√			√
Weerahandi & Zidek (1978)	Unknown	√	√			√
Engemann, Miller & Yager (1995)	Banking	√				√
Hurley & Lior (2002)	Football polling	√				√
Lawrence, Edmundson & O'Connor (1986)	Magazine advertising dollars	√			√	√
Chen, Fine & Hubermann (2003)	Stock Market prices		√			√
Schuenemeyer (2002)	Oil & Gas resources				√	√
Cornell (1996)	Seismic Risk analysis				√	

Table 1: Aggregation Literature Summary



Researchers

(1) Lawrence, et al. (1986)
(2) Schuenemeyer (2002)
(3) Cornell (1996)
(4) Clemen & Winkler (1990)
(5) Staël von Holstein (1970)
(6) Winkler & Poses (1993)
(7) Weerahandi & Zidek (1978)
(8) Chen, Fine & Hubermann (2003)
(9) Lindley, D.V. (1983)
(10) Morris, P.A. (1977)
(11) Engemann, Miller & Yager (1995)
(12) Hurley & Lior (1986)
(13) Chytka (2003)

Figure 1: Context of Current Research

Research Objectives

The role of experts in theoretical environments is critical - their judgments can provide valuable information and insight in areas where limited "hard" data is available. Decision-makers often rely on multiple opinions as a data set when historical or empirical statistics are deficient in a specific decision domain.

The prime concern of most researchers in risk analysis when using multiple experts is how the multiple opinions should be combined or aggregated to ensure adequate capture of diverse judgments. In order to make an expert elicitation data set useful, a means to adjudicate the disparity is necessary. Aggregation provides a means to combine divergent opinions to aid decision-makers in highly uncertain decision domains. To date, aggregation methods that directly address the combination of multiple opinions in an environment where likelihood functions cannot be determined and prior distribution availability is negligible has been deficient. The development of a complete methodology for expert judgment elicitation utilizing aggregation methods in a multi-discipline, multi-expert environment would be a significant asset to decision-makers in domains plagued with high uncertainty. By applying this methodology, the decision- maker can select the best, most cost-effective, risk tolerant solutions to provide the greatest long-term benefits.

The objective of this research is to develop and demonstrate a methodology to mathematically aggregate expert opinion in an environment where likelihood functions and expert creditability assessments are not available. To meet this objective, expert elicitation of uncertainty assessment must be queried from subject matter experts in disciplines of aerospace design. Uncertainty assessments must be quantified into probability distribution form. Because empirical statistics are not available, a technique to determine subjectively

the weighting factors assigned to each uncertainty assessment must be developed and deployed prior to aggregation. The linear opinion pool aggregation algorithm will be applied to the coupling of weighting factors and uncertainty distributions resulting in a "consensus" distribution of multiple experts. The methodology will be validated through subjective assessment by decision-makers on the usefulness of the combined response to enhance their decision strategies in risk assessments of future space transports.

CHAPTER III

RESEARCH METHODOLOGY

Approach

The thrust of this work is the development of a functional aggregation methodology to combine opinions from multiple experts who have assessed the uncertainty associated with multidisciplinary conceptual space vehicle design parameters. In most of the expert elicitation utilizing aggregation applications, the case studies involved scenarios in which the result could be validated because the phenomenon was either a past event of near term future event. Confirmation of the forecasted event allows for the development of likelihood functions, and precise expert reliability calculations. In the present research application, the confirmation of occurrences being assessed is infeasible; the designs are in the distant future. Due to this constraint, the validation of the methodology becomes subjective. The validation component of this research is a compilation of validation paradigms from Pederson et al. (2000), Sargent (1999) and Monroe (1997). The expert elicitation methodology deployed for this study is guided by the research of Conway (2003) and Monroe (1997). The linear opinion pool aggregation methodology is a synthesis of the approaches used by the researchers detailed in Chapter II but is draws specifically from the work of Rantilla and Budescu (1999), Clemen and Winkler (1997) and Genest and Zidek (1986).

For a more holistic capture of the research methodology, the design, development and deployment of the data acquisition technique (questionnaire), the design and deployment of the expert elicitation process as well as the aggregation methodology itself will be detailed in this section. The methodology in its entirety is represented in Figure 2.

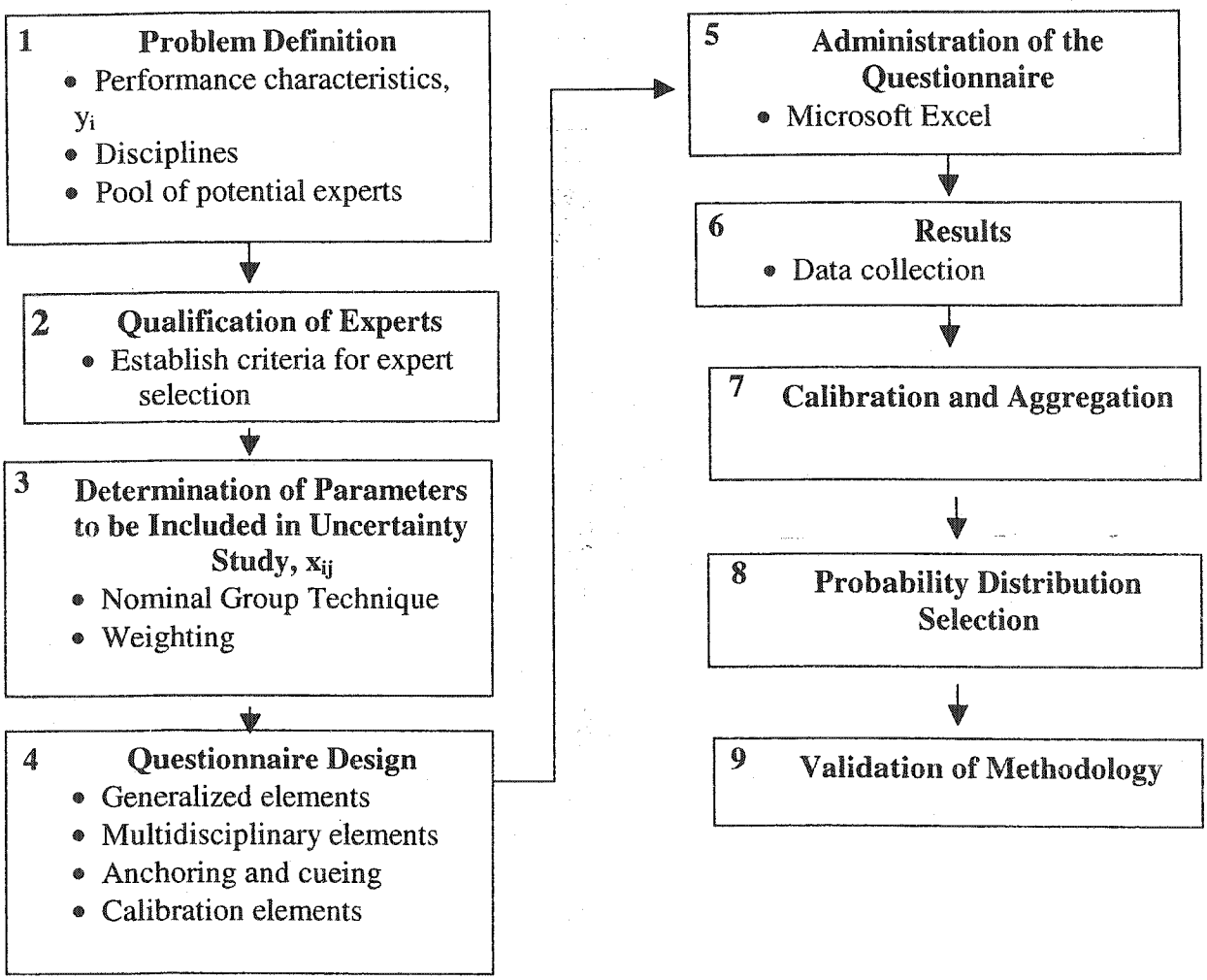


Figure 2: Outline of Research Methodology

Questionnaire Instrument

The questionnaire instrument for this research is of vital importance – it is the mechanism from which input distributions for the aggregation process will be obtained. The quality of information obtained from a questionnaire is directly proportional to the quality of the questionnaire, which in turn is directly proportional to the quality of the question construction process (Peterson, 2000). Determining the type of questions to be asked in a questionnaire is a function of the objective of the research, the nature of information to be collected, and the analysis method of the responses elicited. The two basic types of research

questions are open-end questions and close-end questions. Open-end questions are very general in nature and allow the participants to freely provide any answers they believe are appropriate. The open-end question structure works remarkably well when a researcher has no precognition of how a participant will answer a question or when researchers want to ensure the answers will not unduly be influenced by the presence of predetermined answer alternatives (Cox, 1980).

An alternative to the open-end question is the closed-end question. The distinguishing characteristic of a closed-end question is that answer alternatives are provided to the respondent. Much of the questionnaire strategies center around two closed-end forms: Dichotomous questions and Monadic scale (Peterson, 2000). Dichotomous questions are closed-end questions with only two possible answers or response categories. This question structure is best suited for obtaining demographic or behavioral information, they are not recommended when asking questions about psychological characteristics, such as attitudes, preferences or opinions (Ayyub, 2001). Monadic questions are a closed-end questions in which the answers are somehow graduated to measure a continuous construct, such as an attitude, opinion, perception or preference. The number of answer alternatives from closed-end questions can range from two to infinity however, in practice, seldom are there more than nine possible answers (Peterson, 2000).

In addition to constructing the question configuration (open-end versus closed-end), it is important to incorporate elicitation heuristics into the question structure. As the literature review indicates, anchoring and cueing is a common heuristic used when eliciting expert opinion (Lichtenstein & Newman, 1967; Budescu & Wallsten, 1987). The approach of cognitive cueing may be useful in helping experts to achieve anchoring of their opinions.

Cueing elicits specific judgment patterns from experts based upon experience or information and can aid in the adjustment of anchored values. The solicited experts will be asked to anchor their opinion of the parameter of interest via a mode (most likely) value, then cued to provide a response of least likely and most likely. To be effective, anchoring must be consistent, repetitious and use identical methods for eliciting a particular response pattern.

Once the questioning structure has been established, an appropriate scaling measure needs to be defined. A rating scale is a closed-end question whose answer alternatives are graduated to measure a continuous construct, such as an attitude, opinion, intention, perception or preference (Peterson, 2000). The most commonly used scales are unidimensional scales, scales that measure single predefined attributes. The unidimensional scales are further categorized into comparative scales and monadic scales. Comparative scales are simply rank-ordered while monadic scales are interval or ratio scales. In comparative scaling, the objects being scaled are directly compared to one another; consequently, the objects are scales relative to each other. A major benefit to comparative scaling is that it detects small difference among objects (Trochim, 2000). However, a major limitation to this scaling method is that it is not possible to generalize beyond the object being scaled. In monadic scaling, each object is measured independently of the other thus providing considerable scaling freedom. Although monadic scaling allows for scaling freedom, it prohibits the detection of the fine discriminations that are possible in comparative scaling (Peterson, 2000).

The appropriate number of rating scale categories is dictated by such factors as the mode the questionnaire will be administered, characteristics of the objects being scaled, the function of the research, and the ability of the researcher to handle multiple scaling

categories (Cox, 1980). In general, most rating scales have fewer than 12 categories with more than two thirds having either five or seven categories (Peterson, 2000). However, in practice, there is a widely held belief that the proper number of rating scales should be 7 plus or minus 2; no consensus exists on the proper number of categories (Peterson, 2000; Cox, 1980; Trochim, 2000). In general, there is no single, optimal number of rating scale categories for all scaling situations (Cox, 1980).

Once the number of rating categories has been determined, the form of the rating scale needs to be chosen. Three specific rating scales are most commonly used in measurement and scaling – semantic differential scaling, the Likert scale and the Stapel scale. The semantic differential scale developed by Charles Osgood is a seven-category, multi-dimensional scale that has only the extreme end points labeled. The semantic differential scale measures the meaning of any stimulus object on three dimensions: evaluation (good-bad), potency (strong-weak), and activity (active-passive) (Peterson, 2000).

Another widely used rating scale is the Likert scale, so named for its developer Rensis Likert. The Likert scale differs from the semantic scale in several distinct ways. First, it is a five-category, not seven-category rating scale and generally consists of two-parts: a declarative statement and a list of response categories. Unlike the semantic differential scale, all scale categories are labeled. The Likert scale is best suited for research involving summated opinion or attitude rating (Trochim, 2000). Several modified versions of the Likert scale have emerged since its inception, however the same general format holds true.

Jan Stapel developed a 10-category, uni-polar rating scale with categories numbered +5 to –5 (Trochim, 2000). This rating schema is aptly named the Stapel scale. Because of its unique numbering system, the Stapel scale is best represented vertically as opposed to

horizontally (as is the format for the Semantic differential scale and Likert scale). The Stapel scale is perhaps the most controversial rating scale and the least used (Peterson, 2000). Advocates of this rating scale assert that finer discriminations are permitted as compared to the other scaling schemes however opponents contend it is confusing and difficult to apply (Cox, 1980; Trochim, 2000).

The questionnaire methodology developed for this research application expands upon the work of Monroe (1997) and the literature cited above. The questioning style of Monroe is that of an open-end structure allowing the expert to answer questions freely and provide anchoring and cueing descriptions. Monroe also adopted Likert scaling for his questionnaire using the 5-point scaling system for both quantitative and qualitative ratings. The Monroe methodology also assumes a default symmetrical triangular distribution associated with an expert's assessed uncertainty about a parameter of interest. For the current research, the Monroe questionnaire methodology was adopted and modified in two distinct ways. First, the format for which respondents were able to provide the high and low values for a variable of interest was more explicitly detailed. The Monroe questionnaire, as structured, allowed the respondents to contradict their evaluations of uncertainty around a variable of interest. Peterson (2000) asserts that a properly structured questionnaire prohibits contradiction and ambiguity. The questionnaire structure was altered to reflect clarity in defining the uncertainty of the variable. Figure 3 presents the set of instructions for the experts responding to the questionnaire; in particular, instruction 3 reflects the primary modification made to Monroe's approach.

A list of (discipline specific model) input parameters whose values are potentially uncertain will be provided on a subsequent screen. You will be asked to evaluate these parameters using the following guidelines.

1. Rate each INPUT parameter uncertainty QUALITATIVELY using a 5-point rating scale (Low, Low/Moderate, Moderate, Moderate/High, High). Focus only on those INPUT parameters that you feel should be evaluated in this manner.
2. If you feel a parameter's default value should be modified, you may provide a new point estimate for the nominal value.
3. If you feel the range of possible values (due to uncertainty, physical limitations, design constraints, etc.) around the nominal value is not symmetrical, please provide your own estimates of minimum and maximum values.
4. Describe the reason for the uncertainty and the reasoning behind the parameter value ranges for the UNCERTAIN INPUTS that you rated. Include a rationale for those parameters to which you have assigned new nominal values. Do this simultaneously while rating each INPUT parameter to document your thinking.
5. Think of any other cues (or reasons that you have not documented) and record that information at this time.
6. Once the INPUT parameters provided have been rated for uncertainty, you may add parameters not shown which you assess to have a level of uncertainty associated with their value. Use the OTHER option listed at the bottom of the INPUT parameter listing for this purpose.
7. After rating all INPUT parameters, next anchor your Low, Moderate, and High QUALITATIVE measures of uncertainty to QUANTITATIVE measures on the 5-point scales (provided).
8. Describe any scenarios that may change INPUT parameter values. Provide the alternate INPUT parameter values that in your judgment would be appropriate for the scenario

Figure 3: Expert Elicitation Questionnaire Instructions

Secondly, a Background section was added to the questionnaire to allow for the calibration of expert assessments prior to the aggregation process. The Background section was added in support of the research efforts of Conway (2003) who developed an Expert Calibration Function to reduce the variability (uncertainty) of the assessment values. The input distributions for the aggregation algorithm are the calibrated distributions resulting from the Expert Calibration Function applied to the raw distributions from the questionnaire responses.

The context of the questionnaire centers on the elicited assessments from various disciplinary experts (E_{ij}) on design parameters evaluated most prone to uncertainty. The experts are queried to evaluate input design parameters (x_{ij}) as well as uncertainty associated with the analysis tools (z_{ij}) they employ in the conceptual design process. The questionnaire participants are asked to provide their assessments of the level of uncertainty associated with

the parameter (high, medium or low) and to adjust the nominal value provided if they feel the value is inaccurate. The nominal value (or adjusted value if provided) coupled with the level of uncertainty assigned to the parameter will determine the probability distribution assigned to the design parameter [(x_{ij}), (z_{ij})] from expert (E_{ij}). The questionnaire results from each disciplinary expert are the input distributions to the Expert Calibration Function developed by Conway (2003). The output of applying the calibration function to the distributions results in reduced variability distributions, which, in turn, are the inputs into the aggregation process.

The platform from which the questionnaire is launched is a Microsoft Excel® (Microsoft Corporation, 2000, Version 9.0 3821 SR-1) workbook. The Excel workbook includes a "tab" spreadsheet for instructions, a "tab" for a sample questionnaire, a "tab" for the Background section, multiple "tabs" for the variables of interest, and a "tab" to tool uncertainty assessment. The use of multiple "tabs" within the Excel shell enables the questionnaire to be exported to the experts in one compact file which makes for ease of use and practicality. The questionnaire is electronically mailed (e-mailed) to each SME for the respective disciplines. The advantage of using e-mail to distribute the questionnaire is that is allows the experts to assess uncertainty ratings on their own time and in a familiar setting – their workspace.

Population

The target population for this questionnaire is the pool of NASA aerospace engineers and design manager teams in multiple NASA locations. The chosen participants are recognized experts in their respective fields of study. In the present instance, familiarity with multidisciplinary launch vehicle design and optimization applications are also a key criteria.

The selection of appropriate subject matter experts by the design managers will be guided by the adherence to characteristics of expertise assembled from the literature. Much of the literature on identification of expertise (Shanteau, et al., 2002; Jackson, 1999; Ayyub, 2001) asserts that no one criteria should be used as a selection basis or disqualifier for the identification of an expert. For example, the literature does not support that "x" number of years of experience or "y" minimum educational background is used explicitly as selection criteria for the identification of experts. While there has been some positive correlation between years of experience or educational background, there is no evidence to support applying this standard universally (Conway 2003). The number of years of experience, educational background, cognitive skills, etc. are criteria to be integrated together in the selection process. No one criterion is considered a disqualifier for expertise; expertise is an integrated summation of the characteristics (criteria) described. The design team managers were given instructions to reflect upon the selection criteria listed in Table 2 as an integrated compilation of characteristics of an expert and then identify discipline specific experts based upon their subjective assessments of an individual's expertise.

Expert Characteristics
Domain knowledge ▪ Years of experience ▪ Educational background
Cognitive skills ▪ Ability to discern usefulness of data
Decision strategies
Expert-task congruence ▪ Appropriate expertise for discipline specific task

Table 2: Characteristics of an Expert

Aggregation Methodology

There has been much research relevant to the basic question of how people aggregate a variety of expert opinions to generate their own judgment and make decisions. Sound research for the selection of an aggregation methodology is explicitly guided by the answers to these 5 questions:

1. Who is doing the aggregation: a normative model, a decision-maker or a group?
2. What is the form of the information elicited and the response the decision-maker generates?
3. What is the nature of events that are relevant to aggregation: is uncertainty epistemic or aleatory?
4. Are there any inherent characterizations that can be made about the information pattern or information sources such as biases or redundancy in information?
5. What combination rule is to be utilized?

1. Who is doing the aggregation; a normative model, a decision-maker or a group?

For the present research effort, a normative model is chosen to facilitate aggregation of the elicited expert opinions. The mathematical aggregation methods can be performed manually, however several commercial off the shelf programs exist that facilitate the aggregation process and allow for higher order sampling techniques. Examples of such software include @RISK® (Palisade Software), Predict!® (Risk Decisions, Ltd.), and Crystal Ball® (Decisioneering, Inc.). @RISK® (Palisade Software, 2002, Version 4.5.2) works well within Microsoft Excel® and is utilized to facilitate the aggregation of multiple distributions.

2. What is the form of the information elicited and the response the decision-maker generates?

Each discipline expert is queried to provide the input parameters they feel are most prone to uncertainty for the design parameter under study. The initial uncertainty rating is qualitative (low, low/moderate, moderate, moderate/high, high). If the expert feels the

design parameter default value provided should be modified, a section is provided to document a new nominal value. The experts are then asked to describe the reason for the uncertainty associated with the design parameter and to document any cues related to the assessment. A sample questionnaire for Operations Support is provided in Appendix A.

After rating the input design parameters, the experts anchor their qualitative assessments of uncertainty to a quantitative measure as shown in Figure 4.

The amount of uncertainty or variation that I associate with Low Uncertainty is:						
Less	**5%**	**7.5%**	**10%**	**12.5%**	**15%**	**More**
The amount of uncertainty or variation that I associate with Moderate Uncertainty is:						
Less	**10%**	**15%**	**20%**	**25%**	**30%**	**More**
The amount of uncertainty or variation that I associate with High Uncertainty is:						
Less	20%	30%	40%	50%	60%	More

Figure 4: Quantitative assessment rating of uncertainty

The increments for this portion of the questionnaire were chosen as an extension of the format validated in the Monroe (1997) research. A slight modification was made to the increments however, to make the ordinal values more uniform. For example, in the Monroe study, the ordinal increments were non-uniform (for low uncertainty – 5%, 7.5%, 10%, 15%, 20%). Morgan and Henrion (1990) assert that when considering patterns relating to uncertainty, people are more comfortable assessing uncertainty with common incremental values. Assessing uncertainties that are not based on common increments imposes undue psychological stress on participants as they wrestle with the decision on which rating is more

appropriate. Additionally, the increments used in this research for low uncertainty ratings were based on an incremental scaling of 2.5%, the moderate uncertainty ratings were based on an incremental scaling of 5% and the high uncertainty rating incremented at 10% intervals. The chosen increments are a reflection of the work of Ayyub (2001) who contends that participants in an elicitation study prefer to think in smaller increments as the cognitive complexity to answer a question increases. As a participant evaluates an answer with a secure comfort level, the discriminators between answers become smaller. If a participant answers a question where the level of comfort in the answer is not so certain, participants generally think in larger incremental values. The incremental values associated with the assessment ratings reflect the cognitive complexity of assessing uncertainty in a conceptual design environment.

To construct the expert elicited uncertainty distribution, the default value provided on the questionnaire or the modified value provided by the expert becomes the "most likely" design parameter value. The quantitative value the expert assigns to the uncertainty associated with the "most likely" value constrains the outer tails of the distribution. For example: *Expert A* assigns a low uncertainty rating to design parameter x_i. *Expert A* agrees that the default value for x_i should be 5.96. The amount of uncertainty *Expert A* associates with a low uncertainty rating is 7.5%. Therefore, x_i has a value range of [5.96 * 7.5% = ± 0.447]; minimum =5.51, most likely =5.96 and maximum =6.41 (see Figure 5). The minimum, most likely and maximum values are calculated but little knowledge is available of what the "shoulder" values of the distribution resemble therefore, a triangular distribution is used.

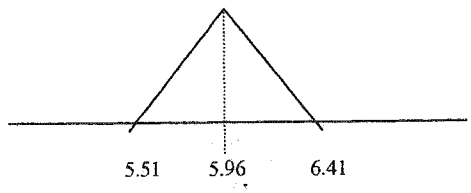


Figure 5: Triangular distribution example

The uncertainty distributions constructed and calibrated for each discipline specific design parameter become the input distributions to the aggregation process as shown in Figure 6.

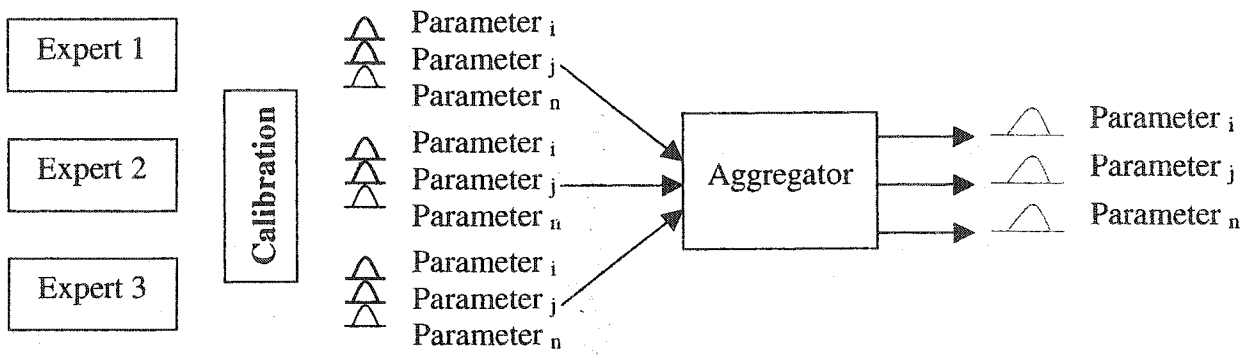


Figure 6: Aggregation Process Model

3.0 What is the nature of events that are relevant to aggregation: is uncertainty epistemic or aleatory?

The type of uncertainty encountered in a decision analysis should help determine which combination rule would be most appropriate. Epistemic uncertainty is the uncertainty associated with events that are uncertain due to internal factors such as incomplete knowledge or unreliable evidence (Rantilla & Budescu, 1999). Epistemic uncertainty can be abridged with increased knowledge about the phenomena and is reducible through research

and better data. Epistemic uncertainties are better suited for behavioral techniques since the sharing of views, opinions and ideas can enhance knowledge about the variable of interest.

Aleatory uncertainty concerns events that are uncertain due to external factors; they cannot be known in full detail nor are they reducible. Aleatory uncertainties are best aggregated using one of the mathematical approaches. The nature of events relevant to this aggregation case are primarily associated with the aleatory uncertainty domain since they are irreducible and not fully known.

4.0 Are there any inherent characterizations that can be made about the information pattern or information sources such as biases or redundancy in information?

Inherent characteristics of the information and redundancy in information may be handled through a calibration function. A calibration function was utilized by Conway (2003) as part of the larger research but is not a component of the aggregation algorithm itself.

5.0 What combination rule is to be utilized?

The deployment of an aggregation methodology to combine multiple expert opinions in a conceptual design environment is not straightforward nor has there been experimentation in this context to test one method's compatibility with conceptual environments over another. Most combination algorithms rely on an expert's prior distribution of the variable of interest or on known distributions for the variable to calculate likelihood functions and determine an expert's credibility. Neither of these two elements are available in the current research domain. Additionally, with no clear consensus among the literature of an aggregation method explicitly applicable to the attributes embedded in the conceptual design environment, the selection of an appropriate combination method is indeed the preference of the analyst. The attributes of the current research case coupled with extensive literature

review supports the use of a mathematical aggregation method, specifically an approach from the opinion pool sector. An extensive literature review reveals the linear opinion pool presents the most straightforward method of combining the opinions of the experts.

The design of the aggregation process itself is rather intuitive. The inputs for aggregation are the calibrated design parameter uncertainty assessments provided by each discipline specific expert. The experts are queried for their assessments of parameter uncertainty via a questionnaire and the expert assessments are calibrated using a calibration function derived from the answers in the Background section of the questionnaire. The calibrated distributions are input into a Microsoft Excel® spreadsheet to be read into a risk assessment model. The calibrated uncertainty assessments are then used as input distributions to the aggregation process.

Prior to the aggregation of the calibrated distributions, appropriate weighting factors are applied to each derived distribution. A weighting factor reflects the perceived credibility of an expert by the decision-maker. Since empirical statistics are not available from which to derive expert credibility, design managers are asked to provide credibility ratings for each expert. For the current research case, discipline specific design managers are queried to provide credibility ratings (weighting factors) which are applied to the distributions of the respective expert. The elicitation of weighting factors by the design managers occurs via an interview process in which the design managers are given a list of the experts who provide uncertainty assessments for design variables. The design managers, in the presence of a facilitator, rank the experts in terms of reliability of prediction on a scale from 0 to 1. For example, discipline *Design Manager A* may rate two experts who will provide uncertainty assessments for design variables for discipline *A*. *Design Manager A* may rank *Expert 1* with

a 0.70 credibility rating while *Expert 2* may receive a 0.30 credibility rating. The combined weighting factors for all experts being assessed by the design manager must equal 1.0. The design manager may base these weighting factors on subjective reasoning and his/her knowledge of each assessor's expertise, experience, and prediction capability (Morgan & Henrion, 1990).

The next task in the aggregation process is to import the calibrated uncertainty distributions and the experts weighting factors into an aggregation platform. The risk assessment model, @RISK® is utilized within the Microsoft Excel® shell to perform mathematical aggregation using the linear opinion pool algorithm. @RISK® allows for the specification of 37 distribution types including Beta, Erlang, Gamma, Normal, Triangular, Uniform, etc. For the current research case, uncertainty assessments are queried in triangular distribution form – the experts assess the minimum, most likely and maximum values for a parameter of interest. The calibration function applied to the initial assessments do not alter the distribution structure therefore the calibrated distributions are also in triangular form. The @RISK® function '*=RiskTriang(minimum, most likely ,maximum)*' converts values in spreadsheet cells into a triangular distribution from which weighting factors and combination algorithms are applied. The application of the aggregation algorithm results in a "most likely" value of the combined assessments. In order to determine the minimum and maximum values of combined responses, sampling of the distributions is performed.

The sampling simulation module within @RISK® samples from each of the expert assessed uncertainty distributions and applies the appropriate weighting factor to each distribution during the sampling process. The advantage of using the @RISK® software is that either a Monte Carlo or Latin Hypercube Sampling technique can be chosen for the

sampling process thus providing a more robust aggregated response. Monte Carlo Sampling is the more traditional sampling technique and is entirely random. Monte Carlo algorithms are available for all the distributions considered feasible for expert assessment (Vose, 2000) however, Monte Carlo suffers from efficiency issues. To increase the accuracy of Monte Carlo simulations Morgan and Henrion (1990) advocate increasing sample size. Consequently, to improve the accuracy of a Monte Carlo simulation a large number of iterations are typically required. This can become quite problematic when computing resources and/or time are limited. One of the variance reduction techniques employed to reduce the number of iterations required to improve computation efficiency is known as Latin Hypercube Sampling. The principle behind Latin Hypercube Sampling is stratification of the input probability distributions. Stratification divides the distribution into equal segments and a sample is then taken randomly from each segment. In this method, sampling is forced to represent values in each interval and thus, is forced to recreate the input probability distributions (Palisade, 2002). Latin Hypercube Sampling provides for faster run times by requiring less iteration for convergence. For this reason, Latin Hypercube Sampling is chosen as the sampling technique for this research.

The @RISK® software also has an embedded module called BESTFIT® which allows a user to call up the BESTFIT® subroutine and have the software perform goodness-of-fit tests on the sampled results and fit the most appropriate probability distribution to the resultant data. Once the simulation has been performed using the Latin Hypercube Sampling technique, the distribution fitting solution BESTFIT® takes the sampled data and finds the distribution function that best fits that data. BESTFIT® tests up to 26 distribution types using

advanced optimization algorithms. Results are displayed graphically and through a statistical report including goodness-of-fit statistics.

Once aggregation of multiple distributions has been applied and Latin Hypercube Sampling performed to determine a resultant distribution, the BESTFIT® module is used in this research to determine the most compatible aggregated probability distribution for the sample data. Since the inputs into the sampling process are probability distributions, a statistical goodness-of-fit statistic is used to verify the selection of the most compatible aggregated distribution response. BESTFIT® ranks all fitted distributions using one or more fit statistics. For continuous sampled data, the Chi-Square, Anderson-Darling or Kolmogorov-Smirnov statistics can be used. Molak (1997) and Vose (2000) contend the best representation of expert opinions is with a discrete distribution therefore the Chi-Square statistic will be used as the goodness-of-fit measure. For discrete sampled data, only the Chi-Square statistic is appropriate (Palisade, 2002). The Chi-Square statistic is the most well known goodness-of-fit statistic but it does have a weakness: there are no clear guidelines for selecting the number and location of the bins (Palisade, 2002). To minimize this weakness, @RISK® has an option that allows the user to set the number of bins to "Auto" and set the bin style to "Equal Probabilities". These choices are selected for the simulation.

Results Evaluation

The selection of the resultant aggregated distribution will be based on goodness-of-fit test statistics. The goodness-of-fit test compares a null hypothesis (H_0) with an alternative hypothesis (H_1). The test consists of computing a statistic based on sampled data. The goodness-of-fit statistic reports a measure of the deviation of the fitted distribution from the input data; the smaller the fit statistic, the better the fit (Palisade, 2002). For this research,

two test statistic values are important, the chi-square fit statistic and the p-value. The p-value or observed significance level, explains how likely it is that a set of N samples drawn from a fitted distribution would generate a fit statistic greater than or equal to a critical value. As the p-value decreases to zero, there is less confidence the fitted distribution could have generated the original data set (Ebeling, 1997). Conversely, as the p-value approaches one, there is no basis to reject the hypothesis that the fitted distribution actually generated the data set. For the chi-square goodness-of-fit statistic, if the p-value for the calculated λ^2 is $p > 0.05$, fail to reject the hypothesis that the deviation from the expected value(s) is small enough that chance alone accounts for it. If the p-value for the calculated λ^2 is $p < 0.05$, reject the hypothesis, and conclude that some factor other than chance is operating for the deviation to be so great. For example, a p-value of 0.01 means that there is only a 1% chance that this deviation is due to chance alone. Therefore, other factors, like linkage, must be involved (Ebeling, 1997).

The calculation of the chi-square test statistic is straightforward. The chi-square test statistic is based upon observed frequencies and expected frequencies of sampled data grouped into interval classes (bins). Samples are drawn from a population of data and observed frequencies (f_o) are documented and expected frequencies (f_e) postulated. A statistic can be calculated from the comparison:

$$\lambda^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Equation 5: Chi-Square Test Statistic

The computed chi-square value can be compared to the chi-square critical value (λ^2_{crit}) available from statistical reference books or computed using Microsoft Excel (CHIINV(probability, degrees of freedom)). If the calculated value is less than the critical value, the hypothesis about the underlying population is not rejected (Hanke & Reitsch, 1998). If the calculated test statistic is larger than the critical value, a poor fit is indicated. To determine the chi-square critical value, the Microsoft Excel CHIINV formula was used based upon a 95% confidence interval.

Table 10 summarizes the chi-square fit statistic and p-value for the aggregated responses. The hypothesis set forth is:

H_0: data fit the expected pattern

H_1: data do not fit expected pattern

If $\lambda^2 >$ critical value reject H_0.

$\lambda^2_{crit}(0.05,87)$	109.77
$\lambda^2_{crit}(0.05,73)$	93.75
$\lambda^2_{crit}(0.05,21)$	36.67

Validation

Validation of engineering research has conventionally demanded "formal, rigorous, and quantitative validation" (Barlas & Carpenter, 1990). Traditional validation methods are based primarily on logical inductive and/or deductive reasoning which works well in predictable, stable, data rich environments. These validation methods have traditionally been classified as objective – based on deviance measures and statistical tests (Law & Kelton,

1991; Sargent, 1999). Objective methods align quite well with the empiricisms that "formal, rigorous and quantitative" validation demands. There are, however, areas of engineering research that rely on subjective assessments, which makes strict adherence to "formal, rigorous and quantitative" validation problematic. Science progresses, according to Thomas Kuhn (1970), when the ruling paradigms cannot provide adequate explanations to scientific problems under investigation and this inadequacy makes way for new paradigms. The inadequacy of objective methods in validating non-empirical environments gave rise to a new validation paradigm – the subjective method. Subjective validation, often used in knowledge based systems and simulation modeling (Sargent, 1999; Bawcom, 1997; Pederson et al., 2000) utilizes conversational, contextual and subjective validation. When observed data does not exist this method must obligatorily be used (Braga, unknown).

Pederson, et al. (2000) prefer to think of objective and subjective validation in terms of epistemological views of knowledge. The logical empiricist validation (objective) is strictly formal, algorithmic, reductionist and a "confrontational process" where new knowledge is either true or false. This view asserts validation is more a matter of formal accuracy than practical use and works well for closed problems such as mathematical postulates or algorithms (Preece, 2001). The relativist validation (subjective) is a semiformal and conversational process, where validation is assessed as the confidence in usefulness of the new knowledge with respect to a purpose (Sargent, 1999). This approach is appropriate for open problems where new knowledge is associated with heuristics and non-precise representations (Pederson, et al., 2000). The validation method for the current research case is an adaptation of a method from the relativist validation paradigm referred to as the Validation Square– a method developed, deployed and validated by Pederson et al.

The Validation Square is a validation method designed to evaluate the effectiveness and efficiency of a research method based on qualitative and quantitative measures. The Validation Square is represented in Figure 7.

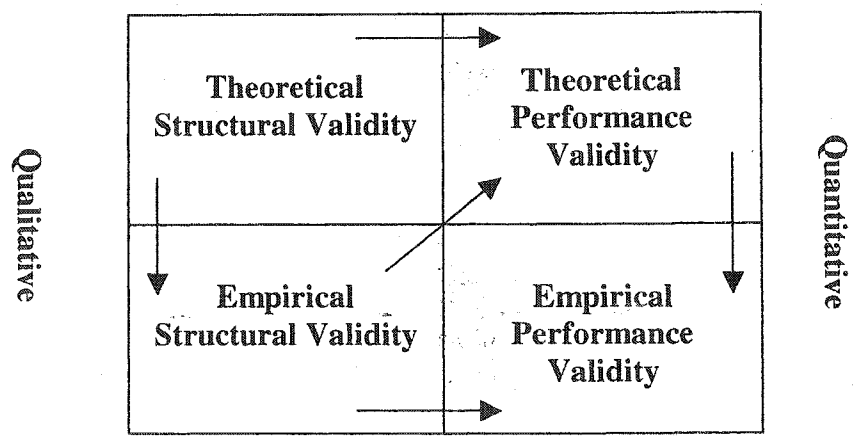


Figure 7: Validation Square (Pederson et al., 2000)

Structural validation is a qualitative process evaluating the effectiveness of the research method at three levels: (1) accepting the individual constructs constituting the method; (2) accepting the internal consistency of the way the constructs are put together in the methods; and (3) accepting the appropriateness of the case study to be used to verify the performance of the method. Theoretical structural validity denotes the "soundness" of the method in a general sense while the empirical structural validity denotes the "soundness" of the method in the applied sense.

Performance validation is a quantitative process evaluating the efficiency of the research method at three levels: (1) accepting that the outcome of the method is useful with respect to the initial purpose; (2) accepting that the achieved usefulness is linked to applying the method; and (3) accepting that the usefulness of the method is beyond the case study.

Performance validation as described by Pederson, et al. 2000 reflects knowledge that can be articulated with statistical grounding, meaning the results of the research method can be evaluated comparatively with a "measure of improvement" allowing a quantitative evaluation. Empirical performance validity refers to the usefulness of the method for some limited case while theoretical performance validity evaluates the usefulness of the method beyond the specific test application. Sargent (1999) offers an argument to the quantitative nature of the performance validation process. Sargent supported by Preece (1994) asserts quantitative evaluation is not always possible in a research domain and thus performance validation must be extended to allow for qualitative measurements. Following this postulate, the current research case will adopt a qualitative evaluation of performance validation.

The Validation Square makes reference to the use of questionnaires and interviews but does not explicitly define how the structural validity and performance validity are to be evaluated. Pederson et al. (2000) do, however, assert that the validation enablers should be consistent with the context from which the application case is embedded. A framework has been established of what needs to be validated in the research methodology, the question now becomes how to validate. Researchers (Sargent, 1999; Hyrkas, Appelqvist-Schmidlechner & Oksa, 2003; Jensen, Klee, & Groenvold, 2002; Preece, 2001) have successfully used panels of experts as validation approaches. Both structured and unstructured interviews (Sargent, 1999; Preece, 2001) as well as diagnostic questionnaires (Hyrkas, Appelqvist-Schmidlechner & Oksa, 2003; Jensen, Klee, & Groenvold, 2002) have been successfully used. The research validation method for the current study will follow the validation success of Sargent and Preece and use an unstructured interview process to determine methodology validity.

The validation method set forth in this research incorporates the principles of the Validation Square and adds another tier to the validation process – content validity of the elicitation instrument. The data acquisition instrument is a vital component to this research therefore the validation process would not be complete without a component to address instrument validity. As discussed, the questionnaire used in this research is an extension of the questionnaire used by Monroe (1997). Monroe validated his elicitation instrument through the use of an independent expert panel whose members evaluated the instrument asynchronously through the use of a questionnaire. Since the Monroe questionnaire was modified and incorporated in this research, re-validation of the questionnaire is prudent.

Hyrkas, Appelqvist-Schmidlechner and Okas (2003) used an expert panel interview process to validate their questionnaire in terms of content validity in a clinical supervision setting. Content validity expresses how well the instrument (questionnaire) represents the content domain being applied. Content validity can be assessed through face validity and expert assessment. Face validity is asking people knowledgeable about the system whether the model or its behavior is reasonable (Sargent, 1999). Expert assessment is achieved by asking experts to review the content of the instrument either qualitatively or quantitatively. For the current research, the validation paradigm will adopt qualitative content validity as the validation process for the data collection instrument.

The validation methodology developed herein adopts principles from the Validation Square philosophy specifically structural and performance validation and couples them with a data acquisition instrument content validation process to create a more holistic validation paradigm. The validation of the aggregation methodology for risk assessment using expert elicitation is therefore, tri-fold. First, performance validity of the methodology is evaluated

at the three levels defined in the Validation Square. An expert panel consisting of the design managers in an unstructured interview process (performed individually) qualitatively comment on (1) the acceptability that the outcome of the method is useful with respect to the initial purpose; (2) acceptability that the achieved usefulness is linked to applying the method; and (3) the usefulness of the method is beyond the case study. Second, the aggregation methodology itself must be proven to be compatible with conceptual vehicle design environments. Assessing the structural validity of the aggregation method is not empirically possible. There does not exist a validation data set from which to compare the results of the aggregation to an observed data point. The objective is to determine the usability and compatibility of the methodology based upon the confidence of design managers in its output. Therefore the aggregation method is determined defensible when it is shown the combination result is reproducible, accountable, subject to peer review, and unbiased. Structural validation is performed in an identical manner as performance validation – individual, unstructured, conversational interviews. Lastly, content validity of the data collection instrument is necessary. Evaluation of face validity through expert assessment is intended to determine that the questionnaire is precise, reproducible, and accountable. Figure 8 represents the validation triad.

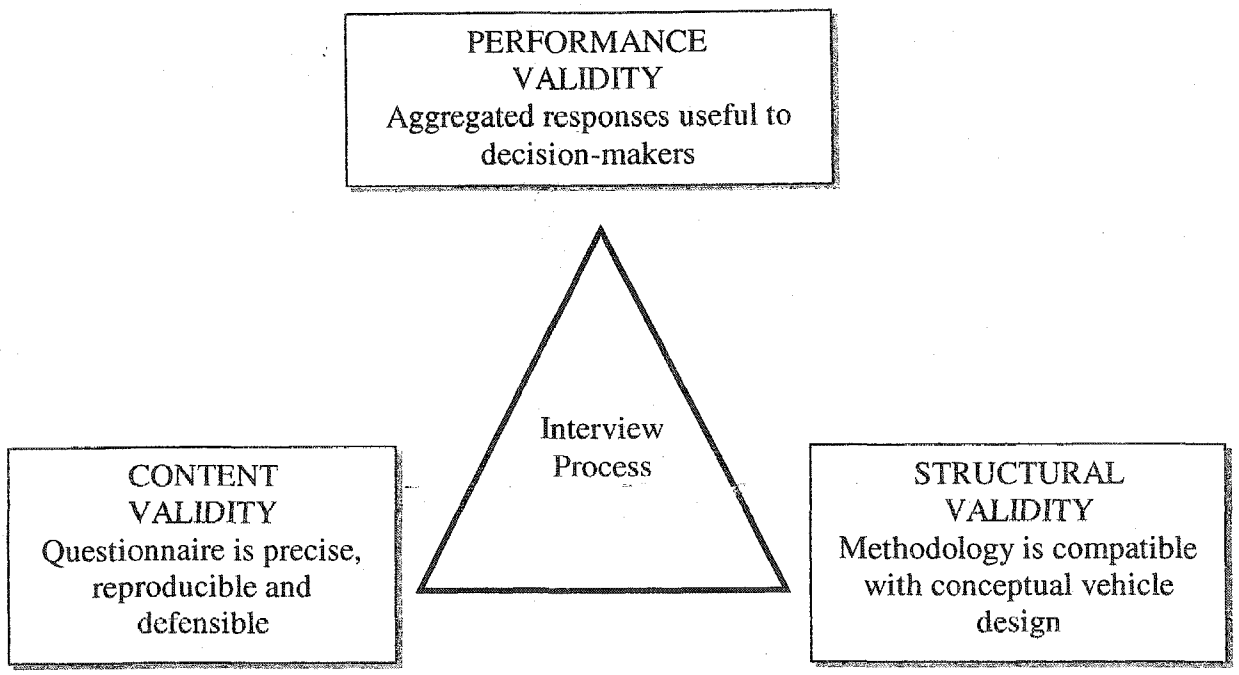


Figure 8: Validation Triad

Although the validation interview process is unstructured and conversational in nature, three specific areas are identified as necessary to be covered to ensure sufficient capture of the elements of the validation triad (see Table 3). The interview process will take place between each member of the expert panel (design team) and the researcher. The researcher will guide the discussion of the interview to capture the minimum elements to satisfy the validation triad. The design managers are given the opportunity to comment on any aspect of the methodology, however the researcher must ensure that at minimum the elements listed in Table 3 are covered.

Performance validity	▪ Feedback on which uncertainty representation they find most useful in their decision-making strategies ▪ Usefulness of method is beyond case study
Structural Validity	▪ Usability and value added of an aggregated response to decision strategies ▪ Applicability of method beyond test case
Content validity	▪ Ease of use of the questionnaire instrument ▪ Appropriateness of questionnaire structure and scaling method chosen ▪ Comprehension of content and context of questionnaire

Table 3: Elements of validation interview

The only material the design managers are given in the interview is a detailed report containing individual uncertainty distributions for discipline specific design variables as well as the aggregated distributions of the combined assessments. The uncertainty distributions and aggregated response representations are necessary to evaluate the performance and structural validity components of the validation methodology. As each expert panel member reviews the documentation and is guided to provide responses to the elements of the validation triad, the researcher captures the responses and documents the comments. If the majority of design managers assert the aggregated responses provide value added to their decision strategies and the aggregation methodology is appropriate to the conceptual domain, the aggregation methodology is considered validated

Issues and Challenges to Research Methodology

The current research case has encountered some issues and challenges that could impact its effectiveness in dealing with large-scale systems. One challenge is the sparse population of identifiable experts in the respective disciplines. The uniqueness of conceptual aerospace design and the very specific knowledge required to perform aerospace analysis

heavily constrains the acceptable population pool. It may be difficult to find more than one experts from which to elicit expert opinion in a discipline. This limitation is addressed in the work of Monroe (1997) and Hampton (2001).

An additional challenge to this research methodology is the inability to empirically validate the "correctness" of the aggregated response. The validation methodology developed herein assesses the usefulness of the aggregated response to decision-makers – it does not determine the correctness of the response. Being able to observe and/or compare the aggregated response to a data point would certainly enhance the strength of this methodology. Morris (2003a), however, acknowledges the near impossibility of this observance and thus asserts this is indeed the "heart of the problem".

A third, and perhaps most daunting issue regarding this research, is the lack of understanding by some in the research community on the difficulties of working in a data vacuum. Many researchers and scientist have difficulty believing that domains exist that do not have perfectly precise and completely available information and thus find it hard to give credibility to research based on subjective data. Baecher (unknown) identifies two elements of research credibility; research must be traceable and defensible. In data rich environments, tracebility is to historical information or empirical statistics and defensibility comes from the postulates, algorithms and laws of physics employed in the research. In conceptual domains (domains with little to no data) tracability refers back to the experts providing the data (assessments) and defensibility refers to the documented assumptions that go into the expert assessments. It is difficult to overcome the barriers of perception that credible research must be tied to empiricisms. Research in domains dominated by subjective data is equally credible and subject to the same rigors of proving tracability and defensibility (Morris, 2003a).

CHAPTER IV

DEMONSTRATION OF METHODOLOGY

Overview

Working with program management from NASA Langley Research Center, two aerospace vehicle disciplines and a concept design vehicle were selected for the application of the aggregation methodology. The development and deployment of an aggregation methodology was in conjunction with a larger research effort incorporating expert judgment elicitation and calibration research. In the larger context, an expert judgment elicitation methodology including background data on experts for the purpose of calibration has been developed. The requirements specified by the Institutional Review Board for the protection of experimental subjects were achieved through careful design and deployment of the questionnaire instrument. The current mechanism for distribution of the data acquisition instrument (questionnaire) is through electronic mail however; the questionnaire is capable of being administered via the World Wide Web.

The vehicle chosen for application of the aggregation methodology is a Two-Stage-To-Orbit (TSTO), staged at Mach 3, conceptual vehicle. This vehicle has a First Stage Booster and a Second Stage Orbiter, both stages having highly uncertain design variables. Two disciplines were chosen to elicit uncertainty assessments for input variables, Weights and Sizing and Operations Support. The Weights and Sizing discipline had 2 identified subject matter experts who were selected by the design managers based upon the expertise criteria outlined. The Operations Support discipline had 3 identified subject matter experts based upon the same expertise criteria.

Weights and Sizing Case Description

NASA Langley Research Center utilizes a Configuration Sizing program (CONSIZ) to size a vehicle and determine the weights of subsystem components. CONSIZ is a program developed specifically at Langley Research Center and provides capability of sizing and estimating weights for a vehicle based upon Weight Estimating Relationships (WERs) derived from historical regression using Shuttle data, finite element analysis and technology readiness level. The CONSIZ program has a predefined initial list of user-defined parameters which make up the input variables to the program. Additional design parameters are necessary to run a CONSIZ model but are provided as "pass through" variables from other discipline applications. For the TSTO staged at Mach 3 vehicle; the Booster has 104 input variables, 54 of which are user defined and the Orbiter has 109 input variables – 58 user defined. A complete listing of the user defined input parameters for the TSTO – Mach 3 vehicle is located in Appendix B.

Operations Support Case Description

For Operational Support analyses, NASA Langley Research Center utilizes a Reliability and Maintainability Analysis Tool (RMAT) to calculate and assess vehicle maintenance burden, ground processing times and manpower requirements for conceptual vehicles. The underlying algorithms for the RMAT computations are regression models built from historical aircraft maintenance data and extrapolations to meet technology readiness level. RMAT is a complex, stand-alone, operational analysis code requiring expert user inputs (Unal, 2002). RMAT utilizes over 200 user defined input variables for an analysis and like CONSIZ, RMAT performs analysis on each element of the vehicle configuration.

Questionnaire Design, Implementation and Deployment

Discipline design team managers agreed that to ensure efficiency in questionnaire content and to reduce the time burden for the experts to provide uncertainty assessments, only those design inputs having the most impact on vehicle performance or operational support need to be queried. To this end, a modified Nominal Group Technique (NGT) was utilized to elicit the most highly uncertain design parameters for each discipline. As discussed, the Nominal Group Technique requires that participants meet in the same location and rank order alternatives synchronously. The modified NGT developed for this research does not require experts to evaluate and rank alternatives synchronously but allows them to rank alternatives at their workstations. The tally of alternatives for the modified NGT is identical to the NGT; the only deviation from common procedure is the allowance of experts to rank alternatives in the privacy of their own workspace.

The modified NGT was implemented by listing all discipline specific user inputs (112 for Weights and Sizing, 200 for Operations Support) in an Excel spreadsheet. Each discipline design team member was given an electronic version of the Excel spreadsheet along with the instructions presented in Figure 9.

Instructions for Classification of Parameter Impact

1. Please examine the list of 'user input" variables provided.
2. On a scale from 1 to 5 (5 least significant – 1 most significant) please rate each of the variables according to your assessment of impact significance on performance characteristics.

Figure 9: Instructions for Parameter Impact Classification

Once the design team managers completed the rankings, the results were tallied. Design team managers supported the use of the Pareto Principle as the discretionary paradigm from which the tally list would be reduced therefore, adhering to the Pareto Principle, the 20 percent of responses deemed to have the most impact on vehicle design or operational support were selected for inclusion in the uncertainty expert elicitation questionnaire. The reduced parameter list for the Booster and Orbiter for each discipline is given in Appendix C.

From the reduced list of input parameters, discipline specific questionnaires were constructed. The questionnaires were electronically mailed to each subject matter expert with a request to complete the questionnaire and return it electronically to the researcher within five days of receipt. The five-day time limit was simply the discretion of this researcher but it did afford the experts sufficient time to complete the questionnaire and not feel too hurried. Each expert, working independently was asked to evaluate each design variable for uncertainty. The expert was first asked to rate the degree of uncertainty associated with the design parameter based on a qualitative 5 unit rating scale (Low, Low/moderate, Moderate, Moderate/high, High). Next, the expert was asked to evaluate the nominal value provided for the variable – the expert could accept this nominal value or provide a value he/she believed more appropriately represented the nominal value for the test case parameter. Additionally, the expert was given an opportunity to establish a non-symmetrical distribution around the nominal value if he/she felt it appropriate.

The expert was next asked to describe the reason for the uncertainty rating for the parameter and the resultant parameter ranges if they were modified. The expert was asked to provide rationale for those parameter values that were altered. Additionally, the expert was

asked to provide any other cues or insights into his/her logic and record that information in the block provided.

Once all input parameters had been assessed, the expert was given the opportunity to add parameters not shown which he/she believed to have a level of uncertainty associated with their value. After rating all input parameters, the experts were asked to anchor their qualitative measure of uncertainty to a quantitative value using the 5-point scale provided.

Institutional Review Board Considerations

Questionnaires were developed for two disciplines for this study: Weights and Sizing and Operations Support. The questionnaires and questionnaire application process utilized in the supporting NASA study were reported to the Institutional Review Board (IRB) representatives at Old Dominion University. The IRB concluded that this research would qualify for an exemption from full IRB procedures for human subject research based on the output from the questionnaires not being harmful or damaging (civil or criminal liability, financial and/or employment implications) in any way to the subject participants.

Expert Judgment Data Collection

The experts provided uncertainty assessments via Microsoft Excel® spreadsheet questionnaires. For each design parameter assessed for uncertainty, the nominal value represented the "most likely" value for the parameter. If the expert assumed a symmetrical distribution about he nominal value, the minimum and maximum values were calculated from the expert's quantitative assessments of uncertainty associated with their qualitative value they assigned to the design parameter. If an asymmetrical distribution was assessed, those values were read directly from the respondents values provided in the questionnaire. The triangular distributions resulting from the "raw" data coupled with the outputs from the

Expert Calibration Function applied to the distributions as part of the Conway (2003) research, resulted in the calibrated distributions that were inputs to the aggregation algorithm.

Calibrated Assessments

The application of the Expert Calibration Function (ECF) is a result of complementary research performed by Conway (2003). Conway derived an ECF by analyzing the responses in the Background section of the questionnaire. Elements of the ECF were subsequently applied to the uncertainty distributions. The development and deployment of an ECF for this research enhances the credibility of the aggregated responses by reducing the variability in the input distributions themselves. The application of the ECF yields uncertainty levels around estimates that are more consistent with an expert's experience and risk philosophy. While the calibrated assessments are inputs into the aggregation process, the ECF is not part of the aggregation methodology per se, therefore will not be expounded on further. However, building upon the work of Conway (2003) the calibrated distributions which are used as inputs to the aggregation algorithm are provided in Appendix D.

Aggregation Process

Prior to the aggregation of multiple assessments using the linear opinion pool method, weighting factors (expert credibility) are assessed for each subject matter expert. The discipline specific design managers (often the elicited experts themselves) were asked to provide credibility assessments for the discipline experts providing the uncertainty assessments. For Weights and Sizing and Operations Support the design managers and the elicited experts were synonymous. A facilitator interviewed each discipline specific design

manager separately and asked for his/her credibility ranking for each of the experts elicited for responses. The results of the interview process follow (Tables 4 and 5):

	Design Manager A	Design Manager B
Expert A	0.30	0.30
Expert B	0.70	0.70

Table 4: Weighting Factors for Weights & Sizing

	Design Manager A	Design Manager B	Design Manager C
Expert A	0.40	0.40	0.40
Expert B	0.40	0.40	0.40
Expert C	0.20	0.20	0.20

Table 5: Weighting Factors for Operations Support

Calibrated distributions are imported into the @RISK® software in basic form (minimum, most likely, maximum values) and triangular distributions are built for each variable assessed for uncertainty using the "RiskTriang" function. An example using Operations Support VAR01 is provided for illustrative purposes in Table 6 and Figure 10.

Variable	Var Name	Nom Value	
VAR 01	Sched Hrs	114750	
		MOST	
Expert	MIN	LIKELY	MAX
Expert A	40,000.0000	120,487.5000	125,000.0000
Expert B	90,000.0000	135,460.3618	140,000.0000
Expert C	90,000.0000	100,000.0000	110,000.0000

Table 6: Operations Support VAR01 Calibrated distribution

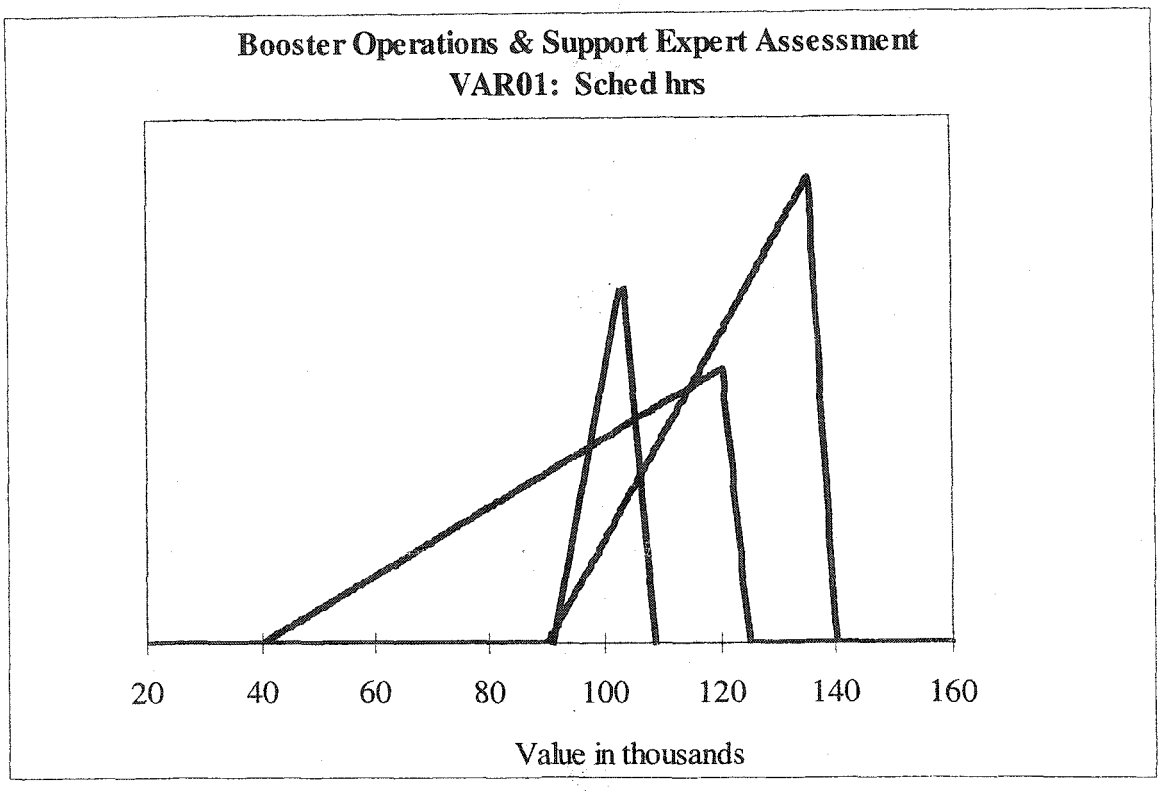


Figure 10: Operations Support VAR01 assessments

Weighting factors, as provided by the decision managers, are applied to each of the expert's assessments (Table 7).

	Variable	Var Name	Nom Value	
	VAR 01	Sched Hrs	114750	
Weighting factor	Expert	MIN	MOST LIKELY	MAX
0.40	Expert A	40,000.0000	120,487.5000	125,000.0000
0.40	Expert B	90,000.0000	135,460.3618	140,000.0000
0.20	Expert C	90,000.0000	100,000.0000	110,000.0000

Table 7: Calibrated distributions with Weighting Factors

A linear opinion pool aggregation algorithm (Equation 6) is coded into a separate input cell.

Aggregated Weighted Distribution
= RiskTriang(40,000, 120487.5,125000)*0.40 + RiskTriang(90000,135460.3618,140000)*0.40 + RiskTriang(90000,100000,110000)*0.20

Equation 6: Linear Opinion Pool Equation

The result of executing this equation is simply an aggregated "most likely" value of the combined distribution. For a more robust answer and to determine the minimum and maximum values of the distribution, a simulation approach is necessary. To perform simulation on the expert assessments, the addition of the variable "RiskOutput" to the aggregation equation is required and should not be considered part of the linear opinion pool equation itself (Equation 7).

Aggregated Weighted Distribution
= RiskOutput("O&S Booster VAR01") + RiskTriang(40,000, 120487.5,125000)*0.40 + RiskTriang(90000,135460.3618,140000)*0.40 + RiskTriang(90000,100000,110000)*0.20

Equation 7: @RISK® Simulation Equation

To run a simulation, a variety of setting may be used to control the type of simulation @RISK® performs. A simulation in @RISK® supports unlimited iterations and multiple simulations (Palisade, 2002). The "Simulation Settings" module allows you to specify the number of iterations to run as well as whether you want to use Monte Carlo or Latin Hypercube sampling. For this research case, 4 iterations were run (500, 1,000, 10,000, 15,000) using the Latin Hypercube sampling technique. The number of iterations chosen

was simply to monitor convergence of results and to compare aggregated values at different sampling iterations. Since increased iterations increases accuracy of simulated results (Vose, 2000), the results from running 15,000 iterations are evaluated.

Data Analysis - Aggregated Responses

The numerical results of performing the aggregation process on the calibrated assessments are summarized in Appendix E. Minimum, most likely, maximum, and mode are represented. The resultant aggregated distributions are formulated by the module BESTFIT® which runs goodness-of-fit algorithms to determine the most compatible distribution that best represents the sampled data. As discussed earlier, the selected distribution from the simulated aggregation process is based upon the Chi-Square statistic. A graphical illustration with test statistics of an aggregated result follows (Figure 11 and Table 8):

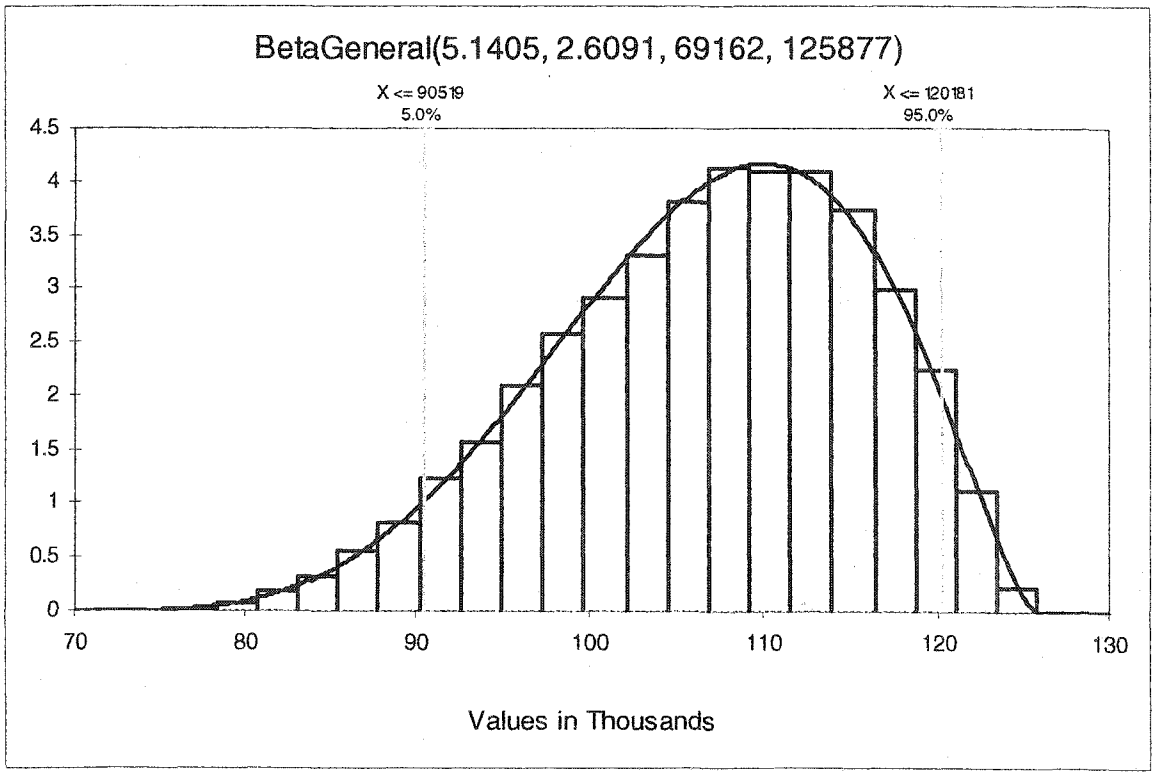


Figure 11: Operations Support VAR01 Aggregated results

Summary Data	
FIT	Beta Distribution
α1	5.153360407
α2	2.646831156
min	69354.10984
max	126007.711
Left X	90579
Left P	5.00%
Right X	120191
Right P	95.00%
Diff. X	2.96E+04
Diff. P	90.00%
Minimum	69354
Maximum	126008
Mean	106784
Mode	109922
Median	107595
Std.	9042.4
Variance	81765319
Skewness	-0.4109

Chi Square Statistics	
Test Value	107.5
P Value	0.0675
Rank	1
C.Val @ 0.75	77.7774
C.Val @ 0.5	86.3342
C.Val @ 0.25	95.4972
C.Val @ 0.15	100.6695
C.Val @ 0.1	104.275
C.Val @ 0.05	109.7733
C.Val @ 0.025	114.6929
C.Val @ 0.01	120.591
C.Val @ 0.005	124.7177
C.Val @ 0.001	133.5121
# Bins	88

Table 8: Operations & Support VAR01 Summary Data

Uncertainty assessments were queried for 33 variables (10 from Weights and Sizing and 23 from Operations Support) and the aggregation methodology applied to the calibrated distributions. Duplicate variables were not aggregated if their values were identical for the Booster and Orbiter. Determining the most appropriate distribution to fit the aggregated responses resulted in 25 variables being most compatible with the Beta distribution, 2 variables best represented with the Triangular distribution and 1 aggregated variable best represented with a Weibull curve. A summary of the variables and the best-fit aggregated distributions are in Appendix F. Appendix G shows graphical representations of the aggregated responses with associated test statistics.

CHAPTER V

RESEARCH FINDINGS

A summary of the BESTFIT® aggregated responses is represented in Table 9.

Variable	Distribution	DF	Fit Statistic	Fail to Reject Reject H_o	p-value
Weights & Sizing					
Booster VAR02	Beta	73	74.14	Fail to reject	0.4408
Booster VAR03	Beta	87	86.49	Fail to reject	0.4954
Booster VAR08	Beta	87	108.7	Fail to reject	0.0577
Booster VAR09	Beta	87	68.94	Fail to reject	0.923
Booster VAR10	Beta	87	88.44	Fail to reject	0.4367
Booster VAR11	Beta	87	92.89	Fail to reject	0.313
Orbiter VAR03	Beta	87	76.65	Fail to reject	0.7784
Orbiter VAR12	Beta	87	64.78	Fail to reject	0.9643
Orbiter VAR13	Beta	87	94.89	Fail to reject	0.2646
Orbiter VAR14	Beta	87	92.89	Fail to reject	0.313
Operations Support					
Booster VAR01	Beta	21	10.84	Fail to reject	0.9658
Booster VAR04	Beta	87	93.57	Fail to reject	0.2958
Booster VAR05	Beta	87	85.11	Fail to reject	0.5372
Booster VAR06	Beta	87	69.73	Fail to reject	0.9125
Booster VAR07	Beta	87	78.58	Fail to reject	0.7289
Booster VAR08	Beta	87	87.51	Fail to reject	0.4646
Booster VAR09	Beta	87	98.54	Fail to reject	0.1871
Booster VAR10	Beta	87	86.65	Fail to reject	0.4904
Booster VAR12	Triangular	73	258.2	Reject	0
Booster VAR14	Beta	87	93.41	Fail to reject	0.2999
Booster VAR15	Beta	87	69.34	Fail to reject	0.9178
Orbiter VAR01	Beta	87	163.9	Reject	1.21E-05
Orbiter VAR06	Beta	87	106	Fail to reject	0.0817
Orbiter VAR09	Triangular	87	113.1	Reject	0.0316
Orbiter VAR10	Weibull	87	81.7	Fail to reject	0.6404
Orbiter VAR11	Beta	87	79.67	Fail to reject	0.6989
Orbiter VAR14	Beta	73	69.92	Fail to reject	0.5805
Orbiter VAR15	Beta	87	104.9	Fail to reject	0.0928

Table 9: Summary of Chi-Square Test Statistics

The majority of fitted data mapped quite well to a beta distribution. The beta distribution is very flexible and useful and can be used when a variable is bounded by two limits. The normal distribution is valid between $-\infty$ and $+\infty$ and the lognormal distribution is valid between 0 and $+\infty$. The beta distribution has long been the distribution of "choice" for subjective assessments (Haldar & Mahadevan, 2000). Because the beta distribution is bounded on both sides, it is often used for representing processes with natural lower and upper limits (Hahn & Shapiro, 1967). The beta function is characterized by two parameters $\alpha 1$ and $\alpha 2$ representing the scale parameter and shape parameter respectively. When $\alpha 1 = \alpha 2 = 1$ the beta distribution essentially becomes a uniform distribution. When $\alpha 1 > \alpha 2$ the distribution becomes skewed towards higher values; essentially mimicking a right skewed triangular distribution. When $\alpha 1 < \alpha 2$ the distribution becomes skewed towards lower values essentially becoming a left skewed triangular distribution. If $\alpha 1$ is set equal to $\alpha 2$, the distribution is symmetric around the mean (Garvey, 2000). Because the beta function can take a wide variety of shapes, the beta distribution is among the most diverse and flexible distribution (Haldar & Mahadevan, 2000). The versatility of the beta distribution and its compatibility with subjective assessments supports the results that 26 out of 29 variables aggregated resulted in a beta distribution fit.

The triangular distribution was a best fit match to two operations support variables (Booster VAR12 and Orbiter VAR09). The triangular distribution is the most commonly used distribution for modeling expert opinion due to its simplicity and is defined in terms of minimum (a), most likely (c) and maximum values (b) (Molak, 1997). The location of c relative to a and b determines how much probability there is on either side of c. The closer

the mode is to the variables maximum possible value *b*, the less likely it is the variable will exceed the mode. The closer the mode is to the variables minimum possible value *a*, the more likely the variable will exceed its mode (Garvey, 2000). In examining possible clues as to why these two variables deviated from the pattern of beta distribution, a look at the input distributions reveals some insight. Both input distributions for Booster VAR12 and Orbiter VAR09 look remarkably similar. Both are heavily skewed with one distribution significantly dissimilar from the other two assessments. This "outlier" distribution is significant enough to skew the goodness of fit away from the beta distribution and cause enough discontinuity to reject the null hypothesis.

Lastly, one variable, operations support VAR10 was a best fit match to the Weibull distribution. The Weibull distribution is often used to represent distributions of failure time in reliability models and is similar in shape the beta distribution (Morgan & Henrion, 1990). The Weibull function, like the beta, has a shape parameter (*a*) and a scale parameter (*b*). It is unclear why this particular variable has a better fit to a Weibull rather than a beta distribution but it is interesting to note the goodness-of-fit statistic indicates the beta distribution is the second most "best-fit" curve of the 22 distributions examined.

Validation

Validation of the aggregation methodology is a one-on-one unstructured interview process consisting of a three-part construct. Content validity is assessed by the decision-makers in regards to ease of use of the questionnaire instrument, appropriateness of the questionnaire structure to the problem domain and comprehension of content and context of the questionnaire. Structural validity is assessed in regards to usability and value added of an aggregated response to decision strategies and the applicability of the method beyond the test

case. Lastly, performance validity is based upon feedback from the decision-makers on which uncertainty representation they find most useful in their decision strategies.

The discipline design managers were separately interviewed and allowed to discuss any aspect of the methodology. This researcher guided the discussion when necessary to ensure the minimum requirements listed in Table 3 were covered. The researcher recorded all responses and comments. The design managers unanimously agreed the questionnaire content, structure and deployment was efficient, well developed and user friendly. Each concurred the questionnaire could easily be applied to other test cases and successfully captured the qualitative and quantitative uncertainty of design parameters. Two specific suggestions for improving the questionnaire were reported by the design managers. The design managers suggest that instead of having the subject matter experts tab through the variables on Microsoft Excel® worksheet tabs, automate the process so that when the experts finish assessing the uncertainty of a variable, the sheet immediately scrolls to the next variable. Additionally, the design managers would like to see the questionnaire transported into the World Wide Web environment. Both of these suggestions have been previously identified in the larger context research from which this methodology is embedded and are to be implemented as an extension of the larger research.

Determining whether the aggregated responses are useful to decision-makers (structural validity) and which uncertainty representation decision-makers find most useful in decision strategies (performance validity) is a completely subject assessment. Design managers were presented 3 representations of the aggregated data – aggregated numerical values of minimum, most likely and maximum, as presented in Appendix E; expert calibrated assessments layered onto one graph as presented in Appendix G; and the BESTFIT®

aggregated responses represented in Appendix G. The discipline design managers were asked to review each of the representations and comment. One design manager believes all three representations are valuable in that each presents a different interpretable reference to the aggregated data. The numerical representation (Appendix E) provides more of a discrete value to the aggregation while the representation of layered calibrated assessments allows the design manager to assess the level of agreement between the experts. The BESTFIT® aggregated representation is valuable "when you are ready to implement a decision". Another design manager focused in more on the aggregated graphical representation and responded this representation gave him a better idea of where to go with his decision strategies instead of relying on "heavy interpolation". In the design manager's opinion, the aggregated response provides mathematical validity to the "eye-ball" method which is so prevalent in conceptual design. Additionally he felt the methodology took "amalgamous data and transformed it into something useful".

Each design manager also commented on the usefulness and applicability of this methodology beyond the test case. One of the operations support design managers hopes to extend this methodology to a full scale conceptual vehicle, integrating each discipline necessary to develop a space transport concept. A weights and sizing design manager asserts this methodology is transportable to planetary exploration projects and would like to see this methodology employed in other projects within the Vehicle Analysis Branch at NASA Langley Research Center.

CHAPTER VI

DISCUSSION

As technology systems continue to evolve in complexity and conceptual designers seek to stretch the limits of feasibility of engineering design, strategies to holistically capture and represent risk associated with such highly uncertain and high consequence enterprises becomes paramount. In order to quantify risk with confidence, better quantification strategies need to be developed (Proffitt, 2003). Coupled with improved quantification strategies is the need for more robust combination methods when multiple uncertainties for the same variable have been quantified. Many aggregation methods exist that combine expert opinions when past data is available from which to ascertain likelihood functions and expert credibility. The current research case is embedded in a domain in which these two factors are absent and, therefore, a combination method which does not rely on likelihood functions and with an alternate way to determine expert credibility has been developed.

The development of this aggregation methodology for uncertainty assessment required the integration of many elements; a thorough understanding of the research domain and uncertainties under investigation, the development of an efficient, relevant and appropriate data elicitation mechanism, an aggregation approach compatible with the type of uncertainty being investigated and the decision model chosen, and the validation methodology had to be congruent with a subjective research paradigm.

The aerospace conceptual design environment is a unique domain in that there is very limited "hard" data to base decisions on and the preponderance to fully know the outcome of an event is difficult. Aerospace conceptual designs are normally initiated 20-30 years from the time the vehicle needs to be in service and the technology projected to be incorporated

into these new designs is unique and revolutionary. For this very reason, the values of design variables and the uncertainties associated with those values is considered unknown. These values are not completely unknowable since in 20-30 years the vehicles should come to fruition, however at the time of a decision milestone the values are not known quantities. This distinction is important in understanding the research domain and uncertainties under investigation in this study.

The data elicitation mechanism used in the current case was a modification of a questionnaire developed by Monroe (1997). The Monroe questionnaire was chosen as a base model because it had been developed and deployed in the conceptual aerospace environment with success in estimating the Weight Estimating Relationships (WERs) for weights and sizing variables. However, in examining the compatibility of the Monroe questionnaire in the deployment of the aggregation methodology, several issues arose. The structure of the Monroe questionnaire allowed the experts to contradict their own assessments of uncertainty and many of the participants of that study did not find the questionnaire time efficient. None of these complications were too great to overcome but a modification of the questionnaire was necessary to ensure time efficiency and prevent expert contradiction of their own assessments. A strength of the questionnaire for application to the current research is the elicitation of values in triangular distribution form. Triangular distributions work remarkably well when knowledge of the "shoulders" of a distribution is unclear. A weakness of the Monroe structure however, was that he disallowed skewed triangular distributions; the structure of the questions forced experts to give a symmetrical answer around the mean. For this study, experts must be allowed to provide skewed distributions to properly capture the uncertainties of a variable of interest. Physical limitations of a variable may constrain an

extreme value (minimum or maximum) thus impacting the uncertainty associated with a mean value. The questionnaire was modified to allow for skewed distributions to acknowledge possible physical limitations of design variables.

Relating to the uncertainty assessments of the experts is the question of how to assign credibility factors to each expert's assessment in the absence of historical data or observable outcomes. The expert weighting factors are a critical component to the opinion pool aggregation algorithms. Some researchers assign credibility factors based upon years of experience or educational level but as the literature indicates, these characteristics are not necessarily qualifiers for expertise. For this research study, it was determined that eliciting credibility factors from the design managers themselves, would be the most appropriate method to gather these factors. This method of subjectively ascertaining credibility factors would incorporate all the relevant characteristics of expertise (see Chapter III, p. 43) and provide the most comprehensive weighting of the experts judgments.

The execution of the linear opinion pool aggregation algorithm provides a "most likely" value of the combined distributions. For a more robust representation of the combination algorithm, simulation sampling was incorporated. The Latin Hypercube sampling technique was run with 500, 1000, 10000, and 15000 iterations; 15000 iterations was the convergence point of the majority of variables. It is interesting to note that on almost all variable sampling, the smaller iterations resulted in either normal or lognormal distributions while the higher iterations (10000 and 15000) converged to a beta distribution. It would be interesting to extend this research to investigate the correlation between input distributions and output distributions and to hypothesize on the relationship between the two.

Perhaps the biggest challenge in the development of the present research was determining a means to validate the aggregation methodology. The distant-future nature of aerospace technology impact rendered classic (objective) validation techniques moot. A means to subjectively validate the usefulness of the aggregated responses to decision-makers was necessary. To access how useful decision-makers found the aggregated responses, a validation technique that allowed free exchange of ideas, thoughts, criticisms and opinions was necessary. The use of an unstructured interview process provided greater depth of detail in the validation analysis; much more detail than if a questionnaire instrument were used. When eliciting qualitative assessments in a questionnaire, participants may not provide as detailed a response as a researcher might like. Frary (unknown) discourages open-end questions on a questionnaire because participants are likely to suppress responses to save on time commitment to the process. Conversational, context driven dialogue enabled a more comprehensive evaluation of the usability of the aggregated response to conceptual decision-makers.

CHAPTER VII

CONCLUSIONS

The development and deployment of an aggregation methodology has resulted in a process that permits aggregation of multiple expert opinions into a single consensus distribution. While research and application of aggregation techniques is not new, the development of an aggregation methodology in the absence of likelihood functions and expert credibility assessments is unique. The use of simulation in the aggregation process expands aggregation from a single point outcome, generally a most likely value to the aggregation of distributions which more effectively represent risk and uncertainty. The application of a distribution-fitting tool such as BESTFIT® adds a level of robustness to the aggregation outcome by allowing an analyst to vary distribution types to compare aggregated outcome ranges. This methodology, applied in the aerospace discipline, which is very dynamic and continuously evolving, should prove an effective aid to decision-making associated with aerospace development.

Research Contributions

The primary focus of this research is the aggregation of multiple opinions in conceptual design environments. The contribution of this research falls within the landscape of aggregation application as well as combination methodology. Prior studies in expert elicitation have been performed in environments where historical data or empirical statistics have been available to determine likelihood functions and to assess expert credibility. These two attributes do not exist in the current research case.

The development of an aggregation methodology which does not rely on prior distributions or known expert credibility factors and the demonstration of the methodology in

a highly dynamic and uncertain domain provides decision-makers with a viable decision strategy to reduce uncertainty in conceptual designs. The aggregation methodology developed and demonstrated herein, provides a method capable of improving the robustness of engineering designs, improve cost estimates of conceptual vehicles and abate variability, which leads to risk and uncertainty. By applying this methodology, the decision-maker can select the best, most cost-effective, risk tolerant solutions to provide the greatest long-term benefits.

Study Limitations & Delimitations

The use of expert judgment elicitation techniques should be reserved for those environments where little "historical" data is known about the parameter of interest. Where sufficient data exists or can be feasibly obtained, traditional statistical strategies are preferable to expert opinion.

Another limitation of this research is the variability of output from the aggregation tool used to combine multiple opinions. Uncertainty estimates of design parameters are queried and uncertainty estimates of error associated with discipline specific analysis tools are elicited; however, the uncertainty associated with the variability from the output from the aggregation tool itself is not investigated. It is not the intent of this research to examine the sensitivity of the combination protocol.

Additionally, this research does not investigate the propagation of uncertainty in the combination of expert opinions. The intent of this research is to develop a viable aggregation methodology that sufficiently combines multiple opinions in conceptual design environments. The propagation of uncertainties in combination methods is outside the intent of this research.

Extensions of Research

The present study has great potential for future expansion. In particular, this research has utilized the triangular distribution for its expert elicitation process. This follows previous work in expert elicitation related to aerospace conceptual design environments (Hampton, 2001; Monroe, 1997). Extending the elicitation process to include other representations of uncertainty assessments such as a beta or exponential distribution may enhance the robustness of the aggregation methodology.

As mentioned in the literature review, the logarithmic opinion pool method is a viable mathematical aggregation method to deploy in subjective probability combination schemes. Advancing this research to include the logarithmic opinion pool method and evaluating the fidelity of results with the linear opinion pool method would be an interesting and valuable comparative analysis.

Lastly, extension of the aggregation methodology developed herein could be applied to domains outside the aerospace conceptual design environment. Many decision domains such as military intelligence, automobile manufacturing, national security/terror analysis and medical/pharmaceutical fields consistently deal with highly uncertain, high consequence decisions. Extension of this methodology to other decision domains may serve to demonstrate the methodologies use as a template and add to the generalizability of this research.

REFERENCES

Apostolakis, G. (1994). A Commentary on Model Uncertainty, Model Uncertainty: Its Characterization and Quantification. *NUREG/CP-0138*, U.S. Nuclear Regulatory Commission.

Ashton, R.H. (1986). Combining the Judgments of Experts: How many and Which Ones? *Organizational Behavior and Human Decision Processes*, 38, 405-414.

Ayyub, B.M. (2001). *Elicitation of Expert Opinions for Uncertainty and Risks*. Boca Raton, FL: CRC Press.

Baecher, G.B. (unknown). Expert Elicitation in Geotechnical Risk Assessment. University of Maryland. Retrieved August 24, 2002 from http://www.glue.umd.edu/~gbaaecher/papers.d/COE.d/ExpertElicitation.doc

Barlas, Y. & Carpenter, S. (1990). Philosophical Roots of Model Validation: Two Paradigms. *Systems Dynamics Review,* Vol. 6, No. 2, pp: 148-166.

Bawcom, D.J. (1997). *An Incomplete Handling Methodology for Validation of Bayesian Knowledge Bases*. Master's Thesis, Air Force Institute of Technology.

Beach, B. H. (1975). Expert Judgment About Uncertainty: Bayesian Decision-making in Realistic Settings. *Organizational Behavior and Human Performance*, 14, 10-59.

Beenen, F. (1970). Psychiatric Diagnosis and Subjective Probability. *Acta Psychologica*, 34, 328-337.

Berstein, P. (1998). Risk at the Roots. *Market Leader*, 2, 40 – 43.

Braga, R. P. (unknown). Model Validation Methods – Special Reference to Crop Simulation Models. Retrieved October 25, 2003 from http://www.esaelvas.vas.pt/ recardo_braga/validation2.pdf

Brier, G. (1950). Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review, Vol. 78, No. 1, pp 1-3.

Brun, Weibecke & Tiegen. (1988). Verbal Probabilities: Ambiguous, context dependent or both? *Organizational Behavior and Human Decision Processes,* Vol. 41, No. 3, pp 390-404.

Buchbinder, B. (1993). Risk Management for the Space Exploration Initiative. *AIAA Paper 93-0377, IN: 31st Aerospace Sciences Meeting and Exhibit, American Institute of Aeronautics and Astronautics*, Reno, Nevada, 1993.

Budescu, D.V. & Wallsten, T.S. (1987). Subjective Estimation of Precise and Vague Uncertainties. IN *Judgmental Forecasting*, G. Wright and P. Ayton [eds.], John Wiley and Sons: Chichester.

Bunn, D.W. (1988). Combining Forecasts. *European Journal of Operational Research*, 33, 223-229.

Chen, Kay-Yut, Fine, L.R. & Huberman, B.A. (2003). Predicting the Future. *Information Systems Frontiers*, 5, 47-61.

Christensen-Szalanski, J.J.,&Beach, L.R., (1984). The Citation Bias: Fad and Fashion in the Judgment and Decision Literature. *American Psychologist*, 39, 75-78.

Clemen, R.T., & Winkler, R.L. (1997). Combining Probability Distributions from Experts in Risk Analysis. *Risk Analysis*, 19, 187-203.

Commission on Engineering and Technical Systems, National Research Council (1999). *Upgrading the Space Shuttle*. National Academy Press, Washington, D.C.

Conway, B.A. (2003). *Calibrating Expert Assessments of Advanced Aerospace Technology Adoption Impact*. Ph.D. Dissertation, Old Dominion University, Norfolk, VA. August 2003.

Cornell, A.C. (1996). Seismic Risk Analysis as as Example of Aggregating Expert Opinion. *Proceedings of the Aspen Global Change Institute Summit*, August, Aspen CO, 1996.

Cox, E.P. III (1980). "The Optimal Number of Response Alternatives for a Scale: A Review". *Journal of Marketing Research*, 17, 407-422.

Cyert, R.M.& DeGroot, M.H. (1987). *Bayesian Analysis and Uncertainty in Economic Theory*. Totowa, NJ: Rowman & Littlefield.

Daan, H., & Murphy, A. (1985). Sensitivity of verification scores to the classification of the predictand. *Monthly Weather Review*, 113, 1384-1392.

Dalkey, N.C. (1969). The Use of Self-Ratings to Improve Group Estimates. *Technology Forecasting* 12, 283-291.

Dawid, A.P. (1982). The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, Vol. 77, Issue 379, 605-610.

Delbecq, A.L., Van de Ven, A.H. & Gustafson, D.H. (1975) *Group Techniques for Program Planning, a Guide to Nominal Group Technique and Delphi Processes,* Scott Foreman.

Druzdzel, M.J. (1989). Verbal Uncertainty Expressions: Literature Review. *Technical Report CMU-EPP-1990-03-02*, Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA.

Du, X., & Chen, W. (1999). An Integrated Methodology for Uncertainty Propagation and Management in Simulation Based Systems Design. *American Institute of Aeronautics and Astronautics Journal*, July.

Ebeling, C.E. (1997). *An Introduction to Reliability and Maintainability Engineering*. New York, NY: McGraw-Hill Companies, Inc.

Engemann, K.J., Miller, H.E., & Yager, R.R. (1995). Decision-making Using the Weighted Median Applied to Risk Management. *IEEE Proceedings of The Joint Third International Symposium on Uncertainty Modeling and Analysis-North American Fuzzy Information Processing Society*, September, College Park, MD, 1995

Feltovich, P.J., Spiro, R.J., & Coulson, R.L. (1997). Issues of Expert Flexibility in Contexts Characterized by Complexity and Change. In P.J. Feltovich, K.M. Ford and R.R. Hoffman (Eds.), *Expertise in Context*. Menlo Park: AAAI Press/The MIT Press.

Frary, R.B. (unknown). A Brief Guide to Questionnaire Development. Office of Measurement and Research Service, Virginia Polytechnic Institute and State University.

French, S. (1981). Consensus of opinion. *European Journal of Operational Research*, 7, 332-340.

Fryback, D.G. & Erdman, H. (1979). Prospects for Calibrating Physicians' Probabilistic Judgments: Design of a Feedback System. *Proceedings from the IEEE International Conference on Cybernetics and Society*, pp. 340-345.

Garvey, P.R. (2000). *Probability Methods for Cost Uncertainty Analysis – A Systems Engineering Perspective*. Marcel Dekker, Inc: New York.

Genest, C. & Zikek, J.V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*, 1, 114-135.

Genest, C. & Schervish, M.M. (1985). Modeling expert judgments for Bayesian updating. *Annals of Statistics*, 9, 53-73.

Gilboa, I., Samet, D., & Schmeidler, D. (2001). *Utilitarian Aggregation of Beliefs and Tastes*. Retrieved November 11, 2002, from http://www.math.tau.ac.il/~schmeid/ PDF/ Gil_Sam_Schmeid_Utilitarian_Bayesian.

Gustafson, D.H., R.E. Shukla, A. Delbecq, , and G.W. Walster (1973). A Comparative Study of Differences in Subjective Likelihood Estimates Made by Individuals, Interacting Groups, Delphi Groups, and Nominal Groups. Organizational Behavior and Human Performance, 9, 200-291.

Gu, X., Renaud, J.E., & Batill, S.M. (1998). An Investigation of Multidisciplinary Design Subject to Uncertainties. *AIAA Paper 98-4747. 7th AIAA/USAF/NASA/ISSMO Multidisciplinary Analysis & Optimization Symposium*, St. Louis, MO, 1998.

Hahn, G.J. & Shapiro, S.S. (1967). *Statistical Methods in Engineering*. New York: John Wiley & Sons.

Haimes, Y.Y (1998). *Risk Modeling, Assessment and Management*. John Wiley & Sons: New York.

Haldar, A. & Mahadevan, S. (2000). *Probability, Reliability and Statistical Methods in Engineering Design*. John Wiley & Sons: New York.

Hammitt, J.K., & Shlyakhter, A.I. (1999). The Expected Value of Information and the Probability of Surprise. *Risk Analysis*, Vol. 19, No. 1, pp. 135-152.

Hampton, K.R. (2001). An Integrated Risk Analysis Methodology in a Multidisciplinary Design Environment. Ph.D. Dissertation. Old Dominion University, Norfolk, VA.

Hanke, J.E. & Reitsch, A.G. (1998). *Business Forecasting. 6th Edition*. Prentice-Hall, Saddle River, NJ.

Hardy, T.L. & Rapp, D.C. (1994). Rocket Engine System Reliability Analyses Using Probabilistic and Fuzzy Logic Techniques. *AIAA Paper 94-2750, presented at the AIAA/ASME/SAE/ASEE 30th Joint Propulsion Conference*, Indianapolis, IN, 1994.

Hertz, D. & Thomas, H. (1984). *Risk Analysis and its Applications*. John Wiley & Sons: New York.

Hodge, R., Evans, M., Marshall, J., Quigley, J. & Walls, L. (2001). Eliciting engineering knowledge about reliability during design - Lessons learnt from implementation. *Quality and Reliability Engineering International*, 17, 169-179.

Hogarth, R.M. (1990). Decision-making Under Uncertainty: The Effects of Role and Ambiguity. *Grant No. N00014-84-C-0018.* Perceptual Science Program, Office of Naval Research, Arlington, Virginia.

Hogarth, R.M. (1978). A Note on Aggregating Opinions. *Organizational Behavior and Human Performance*, 21, 40-46.

Hurley, W.J. & Lior, D.U. (2002). Combining Expert Judgment: On the performance of trimmed mean vote aggregation procedures in the presence of strategic voting. *European Journal of Operational Research*, 14, 142-147.

Hyrkas, K., Appelqvist-Schmidlechner, K., & Oksa, L. (2003). Validating an instrument for clinical supervision using an expert panel. *International Journal of Nursing Studies,* 40, 619-625.

Jackson, P. (1999). *Introduction to Expert Systems.* 3rd Edition. Addison-Wesley, Harlow, England.

Janis, I.L., & Mann, L. (1977). *Decision-making.* New York Press: Free Press.

Jensen, P.T., Klee, M., & Groenvold, M. (2002). Validation of a quesionnaire for self-rating of urological and gynaecological morbidity after treatment of gynaecological cancer. *Radiotherapy and Oncology*, 65, 29-38.

Jones, J. & Marshall J. (2000). An Event Based Database for the Support of a Holistic Reliability Assessment Tool. *European Safety and Reliability Conference*, May, Edinburgh, UK, 2000.

Kader, V.A. (1991). Fuzzy Logic: A Key Technology for Future Competitiveness. International Trade Administration, U.S. Department of Commerce, Washington, D.C.

Keeney, R.L. (1977). The Art of Assessing Multiattribute Utility Functions. *Organizational Behavior and Human Performance*, 19, 276-310.

Kowal, M. (1998). Uncertainty Based Multidisciplinary Design Optimization. *AIAA Paper 98-4915, IN: 7th Symposium on Multidisciplinary Analysis and Optimization, American Institute of Aeronautics and Astronautics*, September, St. Louis MO, 1998.

Kuhn, T. (1970). *The Structure of Scientific Revolutions.* Chicago, University of Chicago Press.

Lasky, K.B. (1996). Model Uncertainty: Theory and Practical Implications. *IEEE Transactions on System, Man and Cybernetics*, 26, 340-348.

Law, A. and Kelton, W. (1991). *Simulation Modeling & Analysis*. 2nd Edition, McGraw-Hill, Inc. New York.

Lawrence, M.J., Edmundson, R.H. & O'Connnor, M.J. (1986). The Accuracy of Combining Judgmental and Statistical Forecasts. *Management Science*, Vol. 32, No. 12, pp.1521-1532.

Leedy, P.D. & Ormrod, J.E. (1985). *Practical Research: Planning and Design*. 7th Edition. Merrill Prentice-Hall, New Jersey.

Lichtenstein, S. & Fischhoff, B. (1977). Do Those Who Know More Also Know More About What They Know? Organizational Behavior and Human Performance, 20, 159-183.

Lichtenstein, S. & Newman, J.R. (1967). Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities. *Psychonomic Science*, 9, 563-564.

March, J.G. (1994). *A Primer on Decision-making*. The Free Press, Macmillan: New York.

Mendenhall, W. & Sincich, T. (1995). *Statistics for Engineering and the Sciences*. 4th Edition. Prentice Hall: Upper Sadler River.

Meyer, M.A., Butterfield, K.B, Murray, W.S, Smith, R. E., & Booker, J.M. (2000). Guidelines for Eliciting Expert Judgment as Probabilities or Fuzzy Logic (Los Alamos National Laboratory report LA-UR-00-218) to appear in Ross, Booker and Parkinson (Eds), Fuzzy Logic and Probability Applications, American Statistical Society, SIAM Series. Forthcoming.

Molak, V. (1997). *Fundamentals of Risk Analysis and Risk Management*. Lewis Publishers, Boca Raton.

Monroe, R.W. (1997). *A Synthesized Methodology for Eliciting Expert Judgment for Addressing Uncertainty in Decision Analysis*. Ph.D. Dissertation. Old Dominion University, Norfolk, VA. August 1997.

Monroe, R.W., Lepsch, R.A., & Unal, R. (2002). Using Expert Judgment Methodology to Address Uncertainty in Launch Vehicle Weight Estimates. *AIAA Paper 2002-5183 presented at 9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Atlanta, GA, 2002.

Morgan, M.G. & Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. London: Cambridge University Press.

Murphy, A.H. & Winkler, R.L. (1974). Credible Interval Temperature Forecasting: Some Experimental Results. *Monthly Weather Review*, 102, 784-794

Moral, Serafin & Sagrado. (1998). Aggregation of Imprecise Probabilities. IN: Bouchon-Meunier, B. (Ed.), *Aggregation and Fusion of Imperfect Information*. Physica-Verlag, New York, 162-188.

Morris, W.D. (2003). Personal conversation, August 18, 2003, 2:30PM.

Morris, W.D. (2003a). Personal conversation, October 28, 2003, 11:00AM.

Morris, P. (1977). Combining Expert Judgments: A Bayesian Approach. *Management Science*, Vol. 23, No. 7, pp. 679-693.

Morris, P. (1986). Observations on Expert Aggregation. *Management Science*, Vol. 32, No. 3, pp. 321-328.

Oberkampf, W.L., & Helton, J.C. (2001). Mathematical Representation of Uncertainty. *AIAA Paper 200-1645, Non-Deterministic Approaches Forum*, April, Seattle, WA, 2001.

Pate-Cornell, M.E. (1996). Uncertainties in risk analysis: Six levels of treatment. *Reliability Engineering and System Safety*, 54, 95-111.

Pederson, K. et al. (2000). The 'Validation Square'-Validating Design Methods. *ASME Design Theory and Methodology Conference*, September, Baltimore, MD, 2000.

Pennock, D.M. (1997). *Market-Based Belief Aggregation and Group Decision-making*. Dissertation Proposal. University of Michigan, Ann Arbor, MI.

Peterson, R.A. (2000). *Constructing Effective Questionnaires*, Sage Publications, Inc. Thousand Oaks, CA

Phillips, L.D & Phillips, M.C. (1990). Facilitated work groups: Theory and practice. Unpublished manuscript, London School of Economics and Political Science.

Preece, A. (2001). Evaluating Verification and Validation Methods in Knowledge Engineering. Retrieved October 25, 2003 from http://citeseer.nj.nec.com/515529.

Preece, A. (1994). Validation of Knowledge-Based Systems: The State-of-the-Art in North America. Retrieved October 25, 2003 from http://citeseer.nj.nec.com/26076.

Proffitt, T. (2003). Keynote Speaker. *The International Society of Parametric Analysts and The Society of Cost Estimating and Analysis 4th Joint International Conference and Educational Workshop*. June, Orlando, Florida, 2003.

Rantilla, A.H & Budescu, D.V. (1999). Aggregation of Expert Opinion. *IEEE Proceedings of the 32nd Hawaii International Conference on System Sciences.*

@Risk User's Guide (2002). Palisade Corporation, Newfield, NY.

Rowe, G. (1992) Perspectives on Expertise in the Aggregation of Judgments. *IN Expertise and Decision Support.* Wright,, G. and Bolger, Fl editors. Plenum Publishing: New York.

Sargent, R.G. (1999). Validation and Verification of Simulation Models. *Proceedings of the 1999 Winter Simulation Conference*. December, Phoenix, AZ, 1999.

Schaefer, R.E. (1976). The Evaluation of Individual and Aggregated Subjective Probability Distributions. *Organizational Behavior and Human Performance*, 17, 199-210.

Scholz, R.W. (1983). *Decision-making Under Uncertainty*. Elsevier: North-Holland, Amsterdam.

Schuenemeyer, J.H. (2002). A Framework for Expert Judgment to Assess Oil and Gas Resources. *Natural Resources Research*, Vol. 11, No. 2, pp. 97-107.

Shanteau, J., Weiss, D., Thomas, R. & Pounds, J. (2002). Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research*, 136, 253-263.

Sousa-Poza, A. (2003). *ENMA 828: Sociotechnical Systems Design*. Department of Engineering Management and Systems Engineering, Old Dominion University, Spring 2003.

Staël von Holstein, C.A.S., (1970). *Assessment and Evaluation of Subjective Probability Distributions.* Economic Research Institute at the Stockholm School of Economics, Stockholm.

Trochim, W.M. (2000). Research Methods Knowledge Base, Cornell University. Retrieved June 30, 2003 from http://trochim.human.cornell.edu/kb.

Tversky, A. & Kahneman, D. (1971). The Belief in the Law of Small Numbers. *Psychological Bulletin*, 75, 105-110.

Unal, R. & Yeniay, O. (2003). Reducing Design Risk Using Robust Design Methods: A Dual Response Surface. *NASA Grant NAG-1-01086 (Old Dominion University Project No: 113091)*. March 2003.

Unal, R. (2002). Development Of Response Surface Models And Technology Support Levels. Final Report, NASA PO # H-30938D (Old Dominion University Research Foundation Project No: 11720). February 2002.

Unal, R. & Conway, B.A. (2000). A Survey to Determine Influence of Design Parameters on Operations & Support Complexity and Cost for Launch Vehicles. Final Report, NASA PO# L-12288 (Old Dominion University Research Foundation Project No: 104881). December 2000.

Vose, D. (2000). *Quantitative Risk Analysis*. John Wiley and Sons: Chichester

Wallsten, T.S., Budescu, D.V. (1983). Encoding Subjective Probabilities: A Psychological and Psychometric Review. *Management Science*, Vol. 29, No. 2, pp.151-173.

Weerahandi, S, & Zidek, J.V. (1981). Multi-Bayesian Statistical Decision Theory. *Journal of the Royal Statistical Society*, Series A, Vol. 144, Issue 1, pp. 85-93.

West, M. (1984). Bayesian Aggregation. *Journal of the Royal Statistical Society- Series A*, Vol. 147, Issue 4, pp. 600-607.

Wilson, M.D. (1989). Task Models for Knowledge Elicitation. *IN Diaper (Eds) Knowledge Elicitation Principles, Techniques and Applications*, pp. 197-220. Chichester, Ellis Horwood.

Winkler, R.L. (1981). Combining Probability Distributions from Dependent Information Sources. *Management Science*, Vol. 27, No. 4, pp. 479-488.

Winkler, R.L., & Clemen, R.T. (1992). Sensitivity of Weights in Combining Forecasts. *Operations Research*, Vol. 40, No. 3, pp. 609-614.

Winkler, R.L. & Poses, R.M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive car unit. *Management Science*, 39, p.1526-1543.

Yager, R.R. (1996). On Mean Type Aggregation. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 26, No. 2, pp. 209-221.

Zadeh, L.A. (1992). Knowledge Representation in Fuzzy Logic. In R.R. Yager and L.S. Zadeh (Eds.), *An Introduction to Fuzzy Logic Applications in Intelligent Systems*. Boston: Kluwer Academic Publishers.

Appendix A: Expert Judgment Elicitation Sample Questionnaire (Operations Support)

Sample Questionnaire – Operations Support

From the RMAT INPUT parameters you have
Variable Scheduled

Nominal 114750

Rate the degree of uncertainty that you associate with this
Low Low/moderate Moderate Moderate/high High
Uncertainty Rating

If you feel this INPUT parameter's default value should be modified, you may provide a new estimate for the INPUT parameter's nominal
New Nominal Value

If you feel the range of possible values around the nominal value is not symmetrical, please provide own estimates of minimum and maximum
Min Max

Now that you have rated the uncertainty for this INPUT parameter, please provide a reason or for your rating. Include a rationale for any change you made to the parameter's nominal

To further document your thinking, please provide any cues (or triggers) that influence your about this

After completing the preceding steps for all parameters you have rated as uncertain, please provide a quantitative explanation of your understanding of Low, Moderate and High uncertainty, using the 5-point scales provided.

The amount of uncertainty or variation that I associate with Low Uncertainty is:
LOW Uncertainty

Less	5%	7.50%	10%	12.50%	15%	More

The amount of uncertainty or variation that I associate with Moderate Uncertainty is:
MODERATE Uncertainty

Less	10%	15%	20%	25%	30%	More

The amount of uncertainty or variation that I associate with High Uncertainty is:
HIGH Uncertainty

Less	20%	30%	40%	50%	60%	More

Appendix B: User Input Parameter List
Two-Stage-To-Orbit launch vehicle Staged at MACH 3

TSTO Launch Vehicle Parameter List

Weights & Sizing -Stage: Orbiter Input Variable Description	Variable Name	Value	Data Source
ballast weight fraction of empty wt	cballast	0	user input
growth allowance fraction	cgrow	0.15	user input
payload weight (lb)	payld	35000	user input
additional down-payload (lb.)	adpay	25000	user input
space radiator area (ft2)	srad	700	user input
mission duration (days), design	tday	10.5	user input
mission duration (days), reserve	tmar	2	user input
number of crew	ncrew	0	user input
maximum man-day capability	tmday	0	user input
nominal fuel cell power (kw)	pfcnom	14	user input
oms delta v for tank sizing (ft./sec.)	delvt	900	user input
oms delta v (ft./sec.) - burn 1	delv1	348	user input
oms delta v (ft./sec.) - burn 2	delv2	0	user input
oms delta v (ft./sec.) - burn 3	delv3	0	user input
oms delta v (ft./sec.) - station appr.	delv_sa	100	user input
oms delta v (ft./sec.) - deorbit	delv_do	366	user input
max dynamic pressure, psf	qmax	700	user input
cruise distance (nmi)	dcruise	0	user input
number of main engines	neng	9	user input
total number of fly-back jet engines	njeng	0	user input
initial t/w, orbiter	tow	1.3113	user input
lift-off t/w, 2-stage vehicle	towi	1.3369	user input
engine power level fraction	pwr	1.04	user input
design max engine power level fraction	pwrmax	1.04	user input
oxidizer-to-fuel ratio	rmix	6	user input
propellant bulk density, o/f=6.0	dbulk	22.54	user input
fuel density (lb./cu. ft.)	d_pf1	4.42	user input
lox density (lb./cu. ft.)	d_lox	71.14	user input
ullage volume fraction	ull	0.015	user input
ullage volume fraction, wing	wull	0.03	user input
wing loading (psf)	wos	65	user input
technology factor - wing str	fwstr	1	user input
technology factor - vertical fin str	fvstr	1	user input
technology factor - body dry str	fbstr	1	user input
technology factor - fuel tank	fpf1tnk	1	user input
technology factor - LO2 tank	flo2tnk	1	user input
technology factor - fuselage TPS	fbtps	1	user input
technology factor - wing & fin TPS	fwtps	1	user input
technology factor - body flap TPS	fbftps	1	user input
technology factor - landing gear	fgear	1	user input

Weights & Sizing - Stage: Orbiter (Continued)			
Input Variable Description	Variable Name	Value	Data Source
technology factor - main engines	fmeng	1	user input
technology factor - propellant feed sys	fpfs	1	user input
technology factor - gimbal actuation	fgim	1	user input
technology factor - main engine ht shld	fhtsld	1	user input
technology factor - he pneumatic sys	fhesys	1	user input
technology factor - RCS	frcs	1	user input
technology factor - OMS	foms	1	user input
technology factor - APU	fapu	1	user input
technology factor - fuel cell sys	ffcell	1	user input
technology factor - ECD	fecd	1	user input
technology factor - hydr conv & distr	fhcd	1	user input
technology factor - control surface act.	fcs	1	user input
technology factor - avionics	fav	1	user input
technology factor - environmental contrl	fec	1	user input
technology factor - internal insulation	finsl	1	user input
technology factor - purge, vent, & drn	fpvd	1	user input
technology factor - range safety	frng	1	user input
technology factor - payload container	fplcon	1	user input

Weights & Sizing - Stage: Booster			
Input Variable Description	Variable Name	Value	Data Source
number of common booster stages	nbst	1	user input
ballast weight fraction of empty wt	cballast	0	user input
growth allowance fraction	cgrow	0.15	user input
ascent time (min)	t_asc	2	user input
descent time (min)	t_desc	20	user input
operating time margin (min)	t_mar	5	user input
number of crew	ncrew	0	user input
maximum man-day capability	tmday	0	user input
electrical power req. (kw), ascent	p_asc	11.3	user input
electrical power req. (kw), descent	p_desc	7.7	user input
nominal electrical power (kw)	pfcnom	11.3	user input
max dynamic pressure, psf	qmax	700	user input
cruise distance (nmi)	dcruise	0	user input
number of main engines	neng	8	user input
total number of fly-back jet engines	njeng	0	user input

Weights & Sizing - Stage: Booster (Continued)			
Input Variable Description	Variable Name	Value	Data Source
lift-off t/w, 2-stage vehicle	tow	1.3372	user input
initial t/w, orbiter	toworb	1.3113	user input
engine power level fraction	pwr	1.04	user input
design max engine power level fraction	pwrmax	1.04	user input
oxidizer-to-fuel ratio	rmix	6	user input
propellant bulk density, o/f=6.0	dbulk	22.54	user input
propellant bulk density, o/f=6.0 (orb.)	dbulk2	22.54	user input
fuel density (lb./cu. ft.)	d_pf1	4.42	user input
lox density (lb./cu. ft.)	d_lox	71.14	user input
ullage volume fraction	ull	0.015	user input
ullage volume fraction, wing	wull	0.03	user input
wing loading (psf)	wos	65	user input
technology factor - wing str	fwstr	1	user input
technology factor - vertical fin str	fvstr	1	user input
technology factor - body dry str	fbstr	1	user input
technology factor - fuel tank	fpf1tnk	1	user input
technology factor - LO2 tank	flo2tnk	1	user input
technology factor - fuselage TPS	fbtps	1	user input
technology factor - wing & fin TPS	fwtps	1	user input
technology factor - body flap TPS	fbftps	1	user input
technology factor - landing gear	fgear	1	user input
technology factor - main engines	fmeng	1	user input
technology factor - propellant feed sys	fpfs	1	user input
technology factor - gimbal actuation	fgim	1	user input
technology factor - main engine ht shld	fhtsld	1	user input
technology factor - he pneumatic sys	fhesys	1	user input
technology factor - RCS	frcs	1	user input
technology factor - OMS	foms	0	user input
technology factor - APU	fapu	1	user input
technology factor - fuel cell sys	ffcell	1	user input
technology factor - ECD	fecd	1	user input
technology factor - hydr conv & distr	fhcd	1	user input
technology factor - control surface act.	fcs	1	user input
technology factor - avionics	fav	1	user input
technology factor - environmental contrl	fec	1	user input
technology factor - internal insulation	finsl	1	user input
technology factor - purge, vent, & drn	fpvd	1	user input
technology factor - range safety	frng	1	user input
technology factor - payload container	fplcon	1	user input

Appendix C: Pareto Reduced Input Parameter Lists (Weights & Sizing and Operations Support)

Weights & Sizing
Booster

Variable Name	Description	Nominal Value
nbst:	number of common booster stages	1
tow:	lift-off t/w, 2-stage vehicle	1.3372
pwr:	engine power level fraction	1.04
fmeng:	technology factor - main engines	1
fbstr:	technology factor - body dry str	1
fpf1tnk:	technology factor - fuel tank	1
flo2tnk:	technology factor - LO2 tank	1
cgrow:	growth allowance fraction	0.15
d_pf:	fuel density (lb./cu. ft.)	4.42
d_lox:	lox density (lb./cu.ft.)	71.14
wos:	wing loading (psf)	65
fwstr:	technology factor - wing str	1

Weights & Sizing
Orbiter

Variable Name	Description	Nominal Value
cgrow	growth allowance fraction	0.15
payld	payload weight (lb)	35000
tow	initial t/w, orbiter	1.3113
towi	lift-off t/w, 2-stage vehicle	1.3369
pwr	engine power level fraction	1.04
fmeng	technology factor - main engines	1
fbstr	technology factor - body dry str	1
fpf1tnk	technology factor - fuel tank	1
flo2tnk	technology factor - LO2 tank	1
fbtps	technology factor - fuselage TPS	1
delvt	oms delta v for tank sizing (ft./sec.)	900
delv1	oms delta v (ft./sec.) - burn 1	348
delv_sa	oms delta v (ft./sec.) - station appr.	100
delv_do	oms delta v (ft./sec.) - deorbit	366
d_pf1	fuel density (lb./cu. ft.)	4.42
d_lox	lox density (lb./cu. ft.)	71.14
wos	wing loading (psf)	65
fwstr	technology factor - wing str	1

Operations Support
Booster

Variable Name	Nominal Value
Scheduled Hours	114750
Shifts per Day	2
Missions per Year	8
MHMA Calibration	1*
Fraction of sequential (independent) work	0.05
Ground Processing	7689
Target minimum vehicle processing days	30
Launch Pad Time in Days	25.3
Number of Crews Assigned per shift	1
MTBM Calibration	0.833 *
Orbit Time	0
Vehicle Integration Time (days)	5.5
Technology Growth	0
Critical Failure Rate	0.0006052 *
Fraction Inherent Failures	0.1836 *

* reflects average value over all

Operations Support
Orbiter

Variable Name	Nominal Value
Scheduled Hours	159897
Shifts per Day	2
Missions per Year	8
MHMA Calibration	1 *
Fraction of sequential (independent)	0.05
Ground Processing	7771
Target minimum vehicle processing	30
Launch Pad Time in Days	25.3
Number of Crews Assigned per shift	1
MTBM Calibration	0.804 *
Orbit Time	252 *
Vehicle Integration Time (days)	5.5
Technology Growth	0
Critical Failure Rate	0.0005745 *
Fraction Inherent Failures	0.1645 *

* reflects average value over all

Appendix D: Calibrated Uncertainty Distributions

Operations & Support
TSTO Mach 3 Booster

			Expert A Calibrated Distribution			Expert B Calibrated Distribution			Expert C Calibrated Distribution		
Variable	**Var Name**	**Nom Value**	a2	c2	b2	a2	c2	b2	a2	c2	b2
VAR 01	Sched Hrs	114750	40000.00	120487.50	125000.00	90000.00	135460.36	140000.00	90000.00	100000.00	110000.00
VAR 04	MHMA Cal	1	0.25	1.05	1.25	0.25	1.18	1.50	0.25	1.00	1.50
VAR 05	FractsequentWk	0.05	0.02	0.05	0.08	0.03	0.06	0.06	0.05	0.10	0.15
VAR 06	GroundProc	7689	150.00	672.00	1250.00	150.00	755.51	1250.00	150.00	640.00	1250.00
VAR 07	TargMinVehProcDays	30	30.44	31.50	32.56	33.58	35.41	37.25	27.00	30.00	33.00
VAR 08	LPadTimeDays	25.3	15.00	26.57	45.00	15.00	29.87	45.00	15.00	25.30	45.00
VAR 09	No.CrewsAss/shift	1	1.00	1.05	3.00	1.00	1.18	3.00	1.00	1.00	3.00
VAR 10	MTBMCal	0.833	0.70	0.87	1.40	0.60	0.98	1.20	0.60	0.83	1.20
VAR 12	VehIntTimeDays	5.5	2.00	5.78	6.00	6.16	6.49	6.83	4.95	5.50	6.05
VAR 14	CritFailRate	0.0006052	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VAR 15	FractInheFailures	0.1836	0.17	0.19	0.19	0.17	0.22	0.26	0.13	0.18	0.24

Operations & Support
TSTO Mach 3 Orbiter

			Expert A Calibrated Distribution			Expert B Calibrated Distribution			Expert C Calibrated Distribution		
Variable	Var Name	Nom Value	a2	c2	b2	a2	c2	b2	a2	c2	b2
VAR 01	Sched Hrs	159897	50000	167892	170000	140000	188756	195000	140000	159897	195000
VAR 04	MHMA Cal	1	0.2500	1.0500	1.2500	0.2500	1.1805	1.5000	0.2500	1.0000	1.5000
VAR 05	FractsequentWk	0.05	0.0200	0.0525	0.0800	0.0300	0.0590	0.0600	0.0500	0.1000	0.1500
VAR 06	GroundProc	7771	160.0000	680.4000	1250.0000	150.000	764.953	1250.000	150.000	648.000	1250.000
VAR 07	TargMinVehProcDay	30	30.4393	31.5000	32.5607	33.5774	35.4145	37.2516	27.0000	30.0000	33.0000
VAR 08	LPadTimeDay	25.3	15.0000	26.5650	45.0000	15.0000	29.8662	45.0000	15.0000	25.3000	45.0000
VAR 09	No.CrewsAss/shif	1	1.0000	1.0500	3.0000	1.1192	1.1805	1.2417	0.9000	1.0000	1.1000
VAR 10	MTBMCal	0.804	0.6000	0.8442	1.1000	0.6000	0.9491	1.2000	0.6000	0.8040	1.2000
VAR 11	Orbit Time	252	255.6905	264.6000	273.5095	282.0498	297.4816	312.9134	226.8000	252.0000	277.2000
VAR 12	VehIntTimeDay	5.5	2.0000	5.7750	6.0000	6.1558	6.4927	6.8295	4.9500	5.5000	6.0500
VAR 14	CritFailRate	0.0005745	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	0.0005	0.0006	0.0007
VAR 15	FractInheFailure	0.1645	0.1500	0.1727	0.1800	0.1539	0.1942	0.2345	0.1151	0.1645	0.2139

Weights & Sizing
TSTO Mach 3 Booster

			Expert A Calibrated Distribution			Expert A Calibrated Distribution		
Variable	**Var**	**Nom**	a2	c2	b2	a2	c2	b2
VAR	tow:	1.3372	1.6027	1.6632	1.7237	1.6382	1.6854	1.7327
VAR	pwr:	1.04	0.8500	0.9500	0.9500	1.2251	1.2604	1.2958
VAR	cgrow:	0.15	0.1000	0.2661	0.3500	0.1519	0.1891	0.2262
VAR	d_pf:	4.42	5.5601	5.8811	6.2020	5.4148	5.5710	5.7273
VAR	d_lox	71.14	91.2117	94.6557	98.0998	87.1509	89.6600	92.1812
VAR	wos:	65	55.0000	70.0000	70.0000	72.7347	81.9271	91.1195

Weights & Sizing
TSTO Mach 3 Orbiter

			Expert A Calibrated Distribution			Expert B Calibrated Distribtuion		
Variable	**Var**	**Nom**	a2	c2	b2	a2	c2	b2
VAR 03	tow	1.3113	1.3208	1.3971	1.4733	1.6064	1.6528	1.6991
VAR 12	delv1	348	429.3384	463.0333	496.7283	426.3213	438.6250	450.9287
VAR 13	delv_sa	100	123.3731	133.0556	142.7380	111.8995	126.0417	140.1838
VAR 14	delv_do	366	451.5455	486.9833	522.4211	428.9624	461.3125	493.6626

Appendix E: Numerical Results of Aggregation

Weights & Sizing: Booster
Aggregated Uncertainty Assessments

Variable VAR02	Name -tow	Nom Value 1.3372			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	1.6027	1.6632	1.7237	
0.7	Expert B	1.6382	1.6854	1.7327	
Aggregated Response		**1.6242**	**1.6787**	**1.7341**	**1.6787**

Variable VAR03	Name -pwr	Nom Value 1.0400			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	0.8500	0.9500	0.9500	
0.7	Expert B	1.2251	1.2604	1.2958	
Aggregated Response		**1.1095**	**1.1573**	**1.1962**	**1.1583**

Variable VAR08	Name -cgrow	Nom Value 0.1500			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	0.1000	0.2661	0.3500	
0.7	Expert B	0.1519	0.1891	.02262	
Aggregated Response		**0.1183**	**0.2040**	**0.2649**	**0.2061**

Variable VAR09	Name -d_pf	Nom Value 4.4200			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	5.5601	5.8811	6.2020	
0.7	Expert B	5.4148	5.5710	5.7273	
Aggregated Response		**5.41786**	**5.640**	**5.9152**	**5.5537**

Variable VAR10	Name -d_lox	Nom Value 71.1400			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	91.2117	94.6557	98.0998	
0.7	Expert B	87.1509	89.66	92.1812	
Aggregated Response		**88.0210**	**91.1617**	**94.2747**	**91.1642**

Variable VAR11	Name -wos	Nom Value 65.0000			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	55.0000	70.0000	70.0000	
0.7	Expert B	72.7347	81.9271	91.1195	
Aggregated Response		**67.6310**	**76.8500**	**85.4620**	**76.9380**

Weights & Sizing: Orbiter
Aggregated Uncertainty Assessments

Variable	Name	Nom Value			
VAR03	**-tow**	**1.3113**			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	1.3208	1.3971	1.4733	
0.7	Expert B	1.6064	1.6528	1.6991	
Aggregated Response		**1.5103**	**1.5761**	**1.6416**	**1.5761**

Variable	Name	Nom Value			
VAR12	**-delv1**	**348.000**			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	429.3384	463.0333	496.7283	
0.7	Expert B	426.3213	438.6250	450.9287	
Aggregated Response		**426.3313**	**445.9480**	**470.4040**	**445.9340**

Variable	Name	Nom Value			
VAR13	**-delv_sa**	**100.000**			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	123.3731	133.0556	142.7380	
0.7	Expert B	111.8995	126.0417	140.1838	
Aggregated Response		**115.4150**	**128.1460**	**140.7320**	**128.1700**

Variable	Name	Nom Value			
VAR14	**-delv_do**	**366.000**			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.3	Expert A	451.5455	486.9833	522.4211	
0.7	Expert B	428.9324	461.31.2	493.6626	
Aggregated Response		**433.8230**	**469.0050**	**505.1810**	**468.8950**

Operations Support: Booster
Aggregated Uncertainty Assessments

Variable	Name	Nom Value			
VAR01	-Schedhrs	114750			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	40000	120488	125000	
0.4	Expert B	90000	135460	140000	
0.2	Expert C	90000	100000	110000	
Aggregated Response		69162	106782	125877	110004

Variable	Name	Nom Value			
VAR04	-MHMACal	1.0000			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	0.2500	1.0500	1.2500	
0.4	Expert B	0.2500	1.1805	1.5000	
0.2	Expert C	0.2500	1.0000	1.5000	
Aggregated Response		2.2500	0.9141	1.3376	0.94331

Variable	Name	Nom Value			
VAR05	-FractsequentWk	0.0500			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	0.0200	0.0525	0.0800	
0.4	Expert B	0.0300	0.0590	0.0600	
0.2	Expert C	0.0500	0.1000	0.1500	
Aggregated Response		0.0239	0.0602	0.0924	0.06039

Variable	Name	Nom Value			
VAR06	-GroundProc	7689			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	150	672	1250	
0.4	Expert B	150	756	1250	
0.2	Expert C	150	640	1250	
Aggregated Response		79	700	1250	700

Variable	Name	Nom Value			
VAR07	-TargMinVehProcDays	30.0000			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	30.4393	31.5000	32.5607	
0.4	Expert B	33.5774	35.4145	37.2516	
0.2	Expert C	27.0000	30.0000	33.0000	
Aggregated Response		30.6516	32.7658	34.7260	32.7735

Operations Support: Booster-Continued
Aggregated Uncertainty Assessments

Variable	Name	Nom Value			
VAR08	-LpadTimeDays	25.3000			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	15.0000	26.5650	45.0000	
0.4	Expert B	15.0000	29.8662	45.0000	
0.2	Expert C	15.0000	25.3000	45.0000	
Aggregated Response		15.0000	29.2110	47.5209	29.013

Variable	Name	Nom Value			
VAR09	-No.CrewsAss/shift	1.0000			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	1.0000	1.0500	3.0000	
0.4	Expert B	1.0000	1.1805	3.0000	
0.2	Expert C	1.0000	1.0000	3.0000	
Aggregated Response		1.0167	1.1697	3.0000	1.625

Variable	Name	Nom Value			
VAR10	MTBMCal	0.8330			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	0.7000	0.8747	1.4000	
0.4	Expert B	0.6000	0.9833	1.2000	
0.2	Expert C	0.6000	0.8330	1.2000	
Aggregated Response		0.6442	0.9433	1.3216	0.93758

Variable	Name	Nom Value			
VAR12	-VehIntTimeDays	5.5000			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	2.0000	5.7750	6.0000	
0.4	Expert B	6.1558	6.4927	6.8295	
0.2	Expert C	4.9500	5.5000	6.0500	
Aggregated Response		4.4074	5.5176	6.2581	5.8875

Variable	Name	Nom Value			
VAR14	-CritFailRate	0.0006			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	0.0006	0.0006	0.0007	
0.4	Expert B	0.0007	0.0007	0.0008	
0.2	Expert C	0.0005	0.0006	0.0007	
Aggregated Response		0.00061	0.00066	0.00075	0.00066

Operations Support: Booster-Continued
Aggregated Uncertainty Assessments

Variable	Name	Nom Value			
VAR15	-FractInheFailures	0.1836			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	0.1700	0.1900	0.1900	
0.4	Expert B	0.1718	0.2167	0.2617	
0.2	Expert C	0.1285	0.1836	0.2387	
Aggregated Response		0.1606	0.1967	0.2306	0.1969

Operations Support: Orbiter
Aggregated Uncertainty Assessments

Variable	Name	Nom Value			
VAR01	-SchedHrs	159897			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	50000	167892	170000	
0.4	Expert B	140000	188756	195000	
0.2	Expert C	140000	159897	195000	
Aggregated Response		101153	154519	181044	158807

Variable	Name	Nom Value			
VAR06	-GroundProc	7771			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	160	680	1250	
0.4	Expert B	150	765	1250	
0.2	Expert C	150	648	1250	
Aggregated Response		142.319	703.9100	1250.000	705.470

Variable	Name	Nom Value			
VAR09	-No.CrewsAss/shift	1.0000			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	1.0000	1.0500	3.0000	
0.4	Expert B	1.1192	1.1805	1.2417	
0.2	Expert C	0.9000	1.0000	1.1000	
Aggregated Response		1.0509	1.3476	1.8779	1.1139

Operations Support: Orbiter-Continued
Aggregated Uncertainty Assessments

Variable	Name	Nom Value			
VAR10	**-MTBMCal**	**0.8040**			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	0.6000	0.8442	1.1000	
0.4	Expert B	0.6000	0.9491	1.2000	
0.2	Expert C	0.6000	0.8040	1.2000	
Aggregated Response		0.6000	0.8794	0.1120	0.8862

Variable	Name	Nom Value			
VAR11	**-OrbitTime**	**252.00**			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	255.69	264.60	273.51	
0.4	Expert B	282.05	297.48	312.91	
0.2	Expert C	226.80	252.00	277.20	
Aggregated Response		258.76	275.23	291.53	275.24

Variable	Name	Nom Value			
VAR14	**-CritFailRate**	**0.00057**			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	0.0006	0.0006	0.0006	
0.4	Expert B	0.0006	0.0007	0.0007	
0.2	Expert C	0.0005	0.0006	0.0007	
Aggregated Response		0.00057	0.00063	0.00066	0.00063

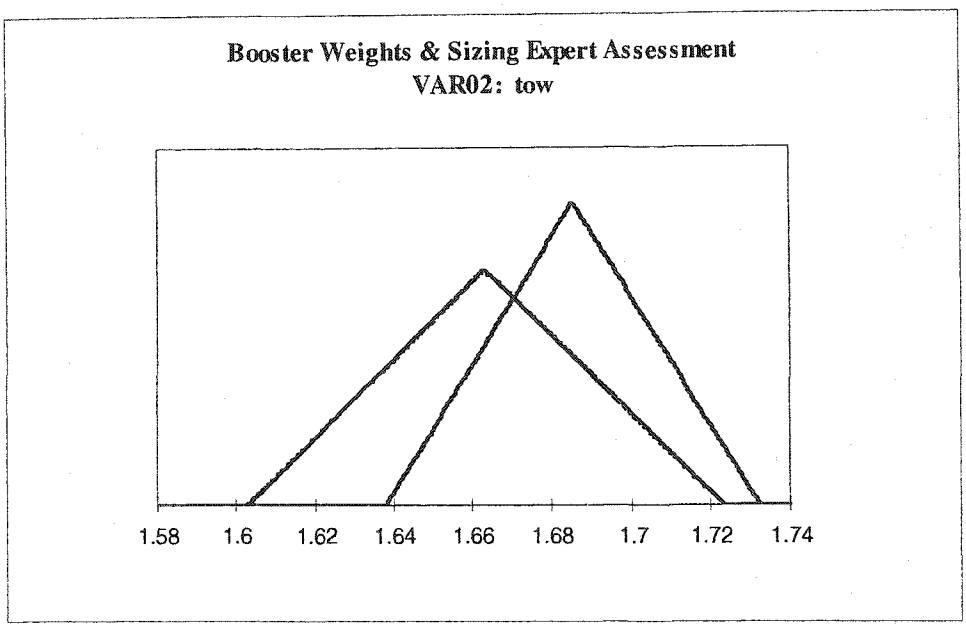
Variable	Name	Nom Value			
VAR15	**-FractInheFailures**	**0.1645**			
Weighting factor	Expert	Min	Most Likely	Max	Mode
0.4	Expert A	0.1500	0.1727	0.1800	
0.4	Expert B	0.1539	0.1942	0.2345	
0.2	Expert C	0.1151	0.1645	0.2139	
Aggregated Response		0.1416	0.1776	0.2131	0.1776

Appendix F: BESTFIT® Distributions by Variable

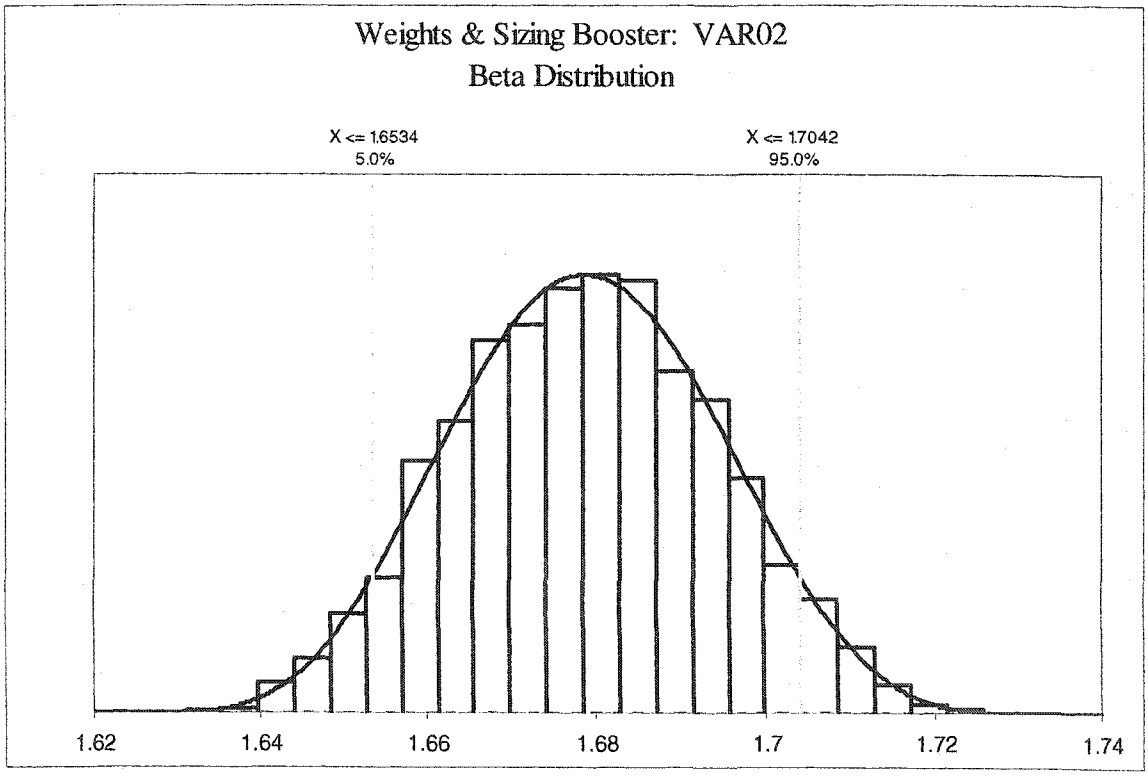
BESTFIT® Distributions by Variable

OPS Booster VAR01	Beta Distribution
OPS Booster VAR04	Beta Distribution
OPS Booster VAR05	Beta Distribution
OPS Booster VAR06	Beta Distribution
OPS Booster VAR07	Beta Distribution
OPS Booster VAR08	Beta Distribution
OPS Booster VAR09	Beta Distribution
OPS Booster VAR10	Beta Distribution
OPS Booster VAR11	Beta Distribution
OPS Booster VAR12	Triangular Distribution
OPS Booster VAR14	Beta Distribution
OPS Booster VAR15	Beta Distribution
OPS Orbiter VAR01	Beta Distribution
OPS Orbiter VAR06	Beta Distribution
OPS Orbiter VAR09	Triangular Distribution
OPS Orbiter VAR10	Weibull Distribution
OPS Orbiter VAR11	Beta Distribution
OPS Orbiter VAR14	Beta Distribution
OPS Orbiter VAR15	Beta Distribution
W&S Booster VAR02	Beta Distribution
W&S Booster VAR03	Beta Distribution
W&S Booster VAR08	Beta Distribution
W&S Booster VAR09	Beta Distribution
W&S Booster VAR10	Beta Distribution
W&S Booster VAR11	Beta Distribution
W&S Orbiter VAR03	Beta Distribution
W&S Orbiter VAR12	Beta Distribution
W&S Orbiter VAR13	Beta Distribution
W&S Orbiter VAR14	Beta Distribution

Appendix G: Aggregated Distributions with Test Statistics



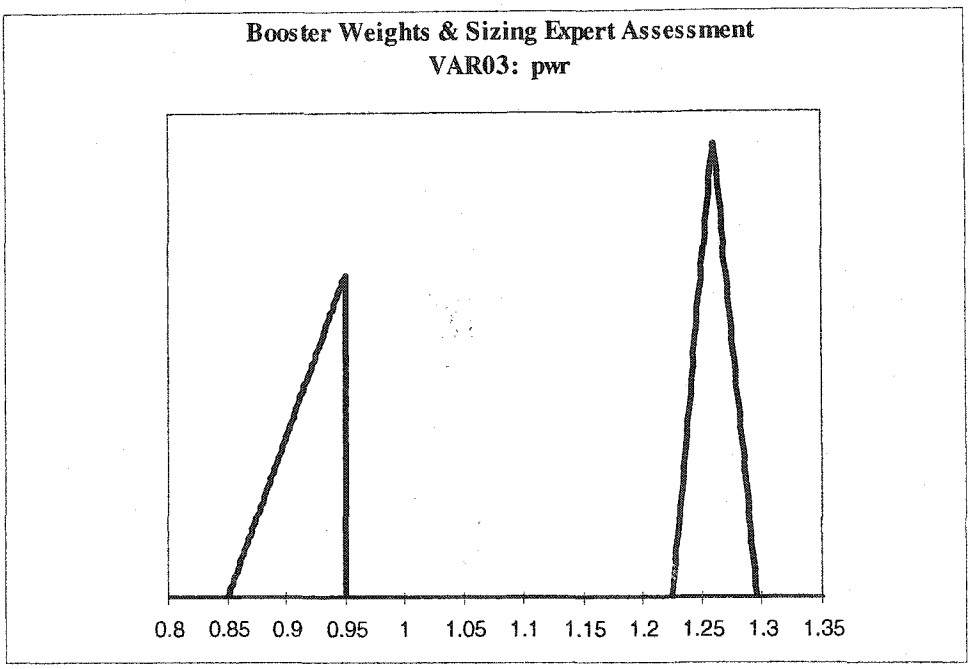
Weights & Sizing Booster: VAR02 Calibrated Expert Assessments



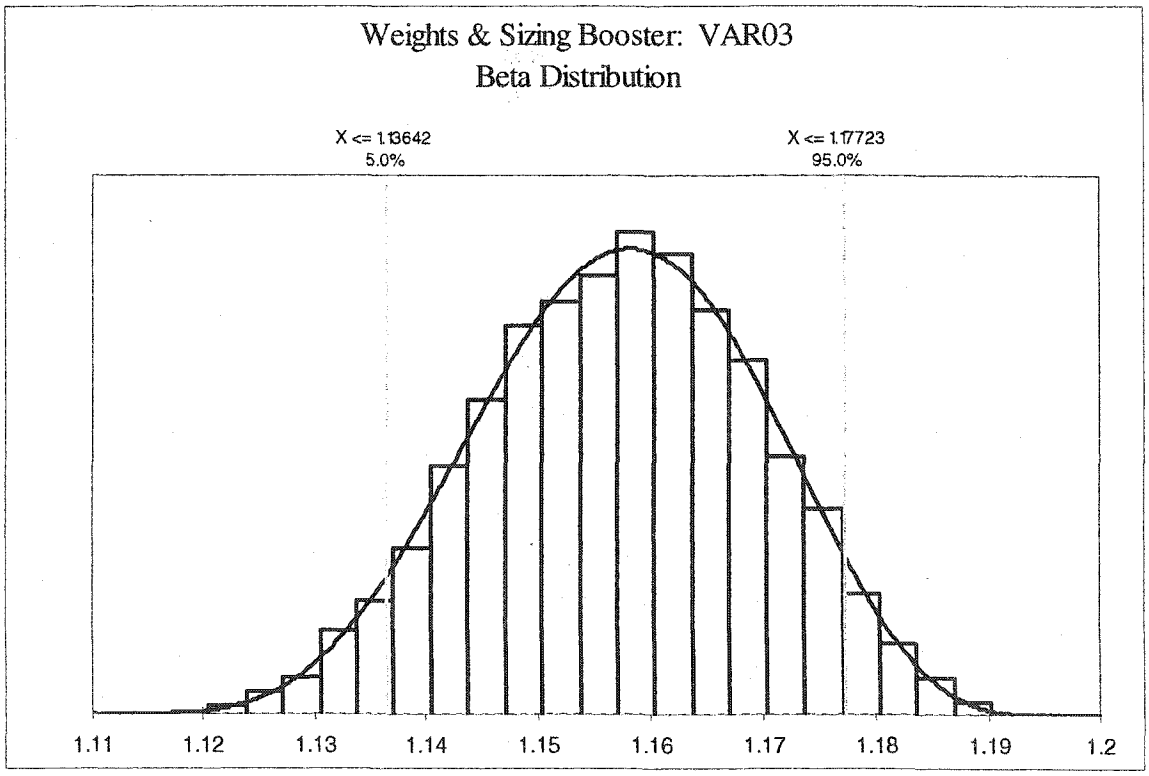
Weights & Sizing Booster: VAR02 Aggregated Response

Fit	Beta
α1	5.834870331
α2	5.919257795
Minimum	1.62422
Maximum	1.7341
Mean	1.67876
Mode	1.67868
Median	1.67874
Std. Deviation	0.015384
Variance	0.00023667
Skewness	0.0075
Kurtosis	2.5934

	Chi-Sq	A-D	K-S
Test Value	74.14	0.2928	0.005005
P Value	0.4408	N/A	N/A
Rank	1	1	1
# Bins	74	N/A	N/A



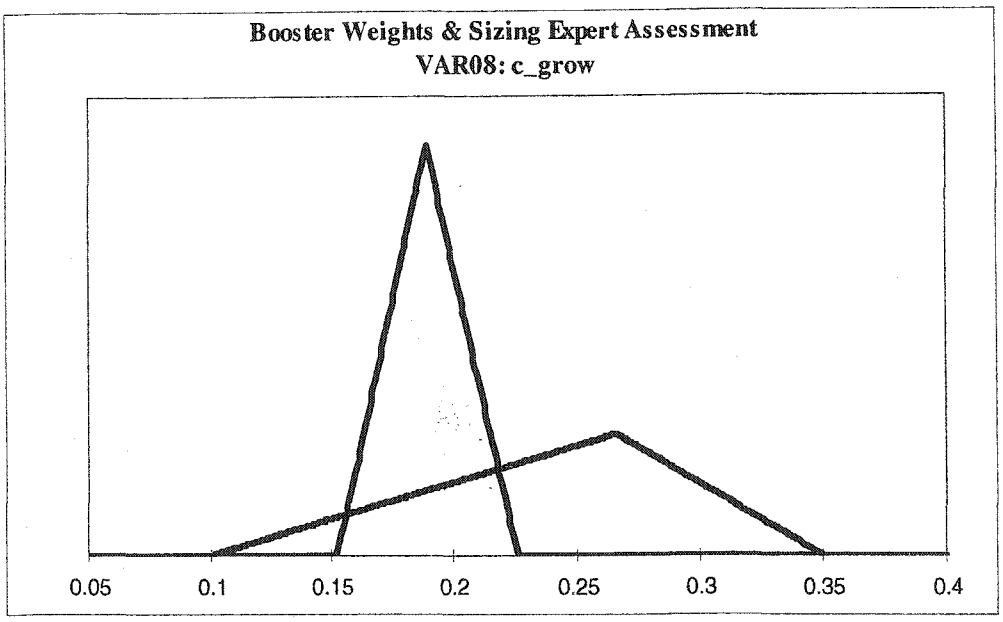
Weights & Sizing Booster: VAR03 Calibrated Expert Assessments



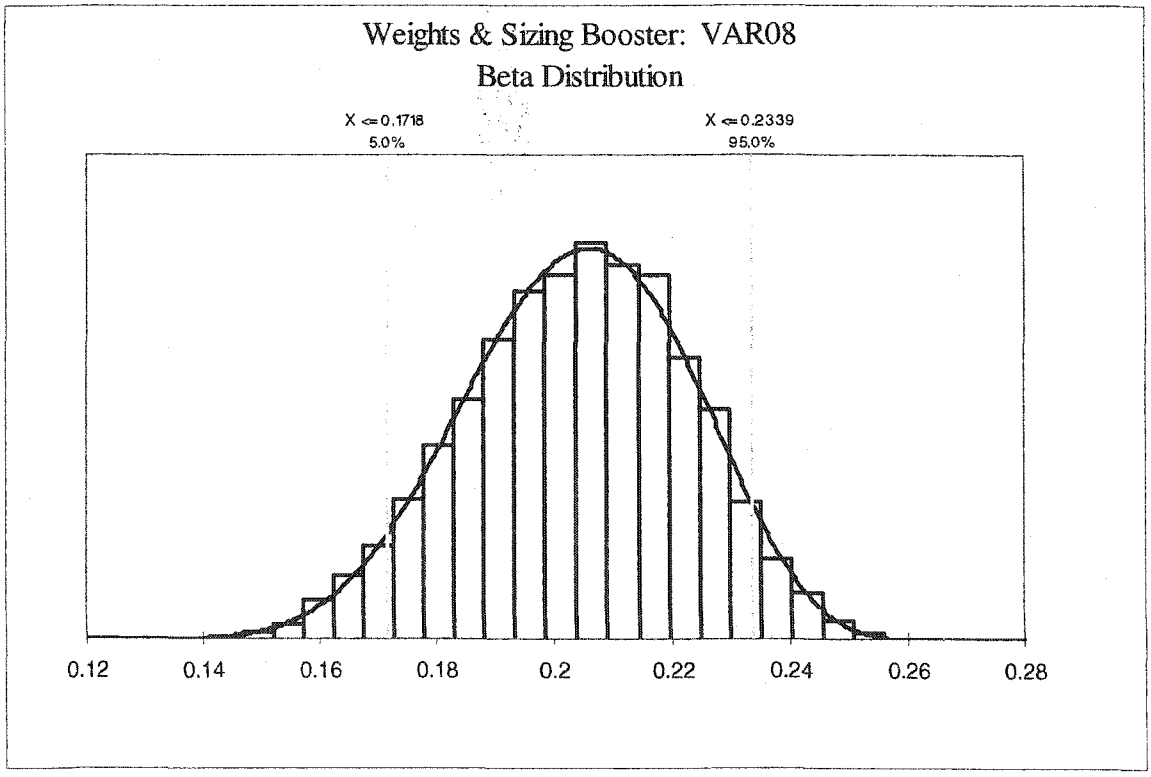
Weights & Sizing Booster: VAR03 Aggregated Response

Fit	Beta
α1	6.140046566
α2	4.989750747
Minimum	1.109497
Maximum	1.196155
Mean	1.157304
Mode	1.158285
Median	1.157582
Std. Deviation	0.012374
Variance	0.00015312
Skewness	-0.1103
Kurtosis	2.5923

	Chi-Sq	A-D	K-S
Test Value	86.49	0.4867	0.004959
P Value	0.4954	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



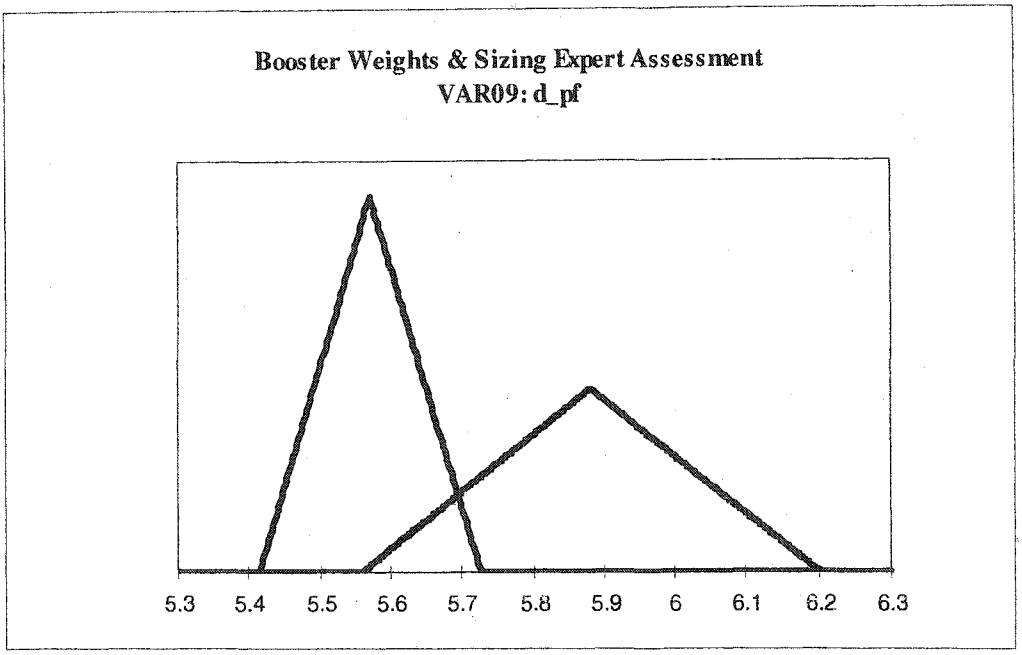
Weights & Sizing Booster: VAR08 Calibrated Expert Assessments



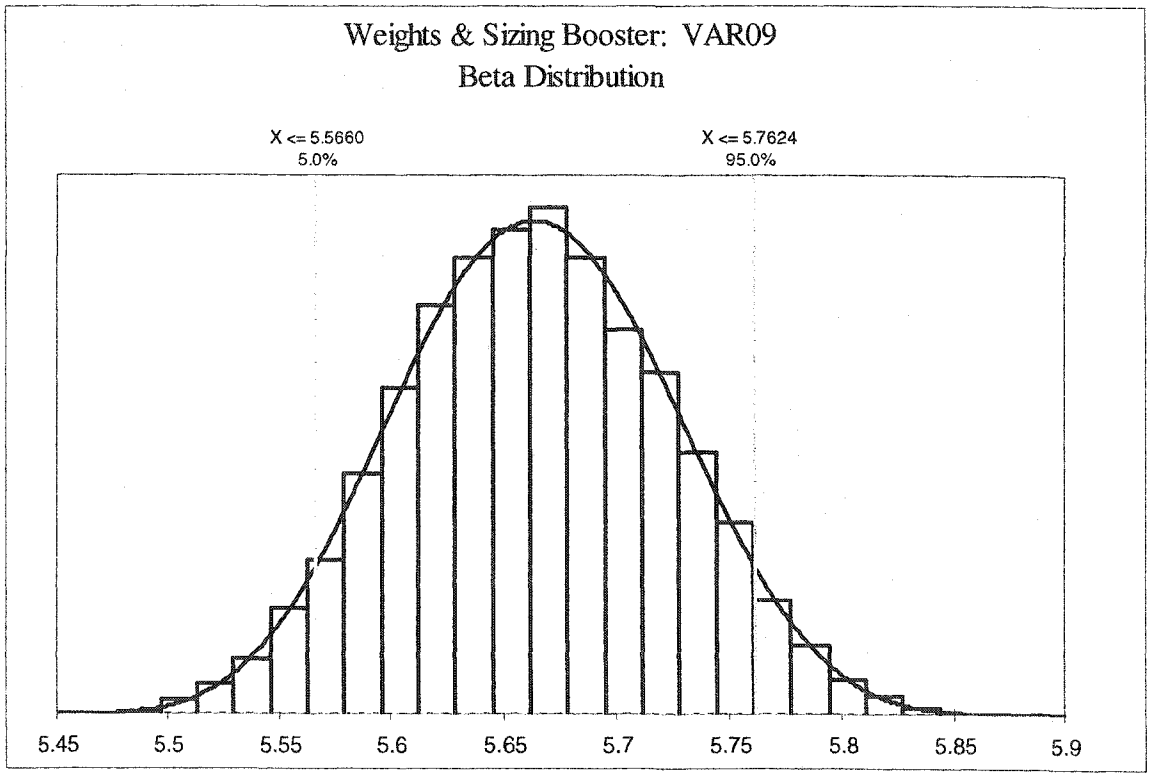
Weights & Sizing Booster: VAR08 Aggregated Response

Fit	Beta
α1	7.999166469
α2	5.69352126
Minimum	0.11832
Maximum	0.2649
Mean	0.20395
Mode	0.20606
Median	0.20457
Std. Deviation	0.018847
Variance	0.0003552
Skewness	-0.1669
Kurtosis	2.6798

	Chi-Sq	A-D	K-S
Test Value	108.7	0.993	0.007553
P Value	0.0577	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



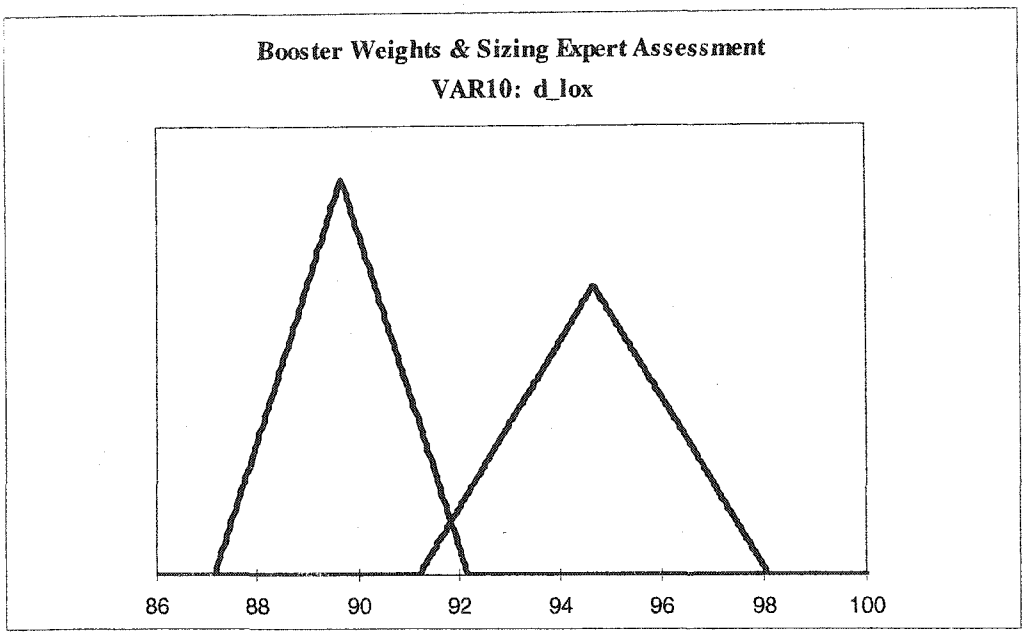
Weights & Sizing Booster: VAR09 Calibrated Expert Assessments



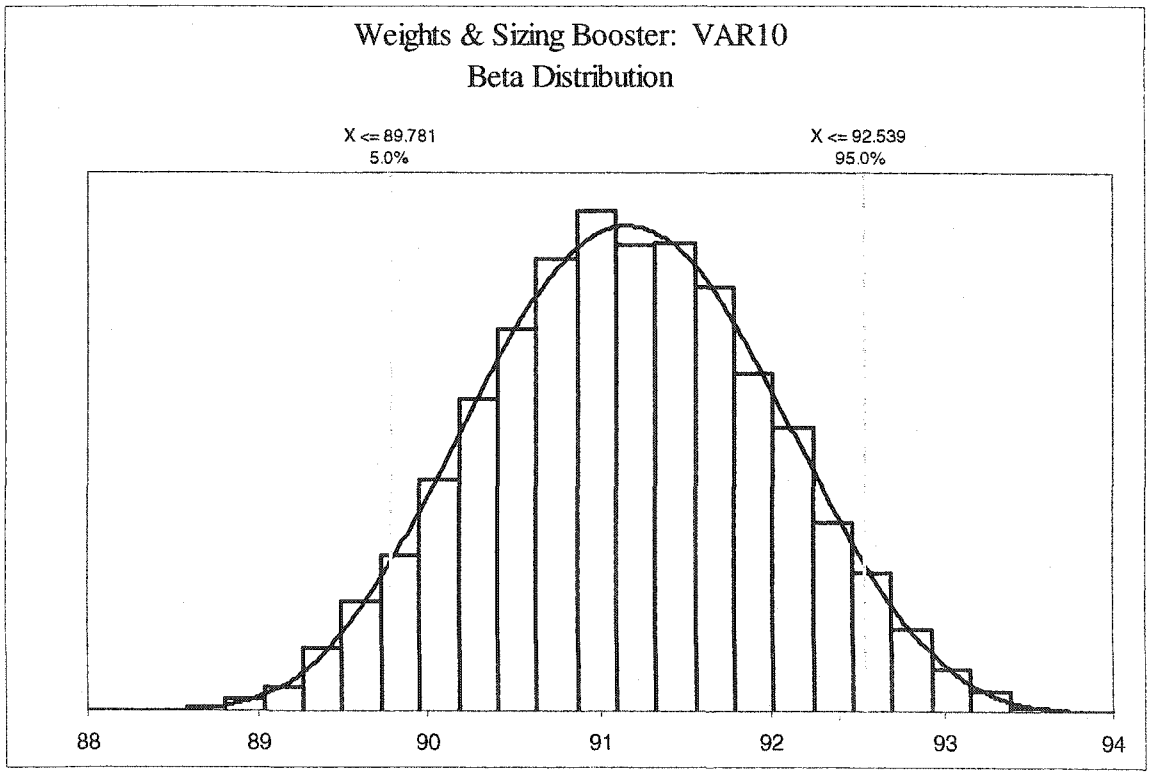
Weights & Sizing Booster: VAR09 Aggregated Response

Fit	Beta
α1	8.141951904
α2	8.308056587
Minimum	5.41786
Maximum	5.91526
Mean	5.66405
Mode	5.6637
Median	5.66394
Std. Deviation	0.059534
Variance	0.0035442
Skewness	0.0091
Kurtosis	2.6916

	Chi-Sq	A-D	K-S
Test Value	68.94	0.4938	0.005596
P Value	0.923	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



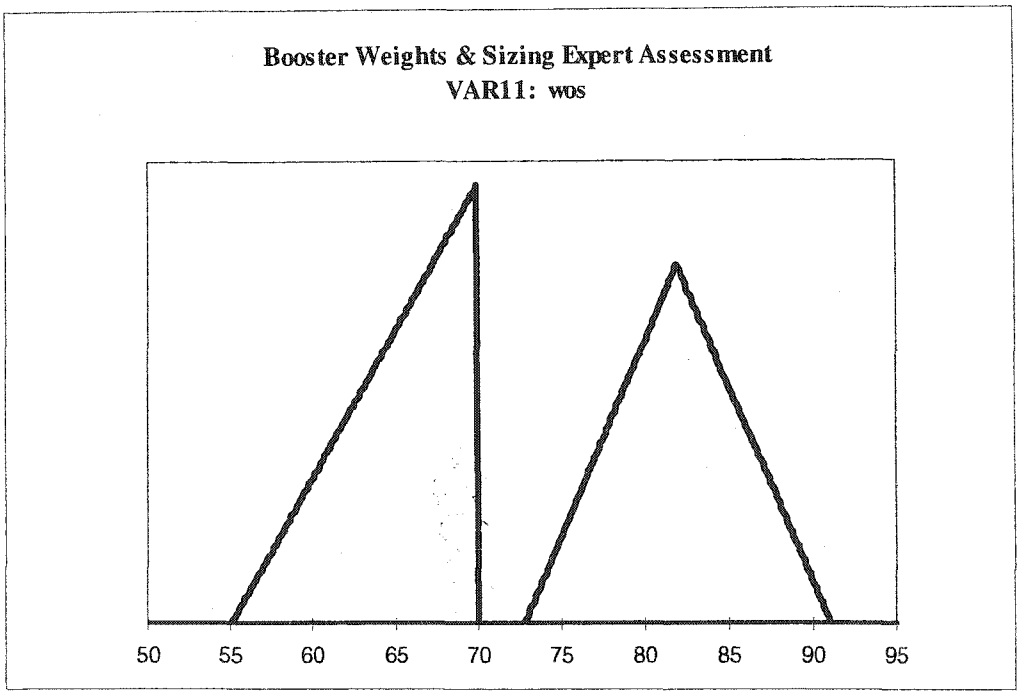
Weights & Sizing Booster: VAR10 Calibrated Expert Assessments



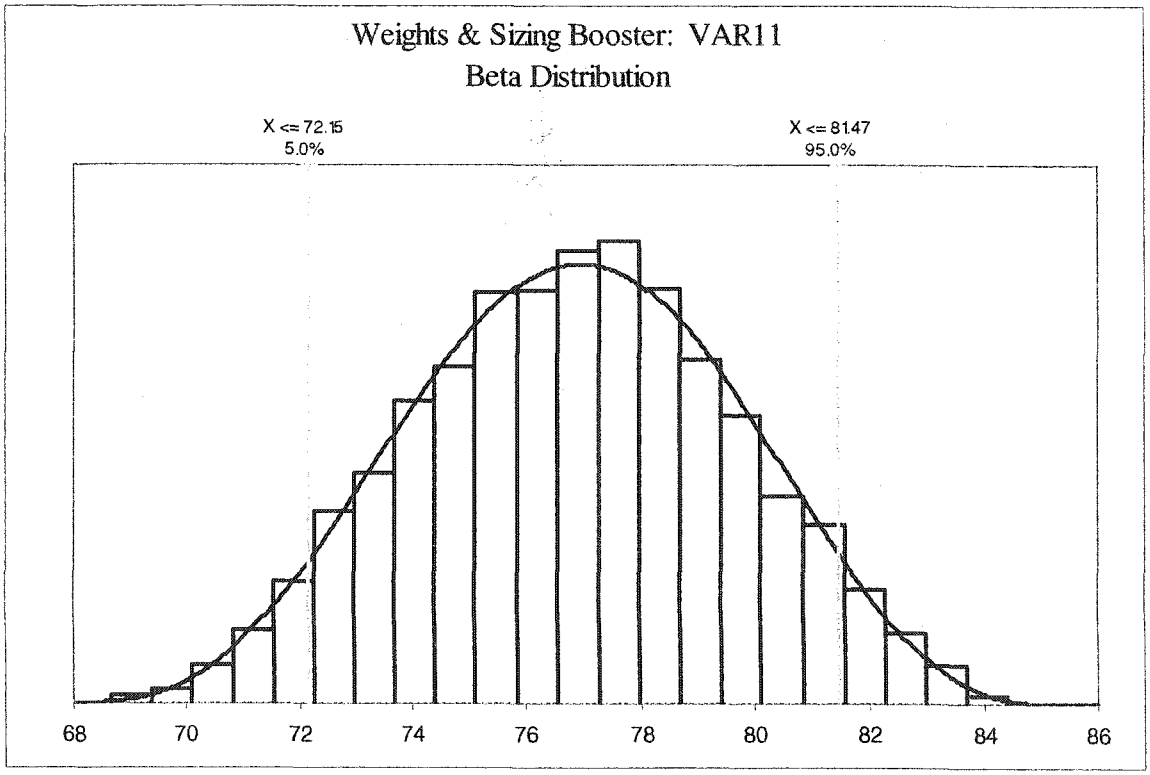
Weights & Sizing Booster: VAR10 Aggregated Response

Fit	Beta
α1	6.523616218
α2	6.466299656
Minimum	88.021
Maximum	94.2747
Mean	91.1617
Mode	91.1642
Median	91.1624
Std. Deviation	0.83598
Variance	0.69886
Skewness	-0.0044
Kurtosis	2.6248

	Chi-Sq	A-D	K-S
Test Value	88.44	0.3761	0.005245
P Value	0.4367	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



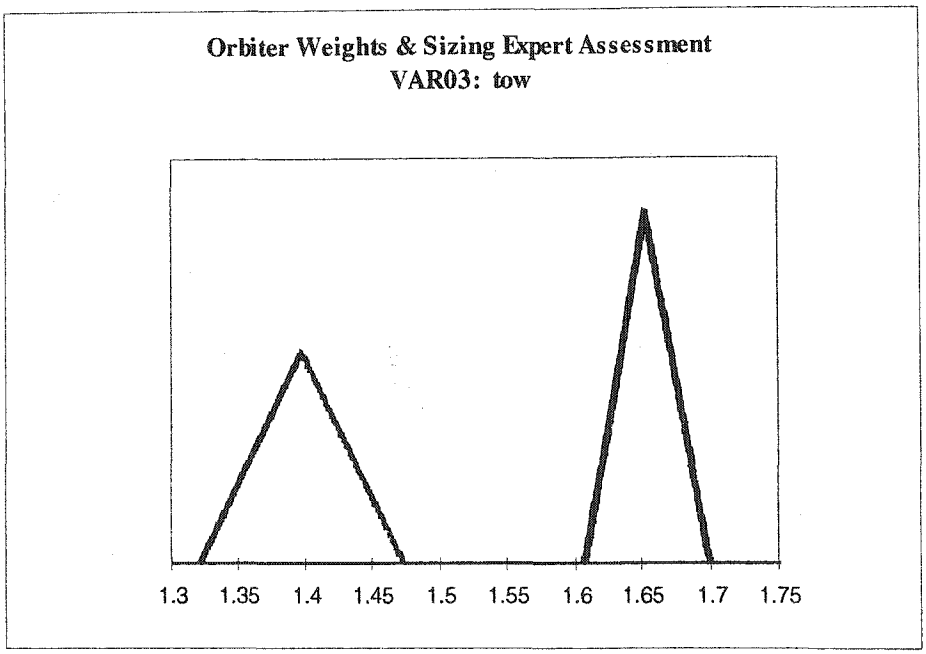
Weights & Sizing Booster: VAR11 Calibrated Expert Assessments



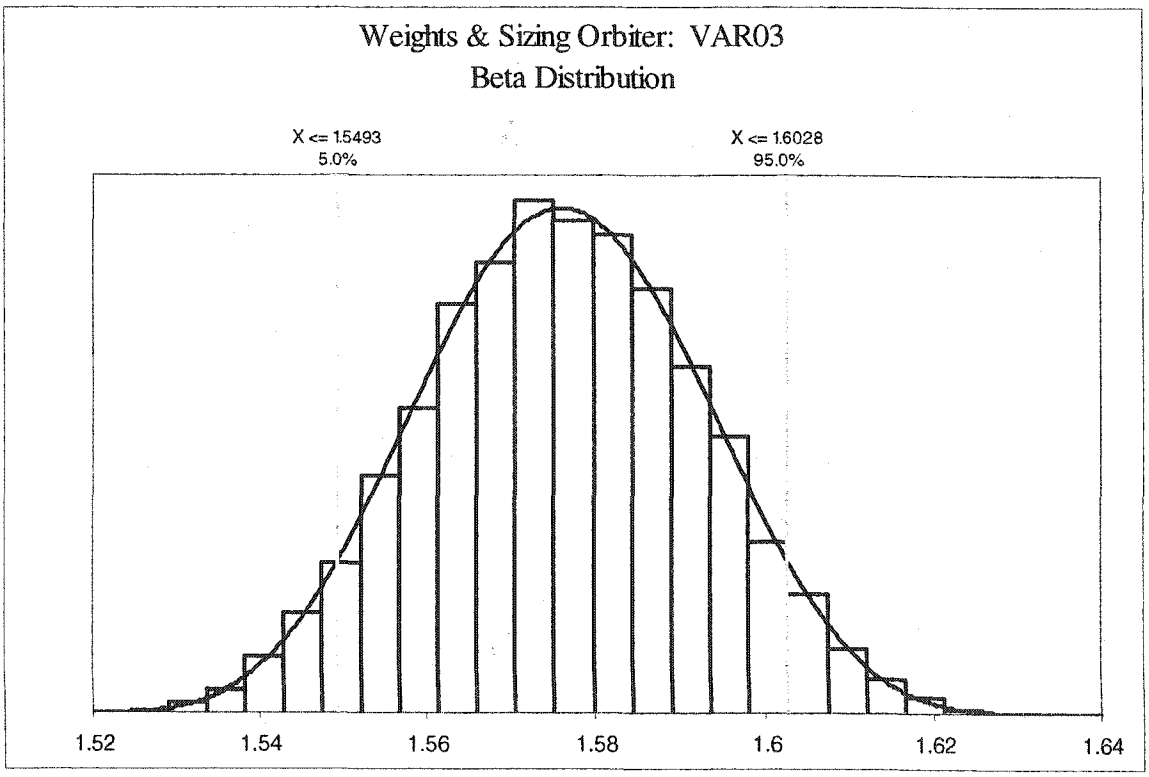
Weights & Sizing Booster: VAR11 Aggregated Response

Fit	Beta
α1	4.624613159
α2	4.319765245
Mean	76.85
Mode	76.938
Median	76.874
Std. Deviation	2.8255
Variance	7.9836
Skewness	-0.0393
Kurtosis	2.4998

	Chi-Sq	A-D	K-S
Test Value	92.89	1.191	0.008877
P Value	0.313	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



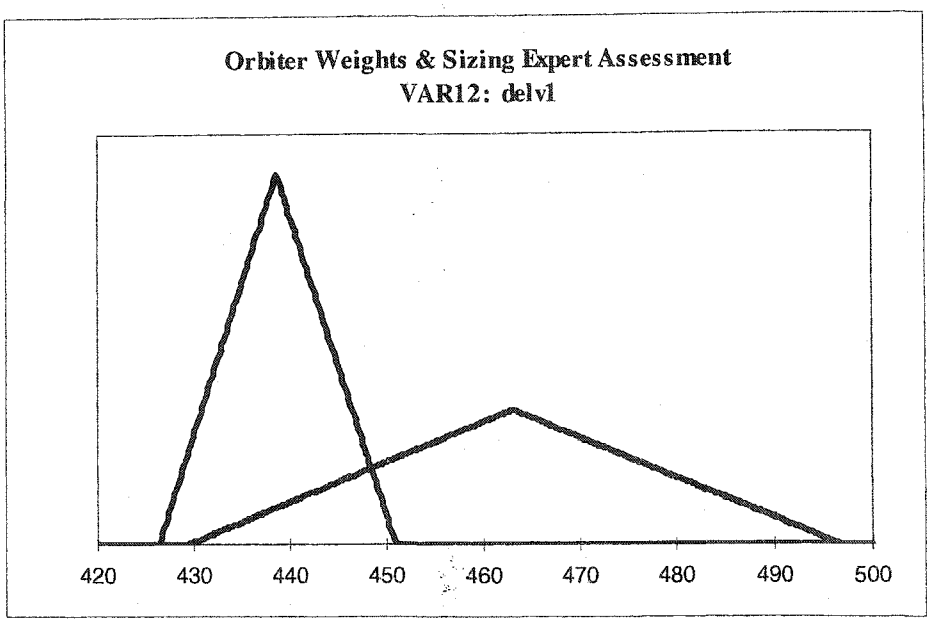
Weights & Sizing Orbiter: VAR03 Calibrated Expert Assessments



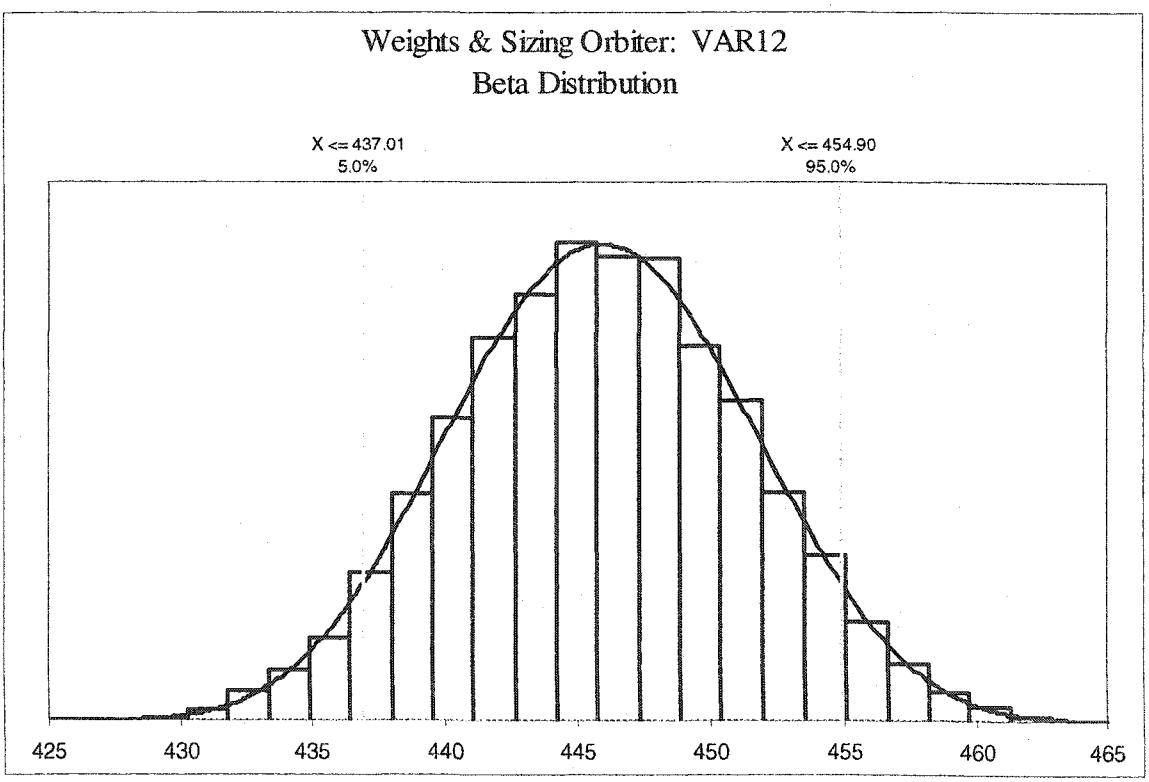
Weights & Sizing Orbiter: VAR03 Aggregated Response

Fit	Beta
α1	7.695706374
α2	7.660219686
Minimum	1.51025
Maximum	1.64156
Mean	1.57606
Mode	1.57608
Median	1.57606
Std. Deviation	0.016234
Variance	0.00026356
Skewness	-0.0022
Kurtosis	2.6731

	Chi-Sq	A-D	K-S
Test Value	76.65	0.2516	0.003759
P Value	0.7784	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



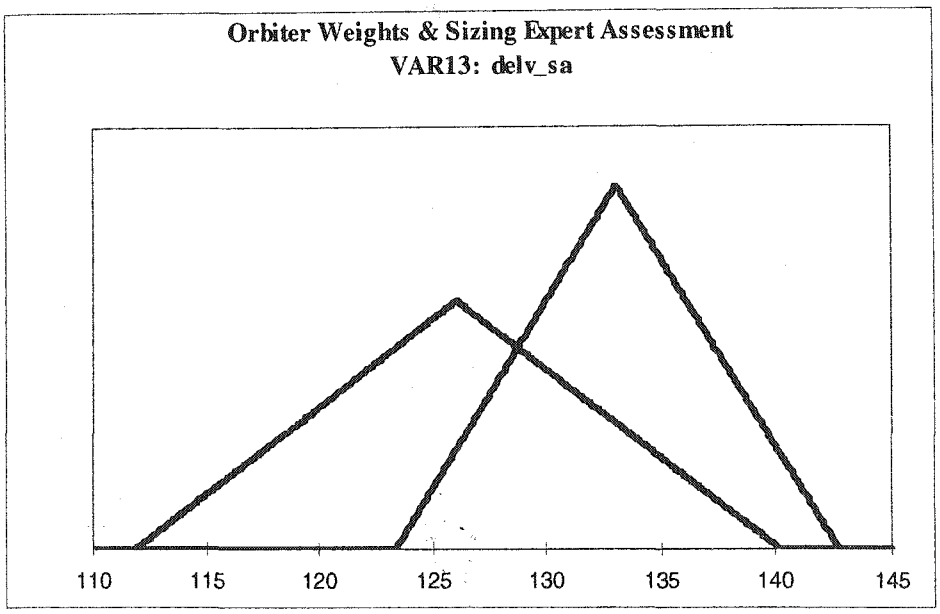
Weights & Sizing Orbiter: VAR12 Calibrated Expert Assessments



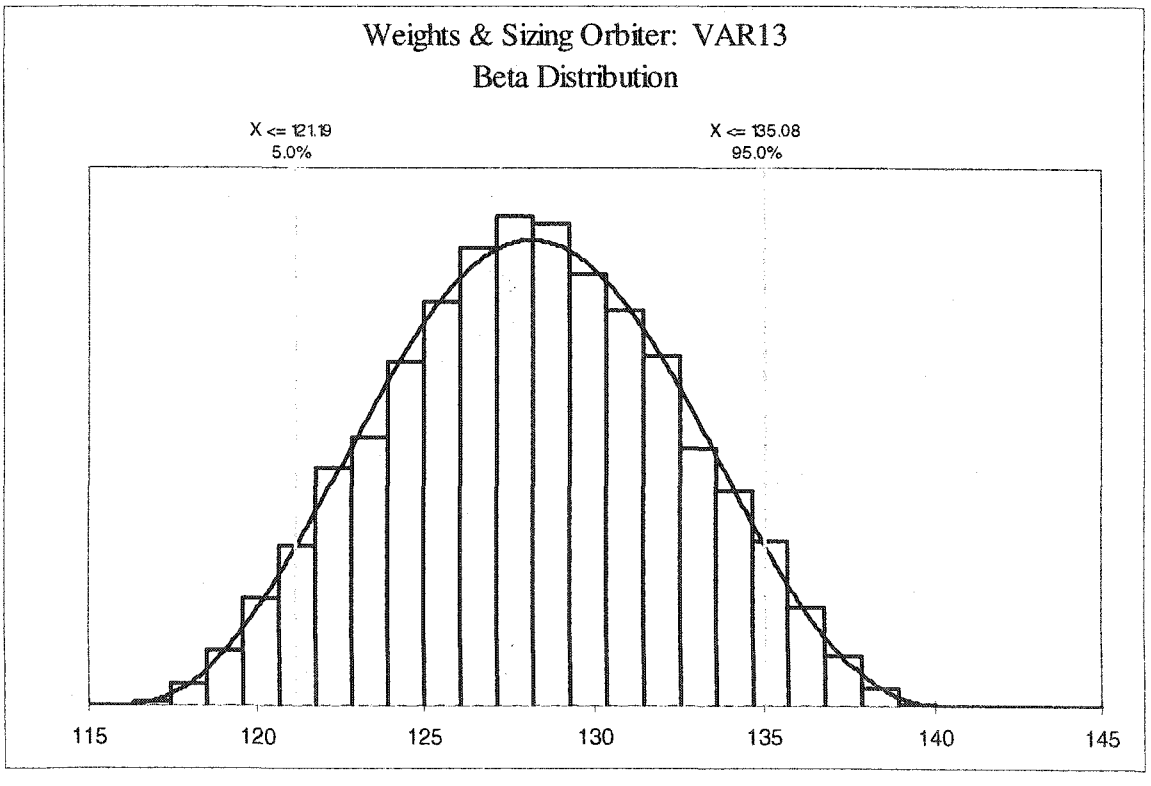
Weights & Sizing Orbiter: VAR12 Aggregated Response

Fit	Beta
α1	9.523694865
α2	9.619120999
Minimum	421.734
Maximum	470.404
Mean	445.948
Mode	445.934
Median	445.943
Std. Deviation	5.4221
Variance	29.399
Skewness	0.0042
Kurtosis	2.7291

	Chi-Sq	A-D	K-S
Test Value	64.78	0.2211	0.004905
P Value	0.9643	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



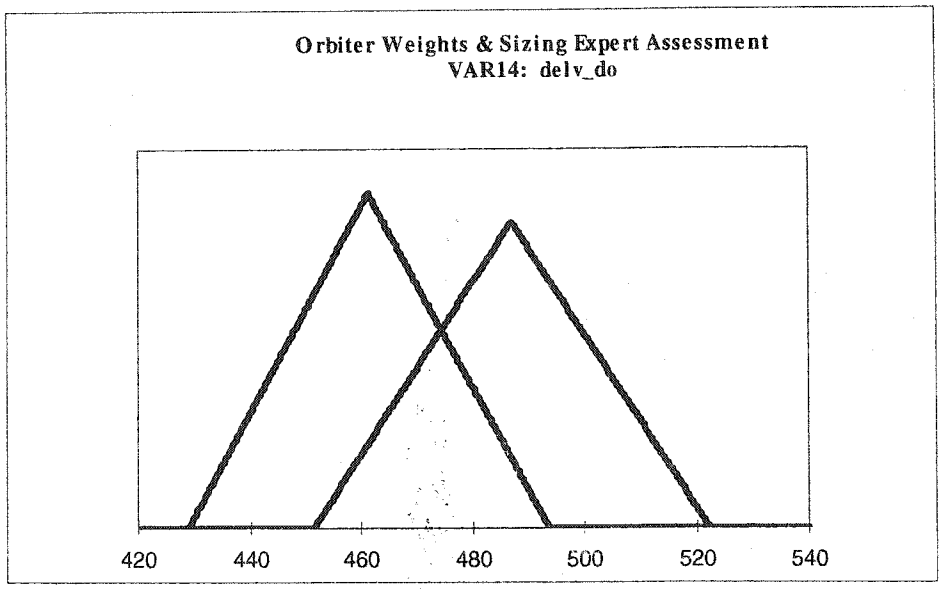
Weights & Sizing Orbiter: VAR13 Calibrated Expert Assessments



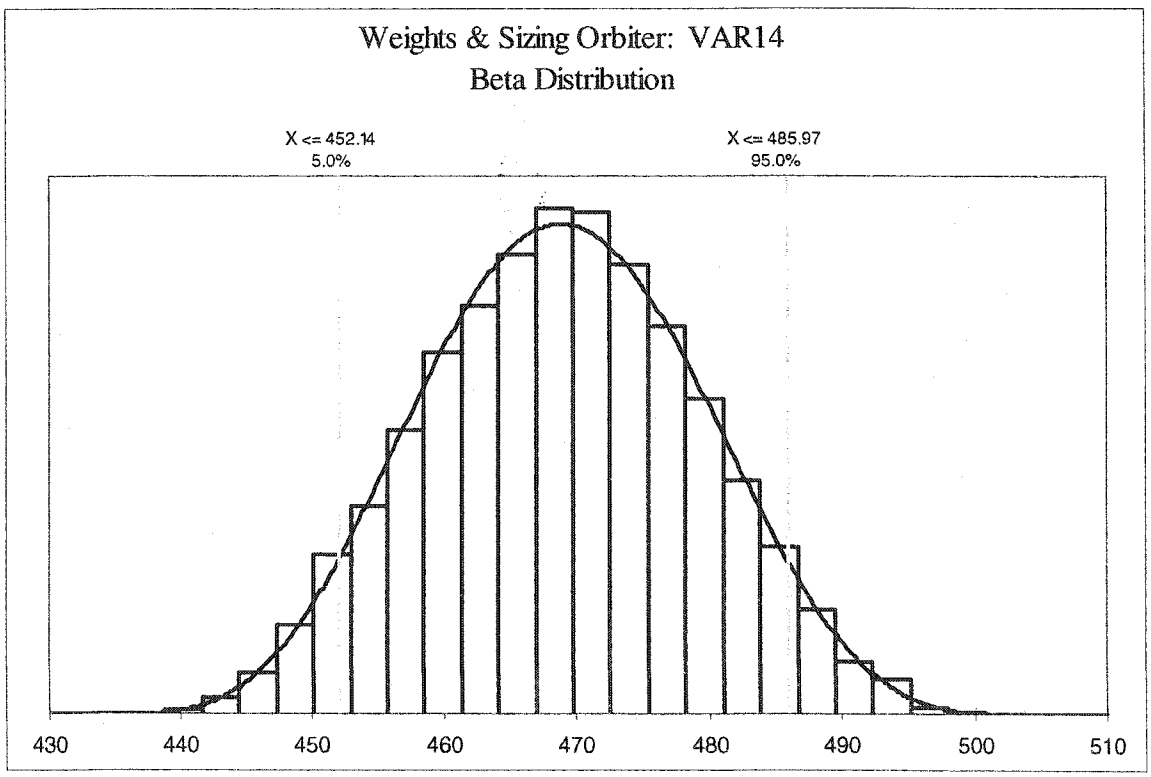
Weights & Sizing Orbiter: VAR13 Aggregated Response

Fit	Beta
α1	4.038315296
α2	3.992221032
Minimum	115.415
Maximum	140.732
Mean	128.146
Mode	128.17
Median	128.153
Std. Deviation	4.2124
Variance	17.744
Skewness	-0.0069
Kurtosis	2.4561

	Chi-Sq	A-D	K-S
Test Value	94.86	1.193	0.008244
P Value	0.2646	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



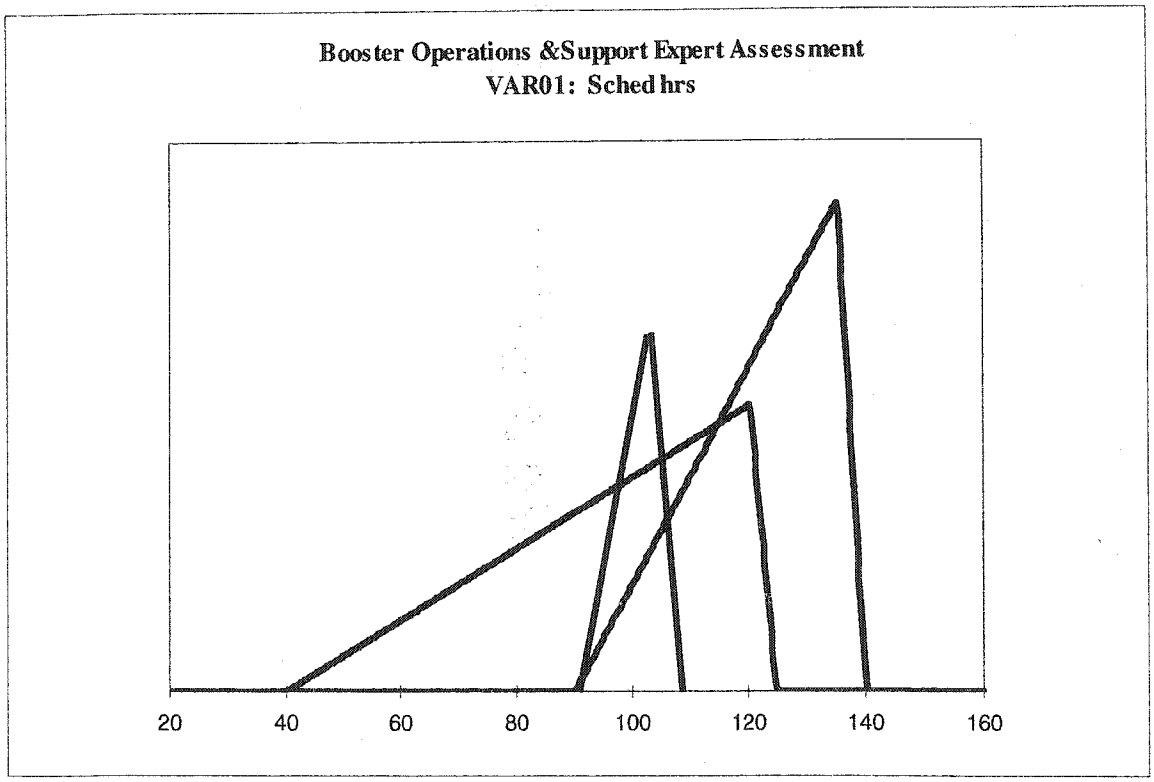
Weights & Sizing Orbiter: VAR14 Calibrated Expert Assessments



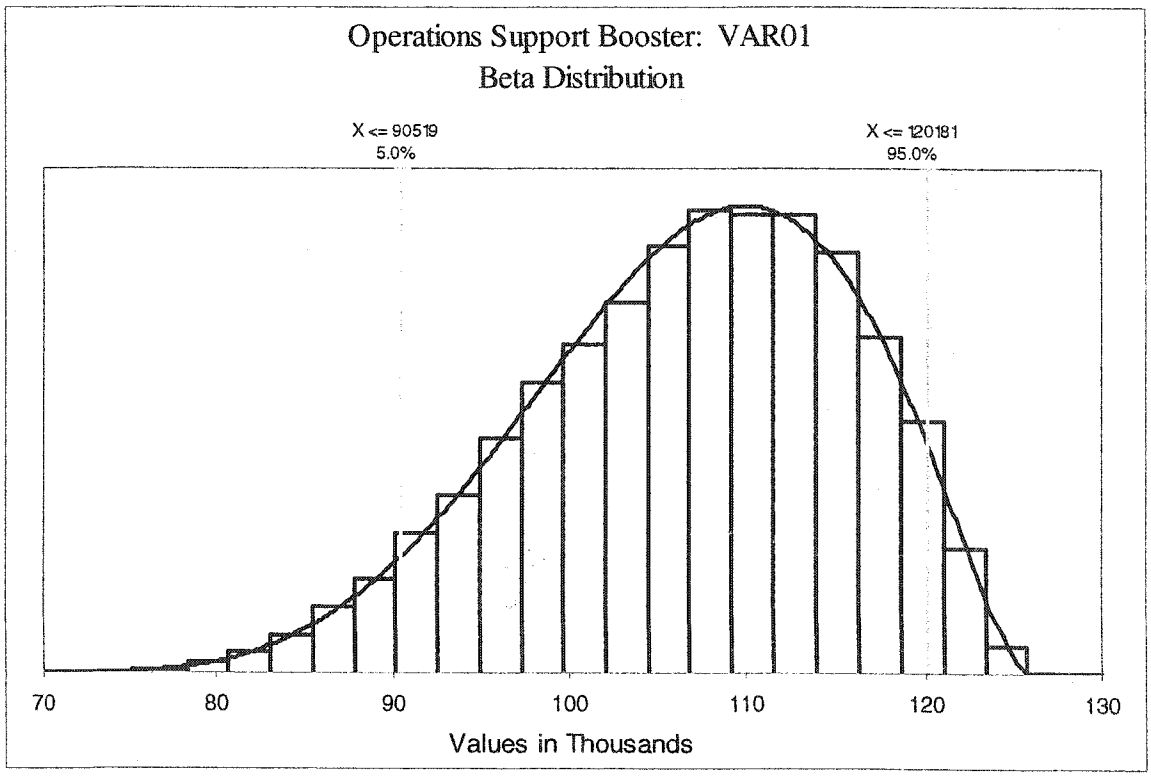
Weights & Sizing Orbiter: VAR14 Aggregated Response

Fit	Beta General
α1	5.474717161
α2	5.629576718
Minimum	433.823
Maximum	505.181
Mean	469.005
Mode	468.895
Median	468.974
Std. Deviation	10.254
Variance	105.149
Skewness	0.0148
Kurtosis	2.5749

	Chi-Sq	A-D	K-S
Test Value	92.89	0.9795	0.008724
P Value	0.313	N/A	N/A
Rank	1	1	2
# Bins	88	N/A	N/A



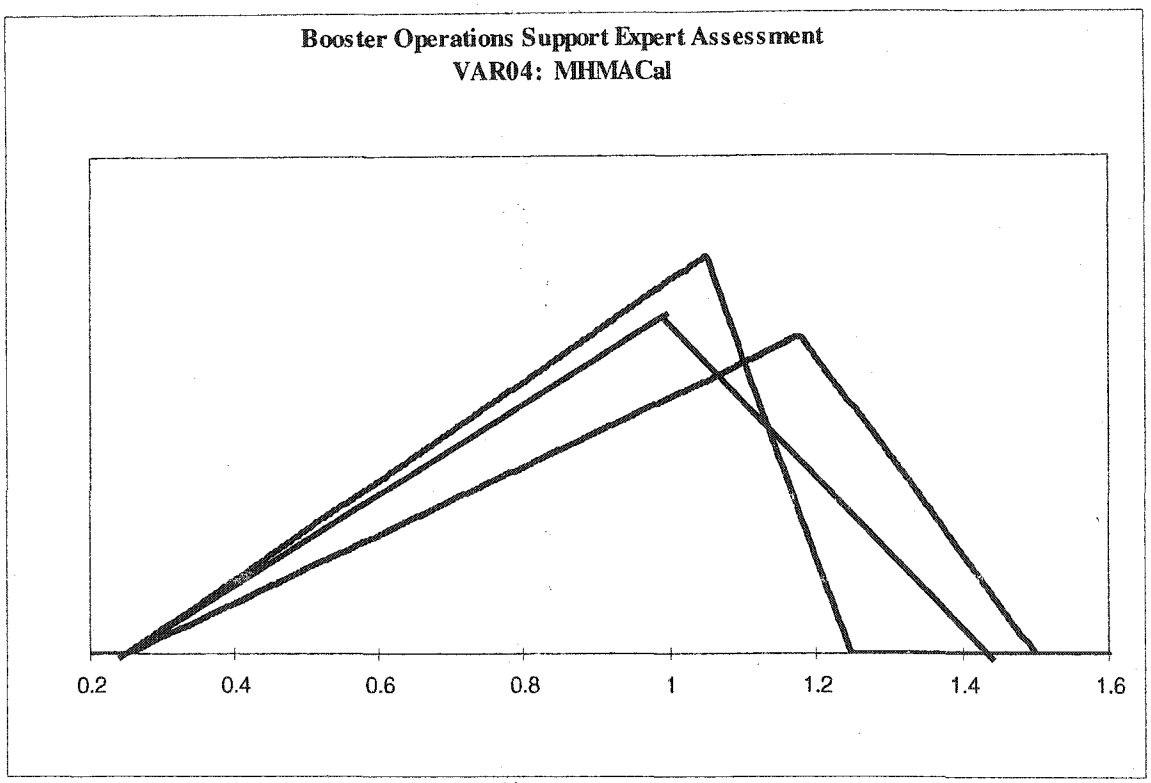
Operations Support Booster: VAR01 Calibrated Expert Assessments



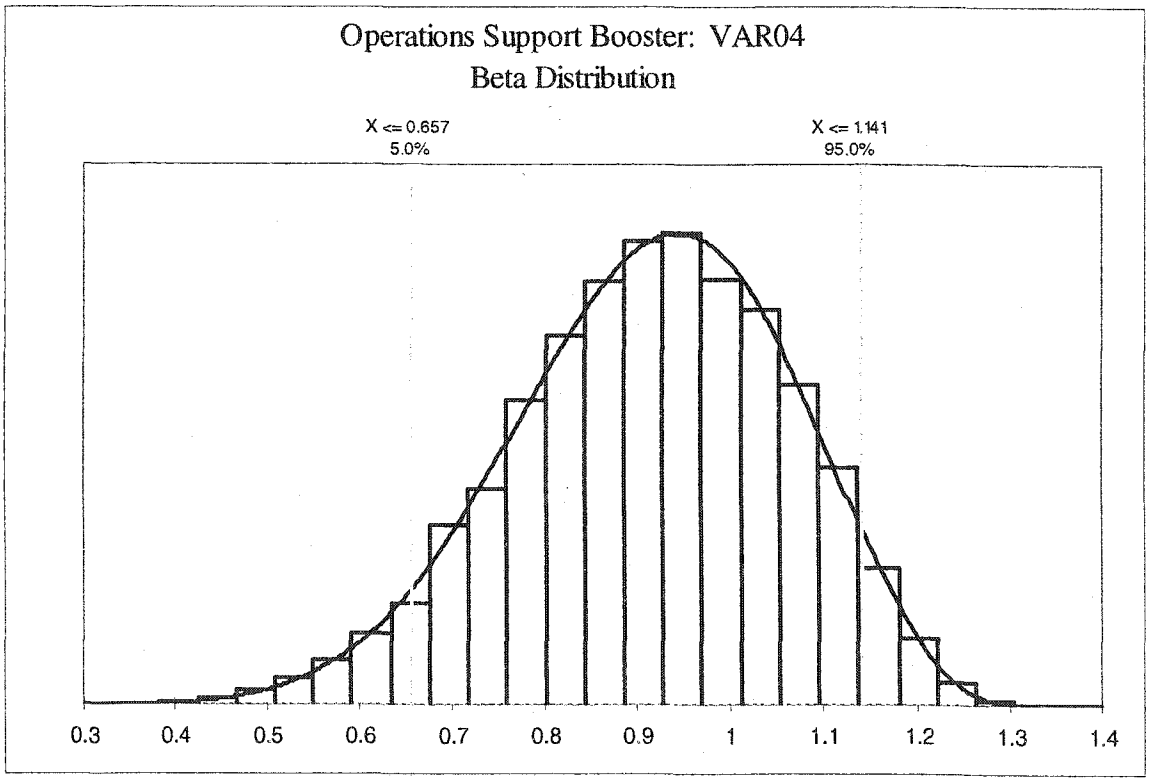
Operations Support Booster: VAR01 Aggregated Response

Fit	Beta
α1	5.140453499
α2	2.609091461
Minimum	69162
Maximum	125877
Mean	106782
Mode	110004
Median	107614
Std. Deviation	9060.8
Variance	82098795
Skewness	-0.4194
Kurtosis	2.6812

	Chi-Sq	A-D	K-S
Test Value	10.84	0.1571	0.02064
P Value	0.9658	N/A	N/A
Rank	1	1	1
# Bins	22	N/A	N/A



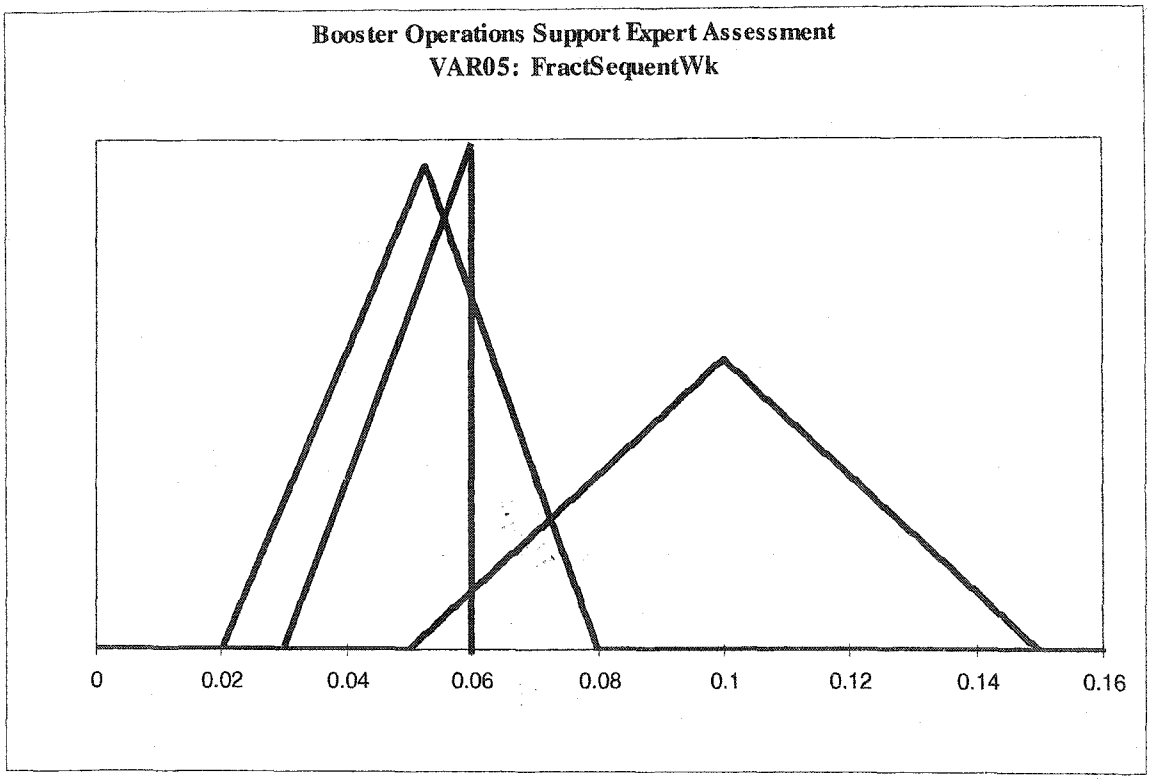
Operations Support Booster: VAR04 Calibrated Expert Assessments



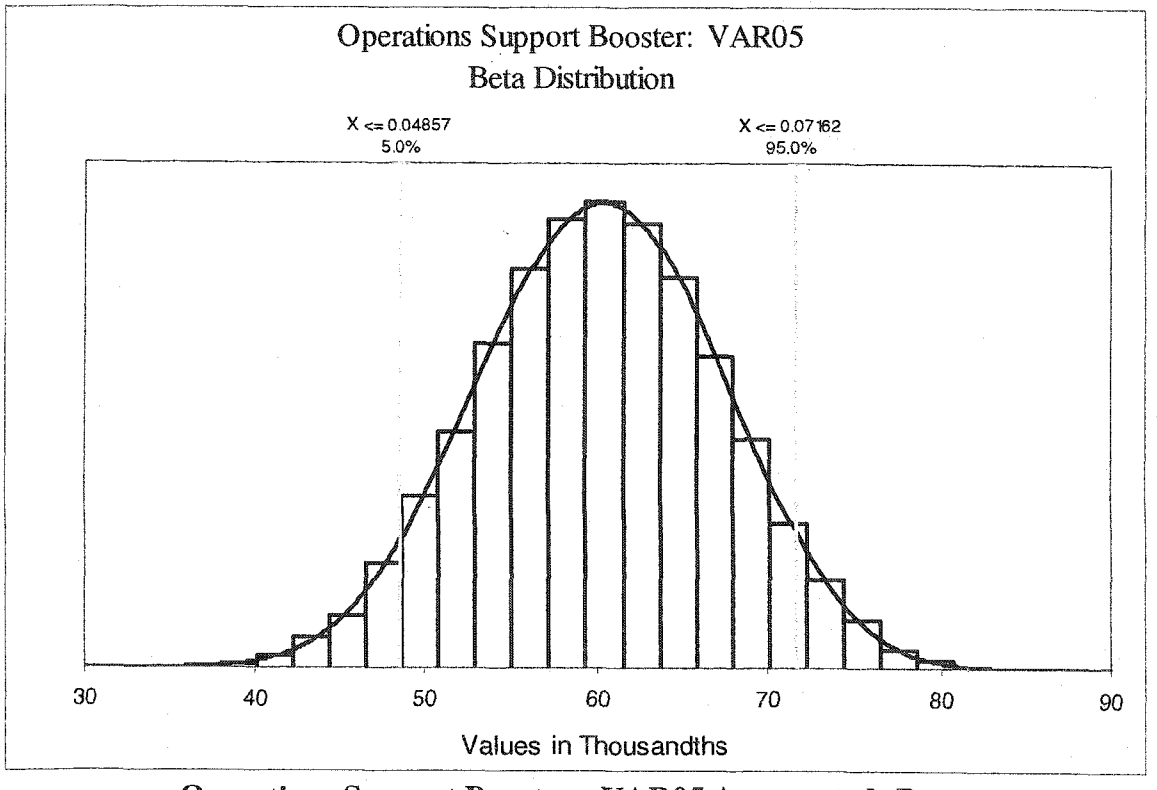
Operations Support Booster: VAR04 Aggregated Response

Fit	Beta
α1	9.337673152
α2	5.03523762
Minimum	0.12871
Maximum	1.3376
Mean	0.91406
Mode	0.94331
Median	0.92266
Std. Deviation	0.14709
Variance	0.021635
Skewness	-0.3005
Kurtosis	2.7823

	Chi-Sq	A-D	K-S
Test Value	93.57	0.3214	0.004555
P Value	0.2958	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



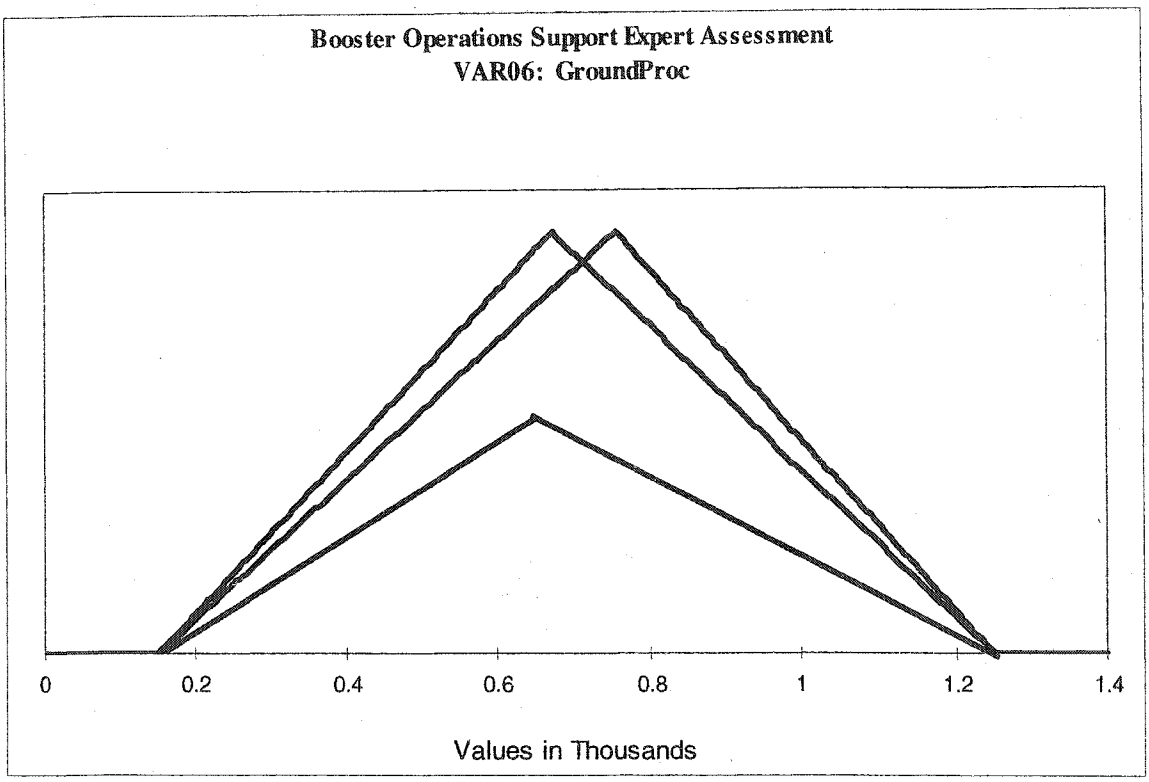
Operations Support Booster: VAR05 Calibrated Expert Assessments



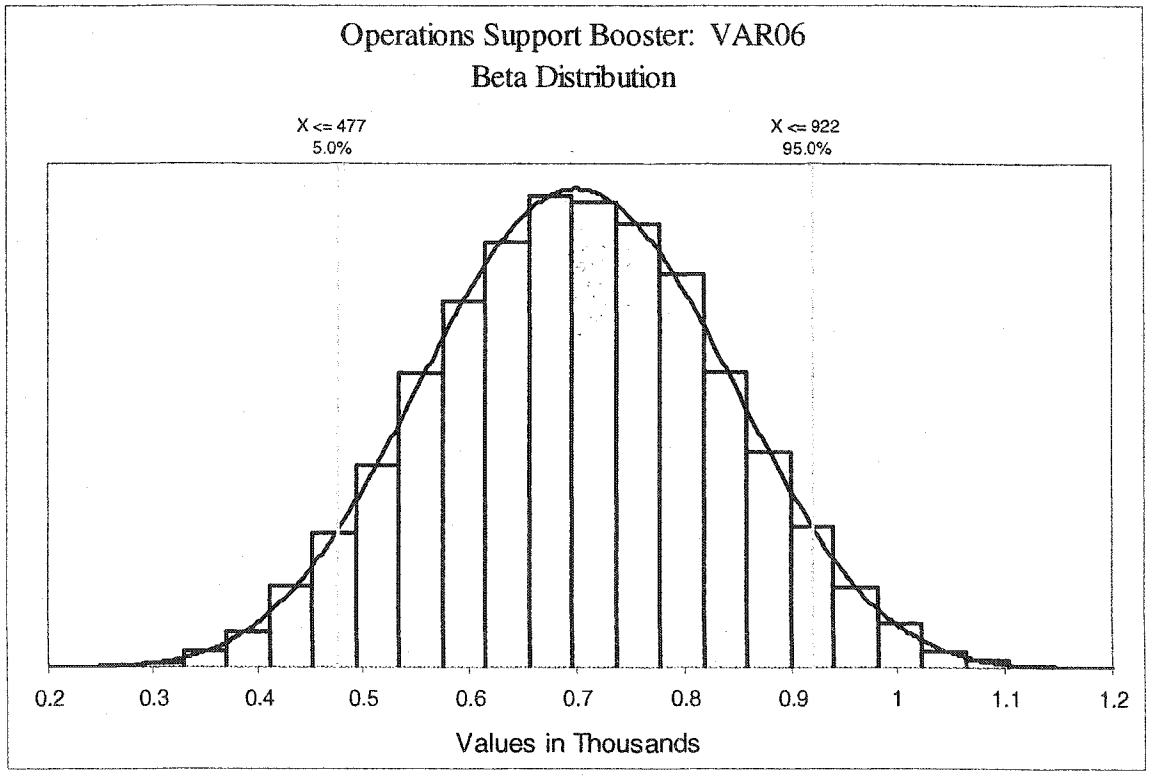
Operations Support Booster: VAR05 Aggregated Response

Fit	Beta
α1	12.13246469
α2	10.78008335
Minimum	0.023926
Maximum	0.09243
Mean	0.0602
Mode	0.060393
Median	0.06026
Std. Deviation	0.0069922
Variance	4.89E-05
Skewness	-0.0464
Kurtosis	2.7716

	Chi-Sq	A-D	K-S
Test Value	85.11	0.1489	0.002961
P Value	0.5372	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



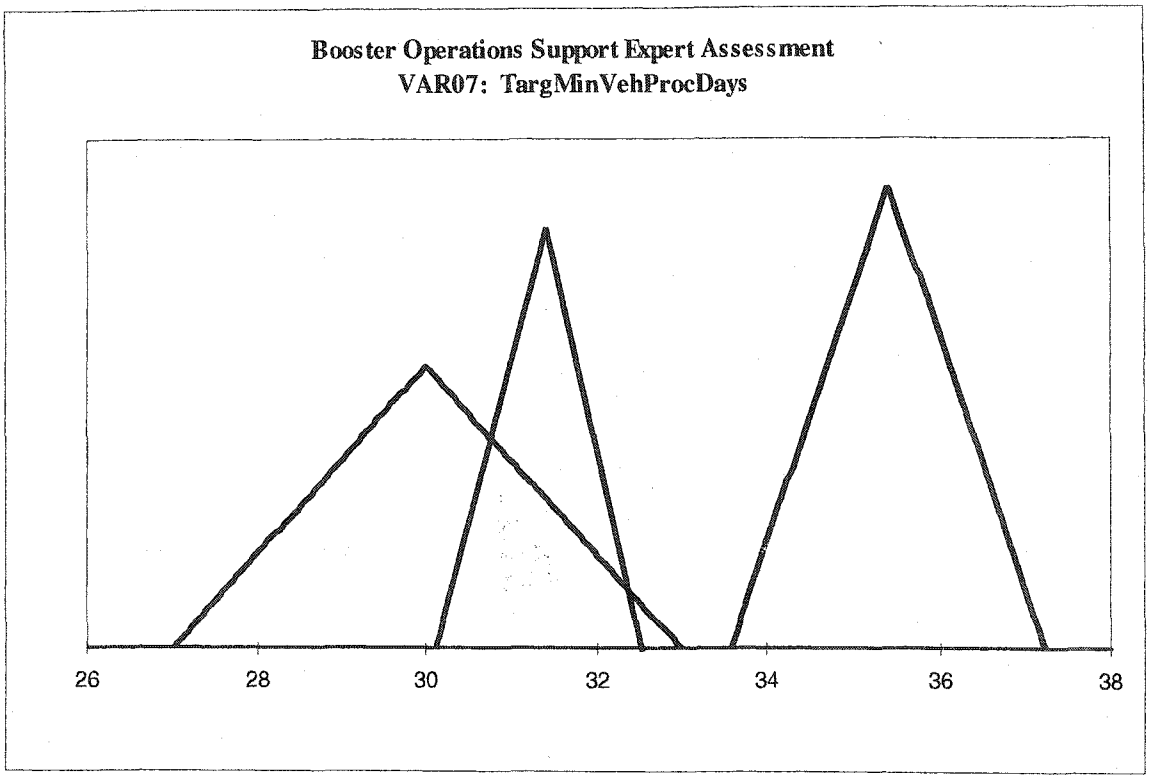
Operations Support Booster: VAR06 Calibrated Expert Assessments



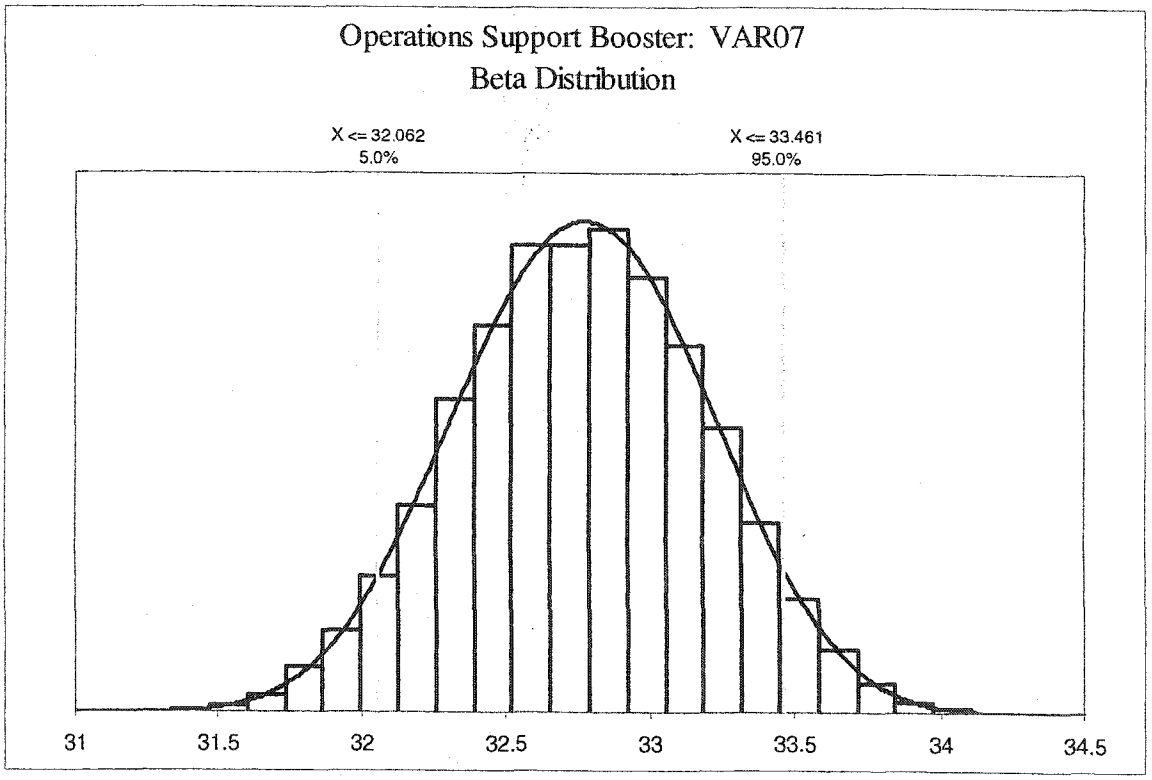
Operations Support Booster: VAR06 Aggregated Response

Fit	Beta
α1	10.03695467
α2	9.924773075
Minimum	78.747
Maximum	1313.6
Mean	699.67
Mode	700.05
Median	699.78
Std. Deviation	134.86
Variance	18187
Skewness	-0.0047
Kurtosis	2.7387

	Chi-Sq	A-D	K-S
Test Value	69.73	0.1982	0.004558
P Value	0.9125	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



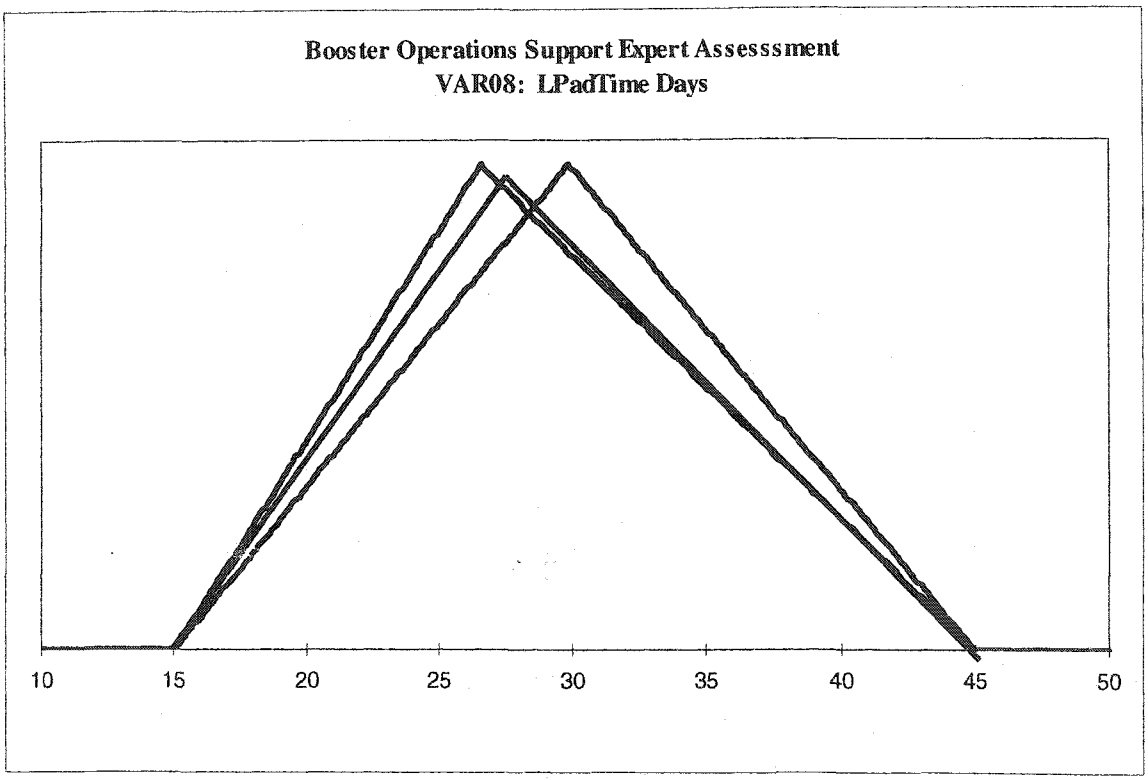
Operations Support Booster: VAR07 Calibrated Expert Assessments



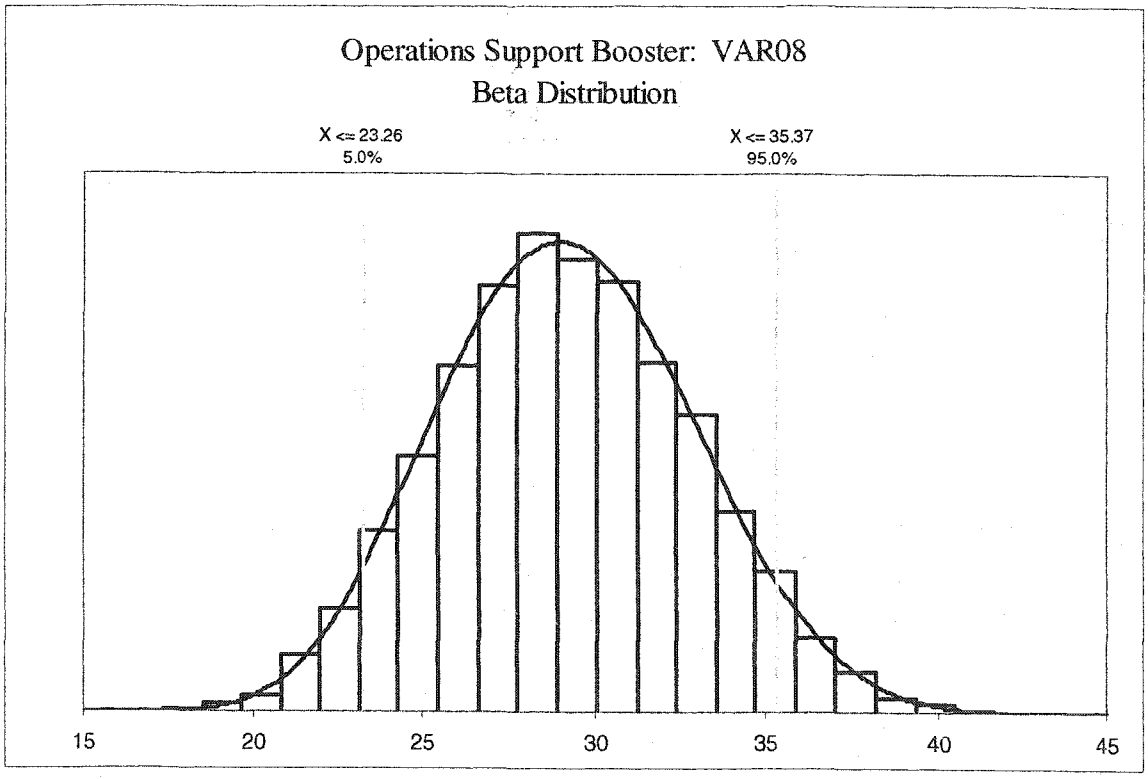
Operations Support Booster: VAR07 Aggregated Response

Fit	Beta
α1	11.43004864
α2	10.59704847
Minimum	30.6516
Maximum	34.726
Mean	32.7658
Mode	32.7735
Median	32.7682
Std. Deviation	0.42423
Variance	0.17997
Skewness	-0.0302
Kurtosis	2.7616

	Chi-Sq	A-D	K-S
Test Value	78.58	0.2857	0.005094
P Value	0.7289	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



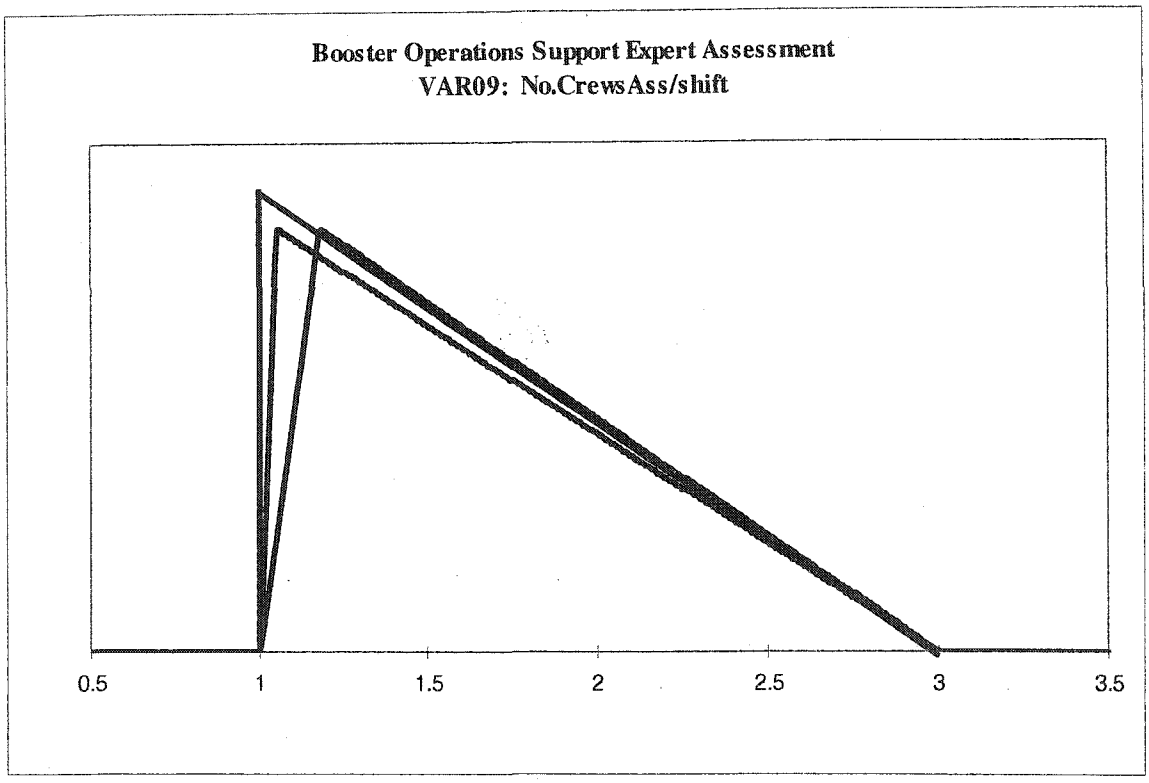
Operations Support Booster: VAR08 Calibrated Expert Assessments



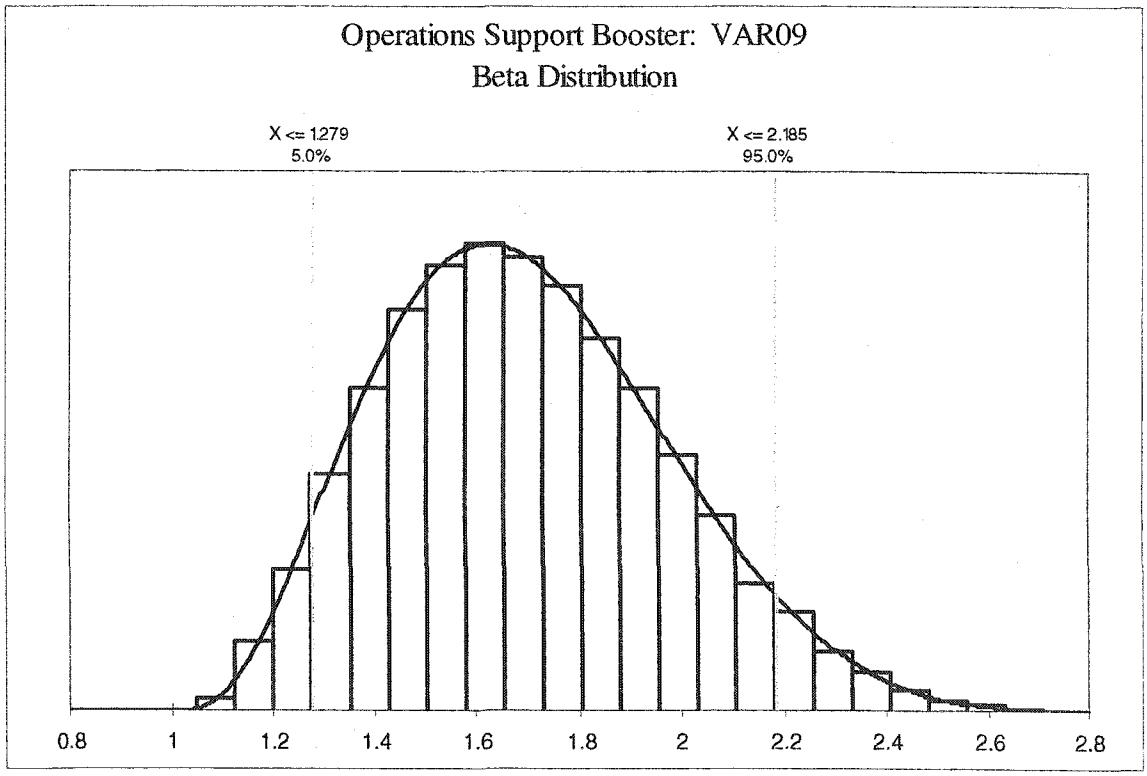
Operations Support Booster: VAR08 Aggregated Response

Fit	Beta
α1	8.626816248
α2	10.60274763
Minimum	14.313
Maximum	47.521
Mean	29.211
Mode	29.013
Median	29.15
Std. Deviation	3.6721
Variance	13.484
Skewness	0.0875
Kurtosis	2.7411

	Chi-Sq	A-D	K-S
Test Value	87.51	0.2318	0.004299
P Value	0.4646	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



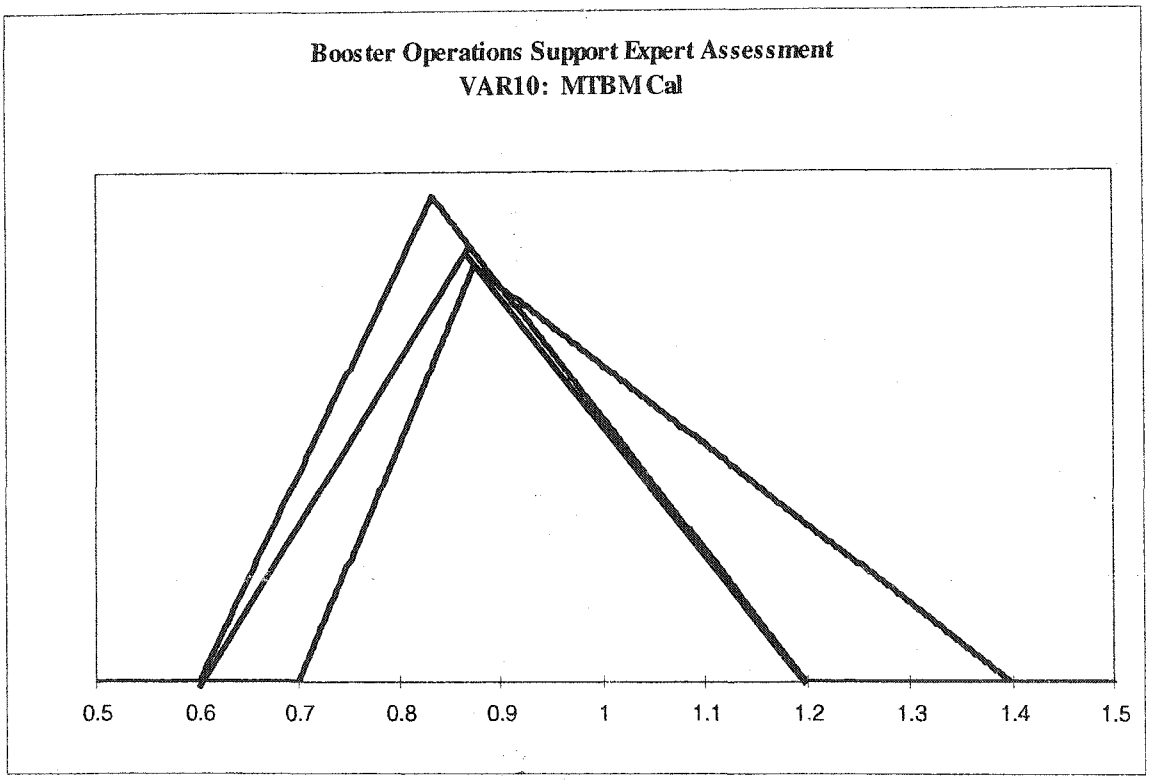
Operations Support Booster: VAR09 Calibrated Expert Assessments



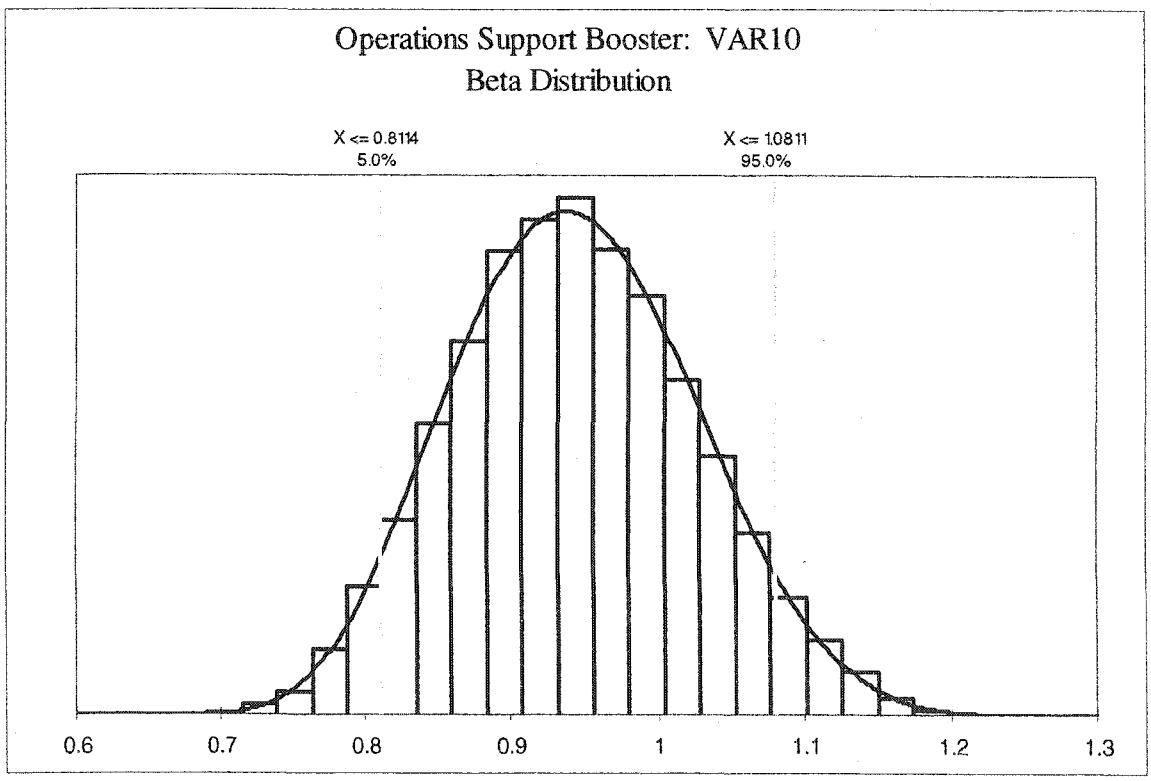
Operations Support Booster: VAR09 Aggregated Response

Fit	Beta
α1	3.663078491
α2	7.063993664
Minimum	1.0167
Maximum	3.0101
Mean	1.6974
Mode	1.625
Median	1.6771
Std. Deviation	0.27604
Variance	0.076198
Skewness	0.3598
Kurtosis	2.7429

	Chi-Sq	A-D	K-S
Test Value	98.54	0.2044	0.004606
P Value	0.1871	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



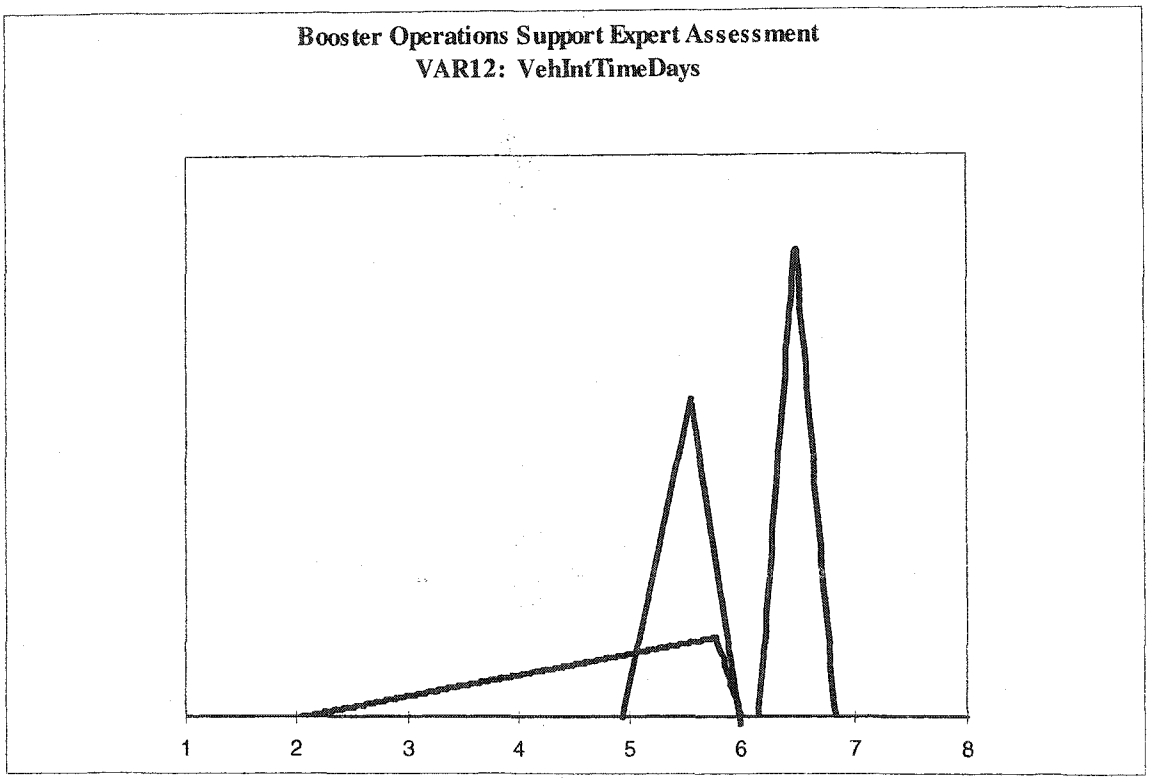
Operations Support Booster: VAR10 Calibrated Expert Assessments



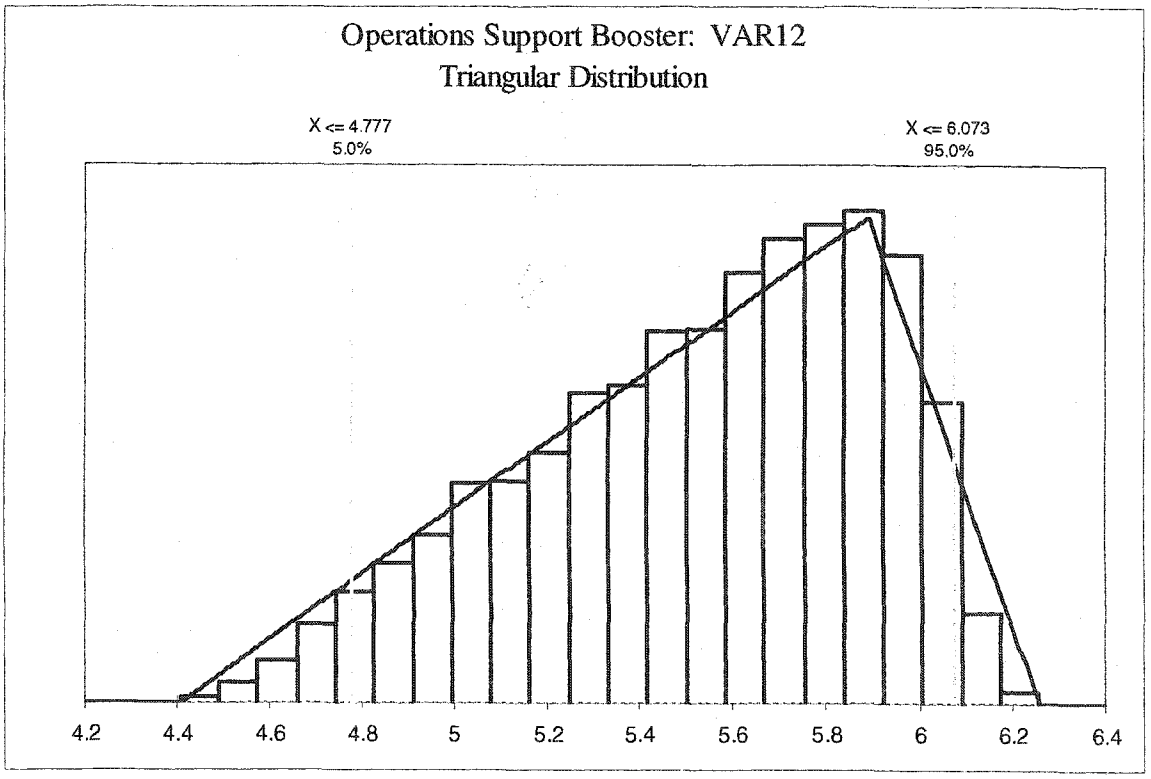
Operations Support Booster: VAR10 Aggregated Response

Fit	Beta
α1	7.024850555
α2	8.888380558
Minimum	0.64424
Maximum	1.32165
Mean	0.94328
Mode	0.93758
Median	0.94158
Std. Deviation	0.081792
Variance	0.0066899
Skewness	0.1083
Kurtosis	2.6994

	Chi-Sq	A-D	K-S
Test Value	86.65	0.4522	0.006008
P Value	0.4904	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



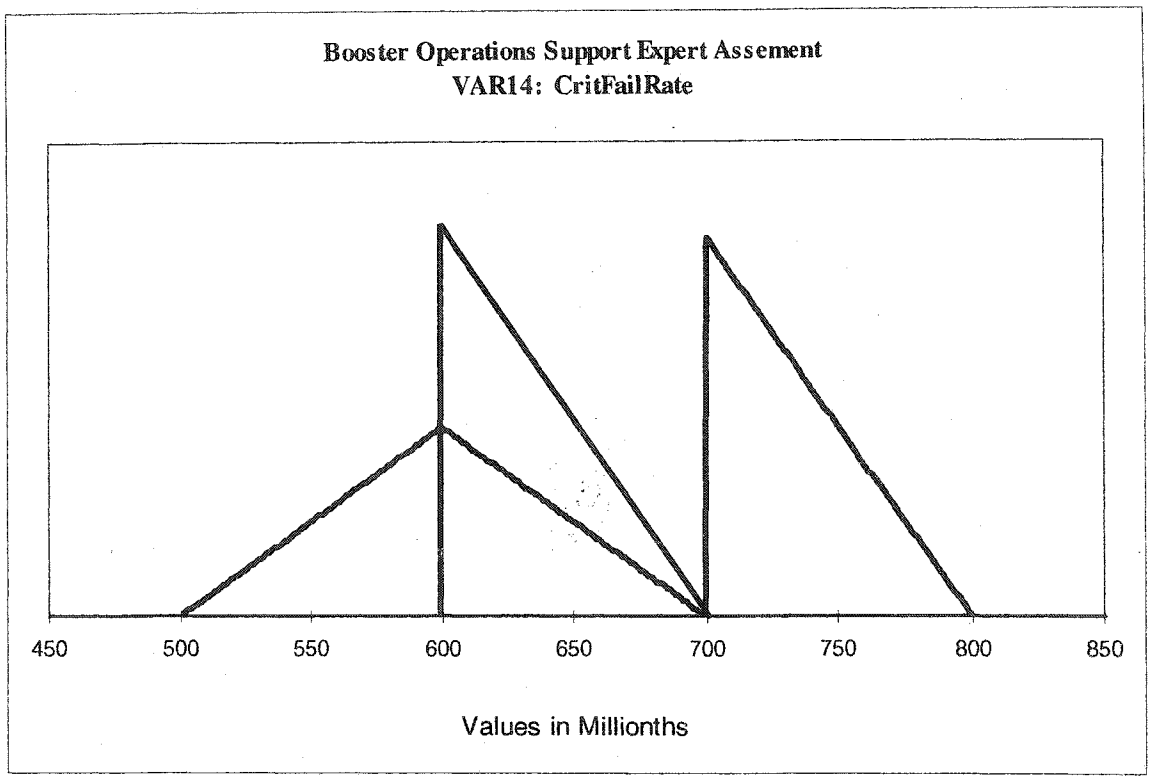
Operations Support Booster: VAR12 Calibrated Expert Assessments



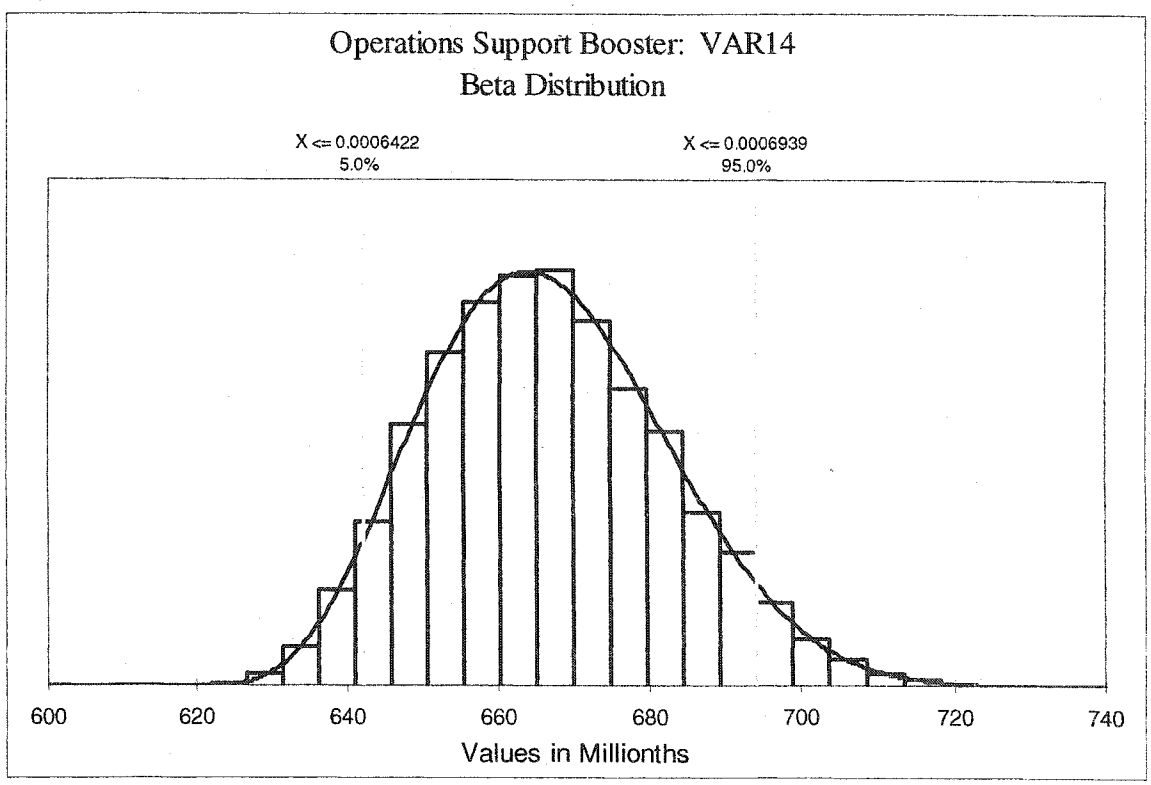
Operations Support Booster: VAR12 Aggregated Response

Fit	Triangular
min	4.407380717
m. likely	5.887459121
max	6.258085919
Mean	5.5176
Mode	5.8875
Median	5.5777
Std. Deviation	0.39976
Variance	0.15981
Skewness	-0.4759
Kurtosis	2.4

	Chi-Sq	A-D	K-S
Test Value	258.2	24.36	0.02559
P Value	0	N/A	N/A
Rank	1	1	1
# Bins	74	N/A	N/A



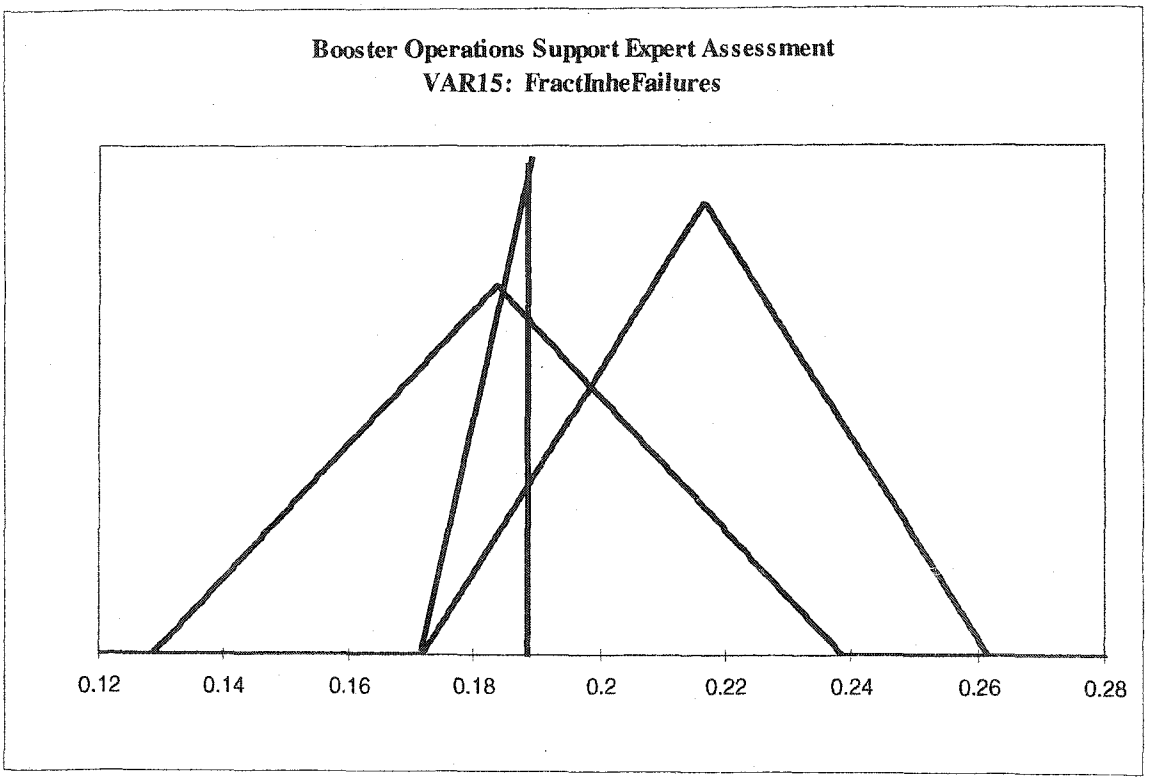
Operations Support Booster: VAR14 Calibrated Expert Assessments



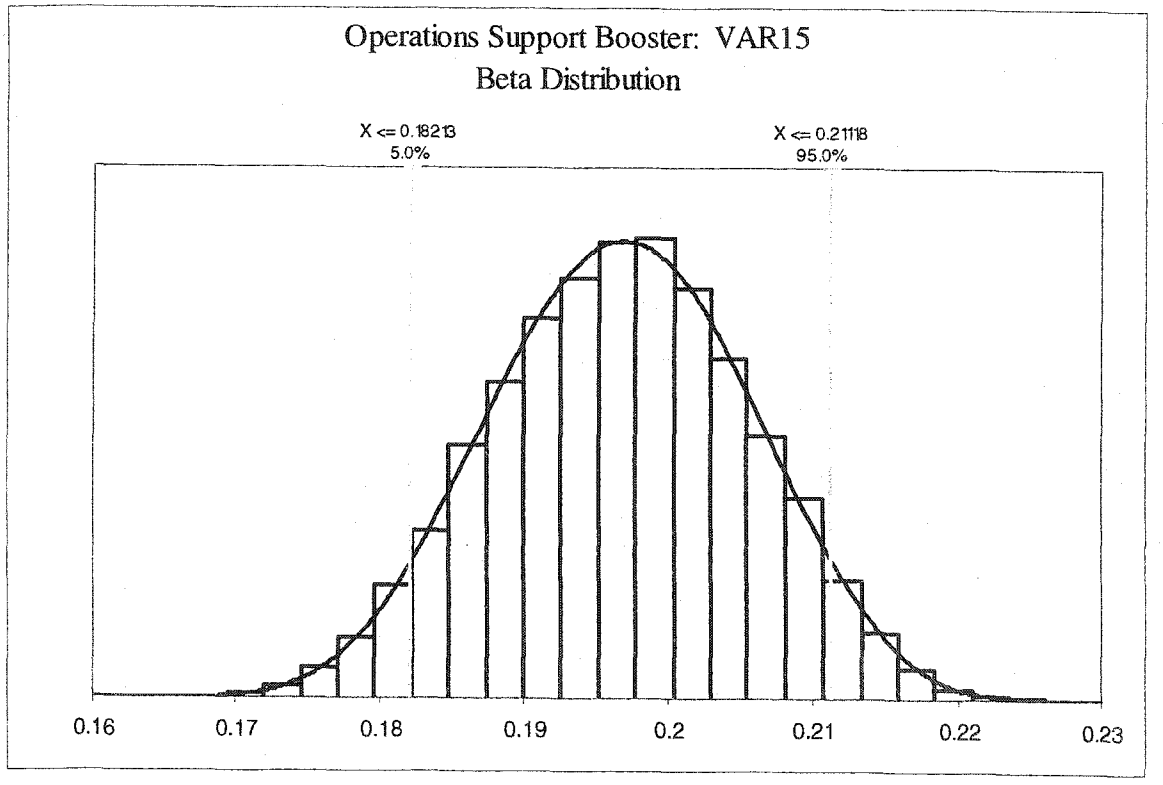
Operations Support Booster: VAR14 Aggregated Response

Fit	Beta
α1	5.950694734
α2	10.64872231
Minimum	0.00061735
Maximum	0.00075493
Mean	0.00066667
Mode	0.000664
Median	0.00066587
Std. Deviation	1.57E-05
Variance	2.47E-10
Skewness	0.2662
Kurtosis	2.7948

	Chi-Sq	A-D	K-S
Test Value	93.41	0.2513	0.005086
P Value	0.2999	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



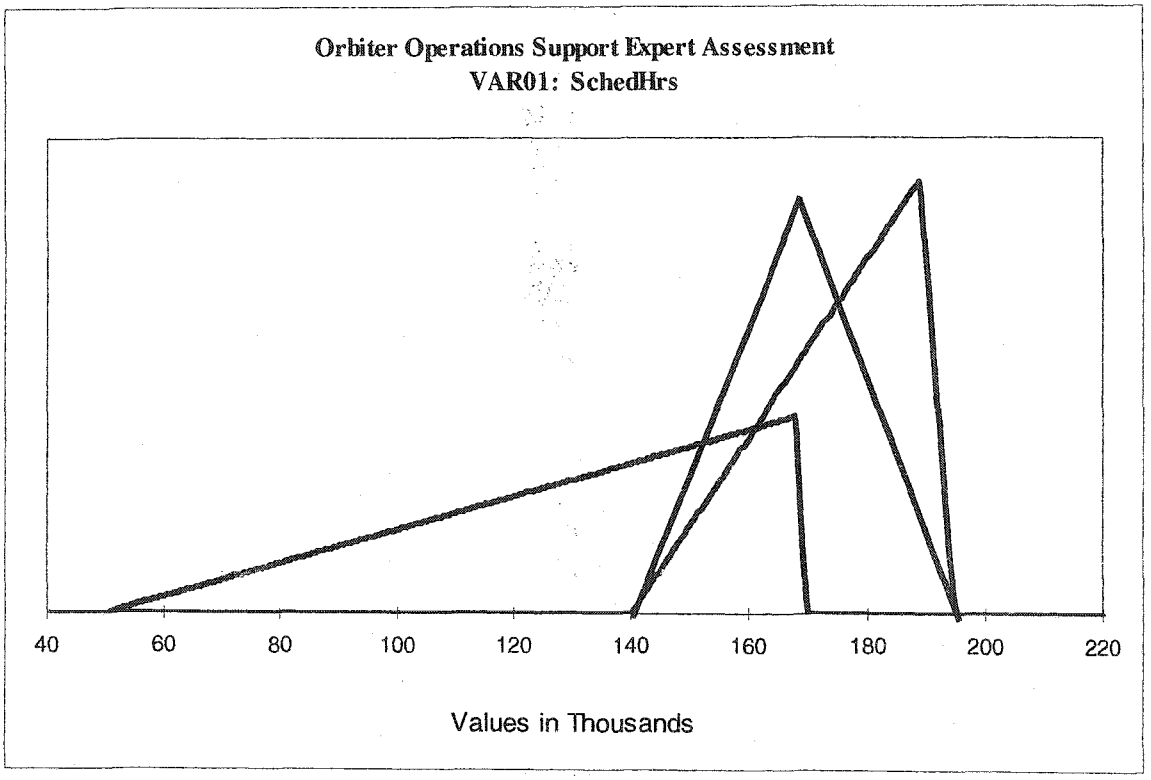
Operations Support Booster: VAR15 Calibrated Expert Assessments



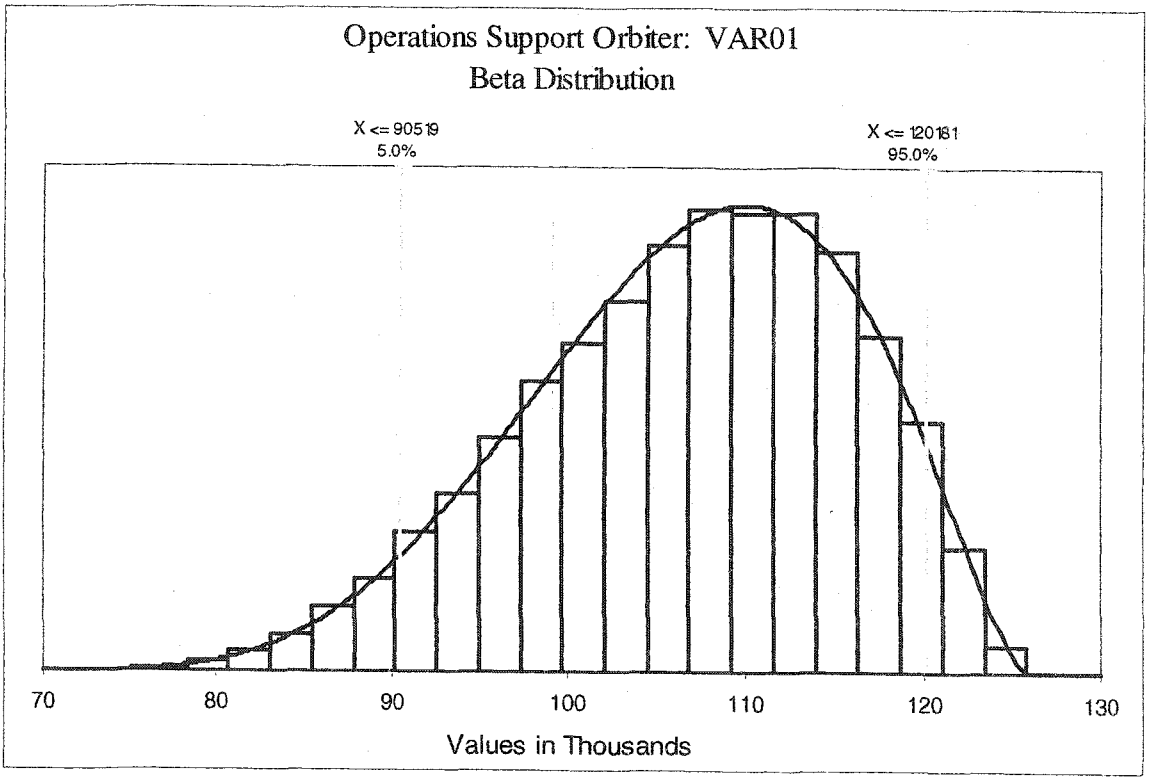
Operations Support Booster: VAR15 Aggregated Response

Fit	Beta
α1	7.614081068
α2	7.14177643
Minimum	0.160643
Maximum	0.230608
Mean	0.196745
Mode	0.19692
Median	0.196797
Std. Deviation	0.0088086
Variance	7.76E-05
Skewness	-0.0303
Kurtosis	2.6634

	Chi-Sq	A-D	K-S
Test Value	69.34	0.3588	0.005855
P Value	0.9178	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



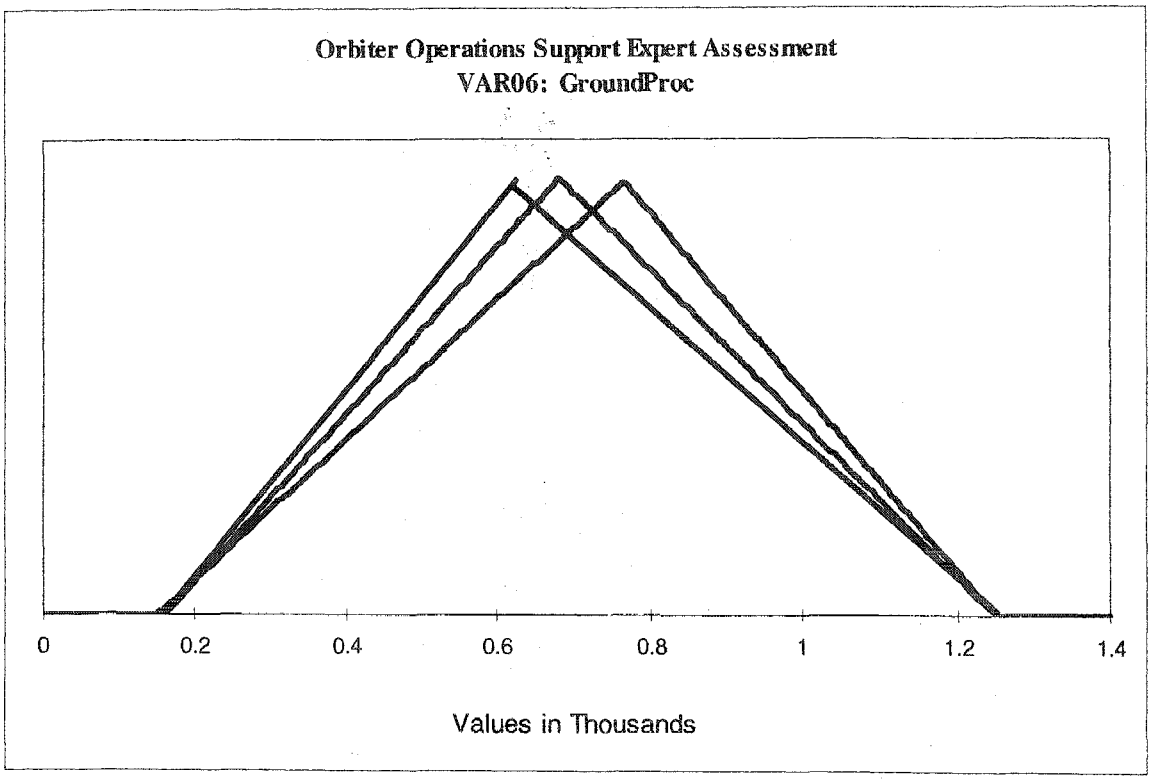
Operations Support Orbiter: VAR01 Calibrated Expert Assessments



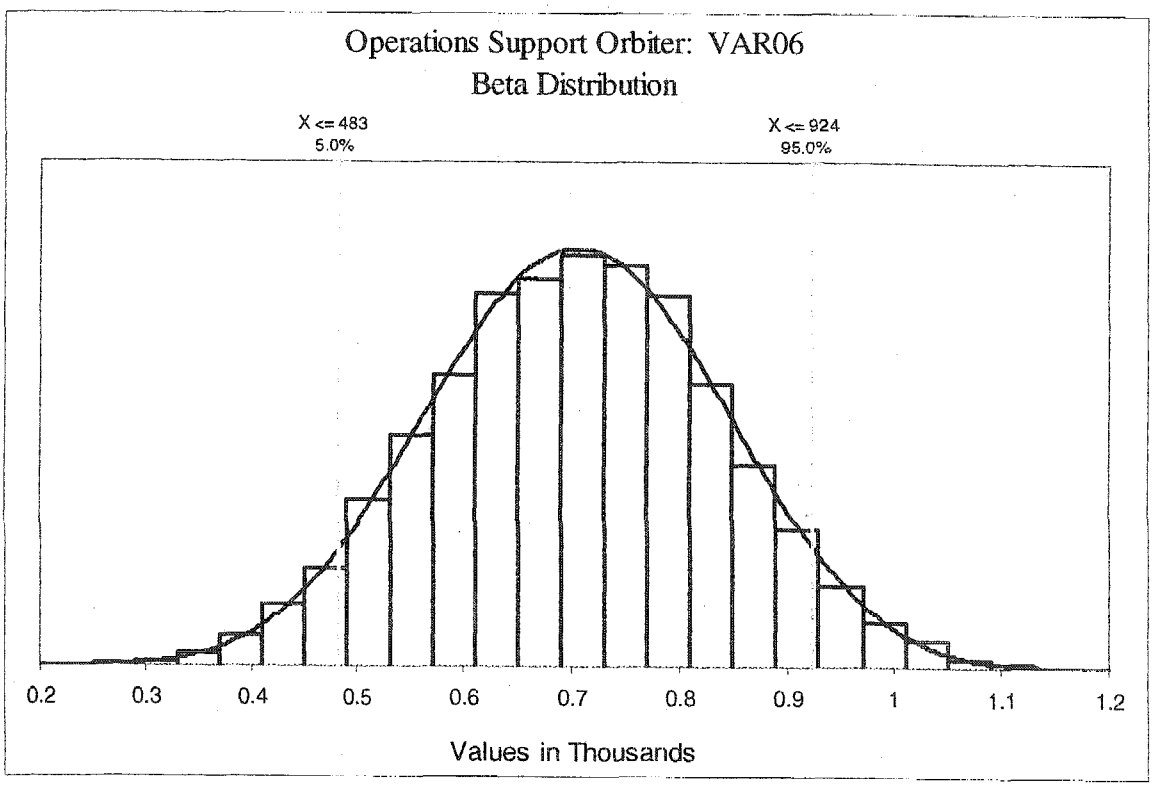
Operations Support Orbiter: VAR01 Aggregated Response

Fit	Beta
α1	5.517947578
α2	2.74263955
Minimum	101153
Maximum	181044
Mean	154519
Mode	158807
Median	155647
Std. Deviation	12363
Variance	152855679
Skewness	-0.4232
Kurtosis	2.7119

	Chi-Sq	A-D	K-S
Test Value	163.9	6.106	0.01569
P Value	1.21E-06	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



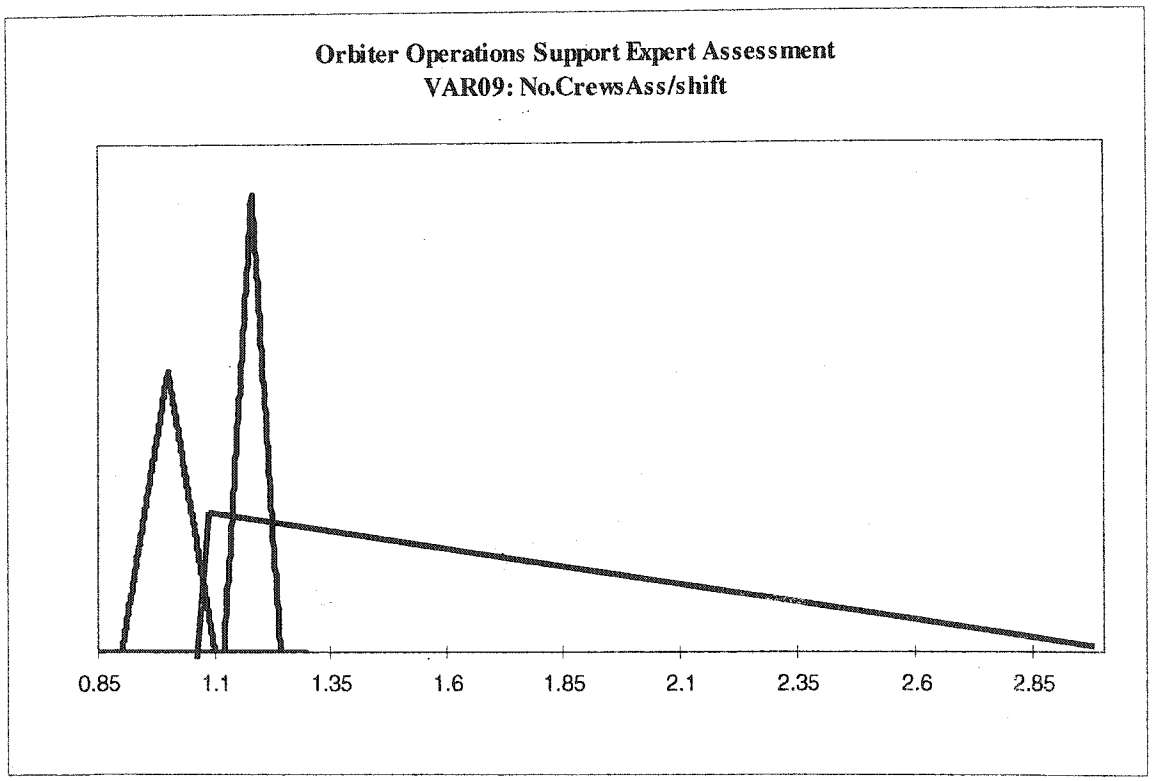
Operations Support Booster: VAR06 Calibrated Expert Assessments



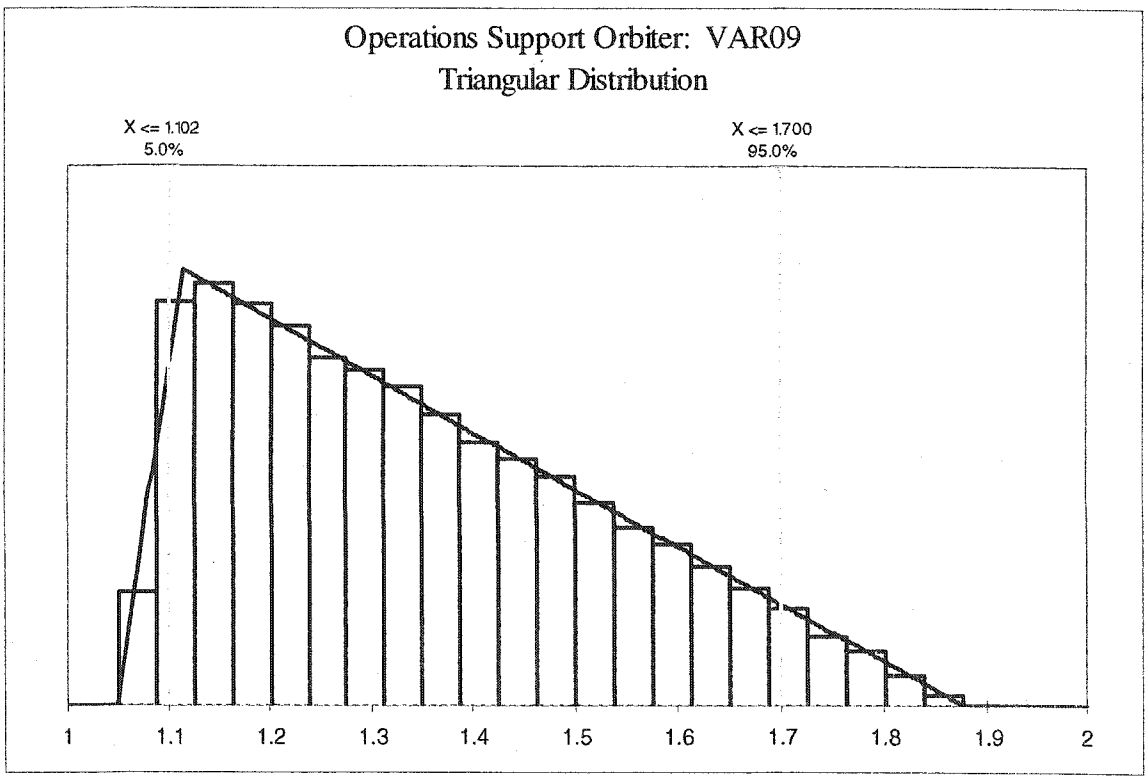
Operations Support Orbiter: VAR06 Aggregated Response

Fit	Beta
α1	12.30813973
α2	11.69327062
Minimum	17.989
Maximum	1355.6
Mean	703.91
Mode	705.47
Median	704.39
Std. Deviation	133.71
Variance	17878.3
Skewness	-0.0197
Kurtosis	2.7784

	Chi-Sq	A-D	K-S
Test Value	106	0.6183	0.006211
P Value	0.0817	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



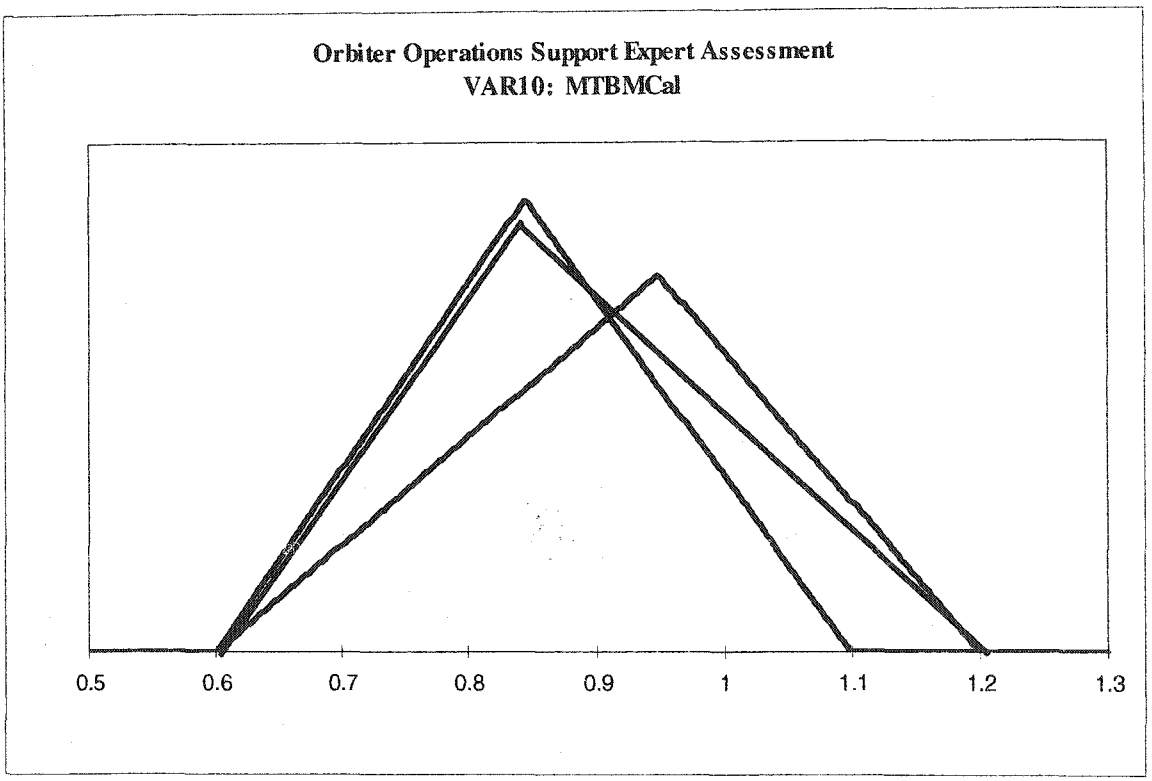
Operations Support Booster: VAR09 Calibrated Expert Assessments



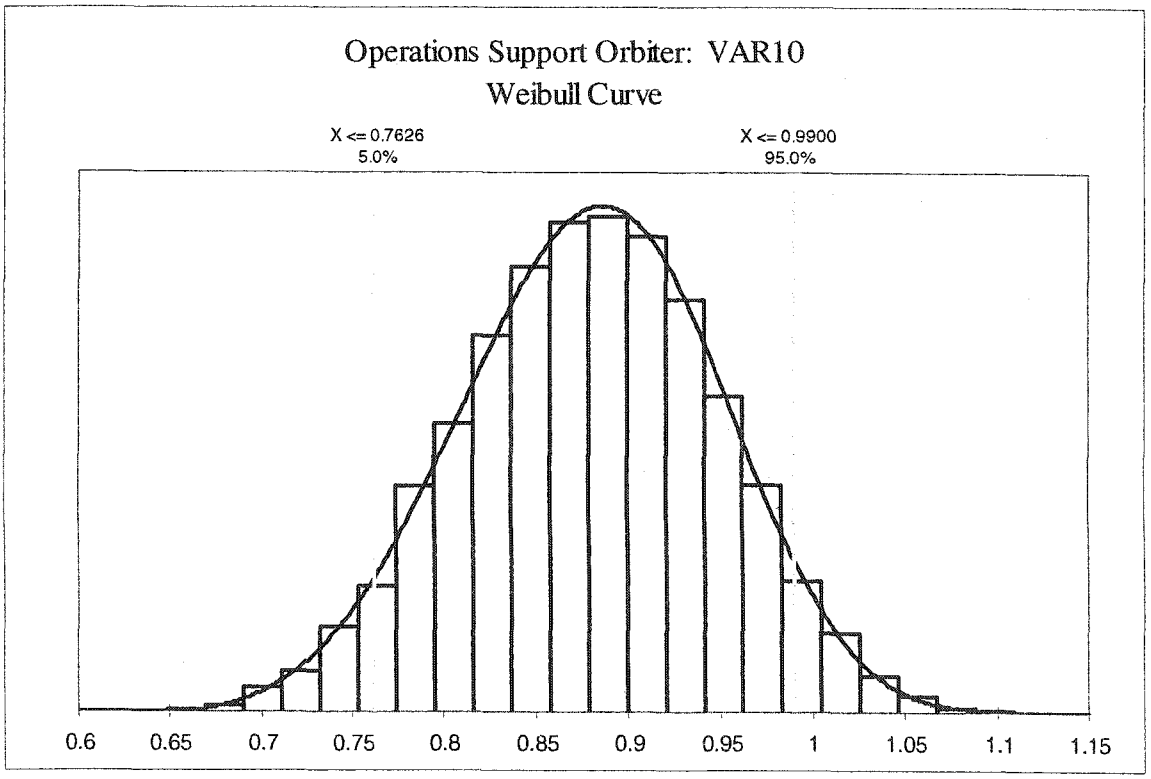
Operations Support Orbiter: VAR09 Aggregated Response

Fit	Triangular
min	1.05088871
m. likely	1.11389028
max	1.877917847
Mean	1.34757
Mode	1.11389
Median	1.31584
Std. Deviation	0.18795
Variance	0.035325
Skewness	0.5538
Kurtosis	2.4

	Chi-Sq	A-D	K-S
Test Value	113.1	1.563	0.006011
P Value	0.0316	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



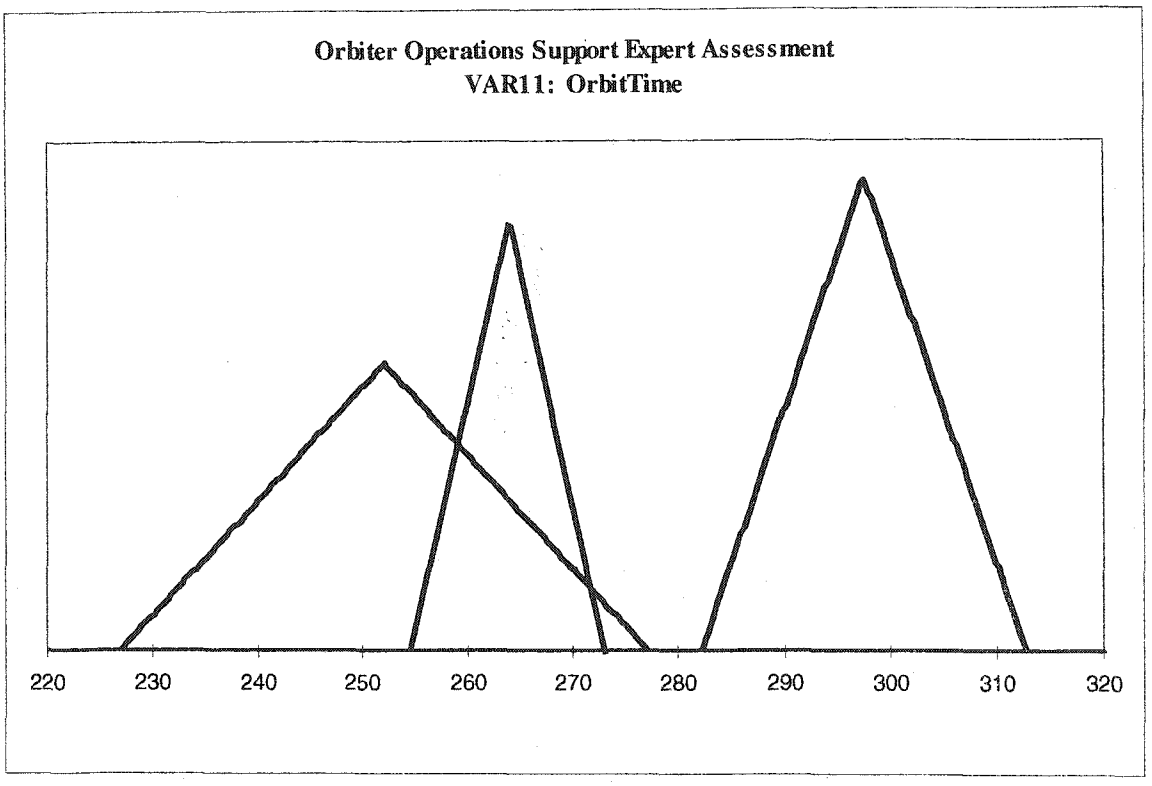
Operations Support Booster: VAR10 Calibrated Expert Assessments



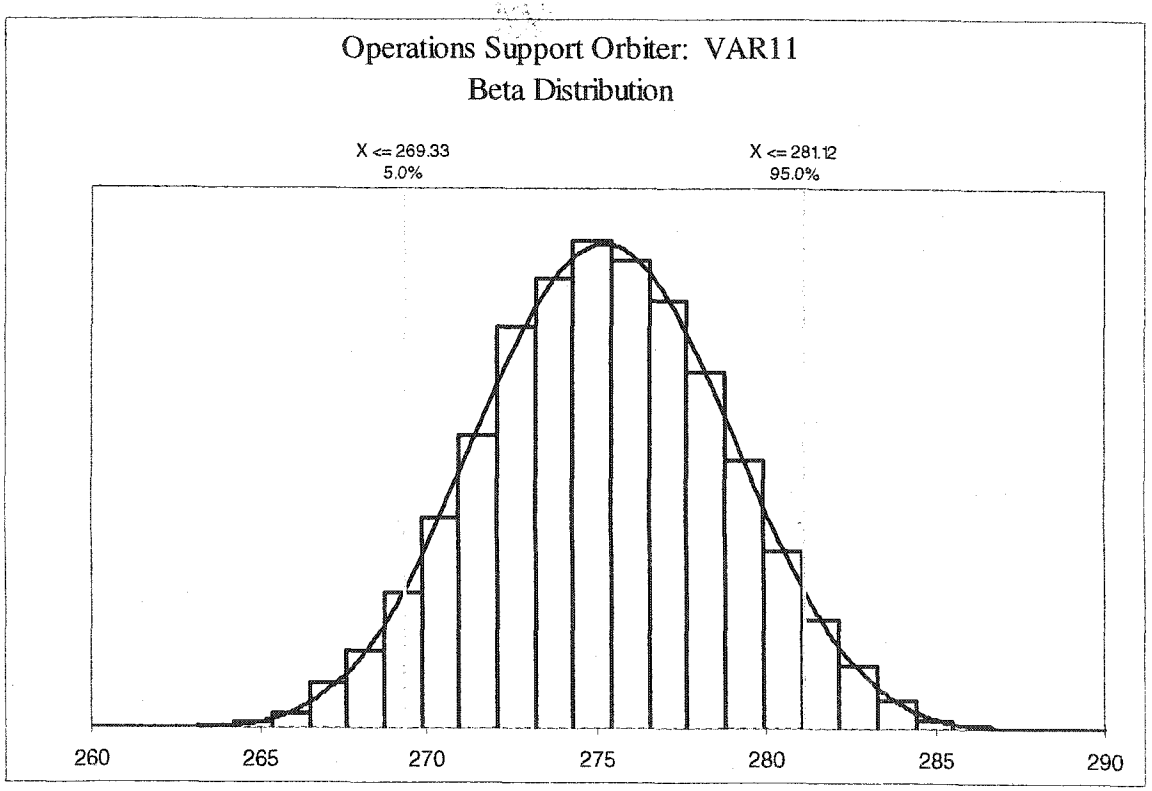
Operations Support Orbiter: VAR10 Aggregated Response

Fit	Weibull
Shift	0.631584272
a	4.039811394
b	0.273194651
Minimum	0.63158
Maximum	1.112
Mean	0.87935
Mode	0.88621
Median	0.88108
Std. Deviation	0.068889
Variance	0.0047457
Skewness	-0.0989 [est]
Kurtosis	2.6678 [est]

	Chi-Sq	A-D	K-S
Test Value	81.7	0.3641	0.00504
P Value	0.6404	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



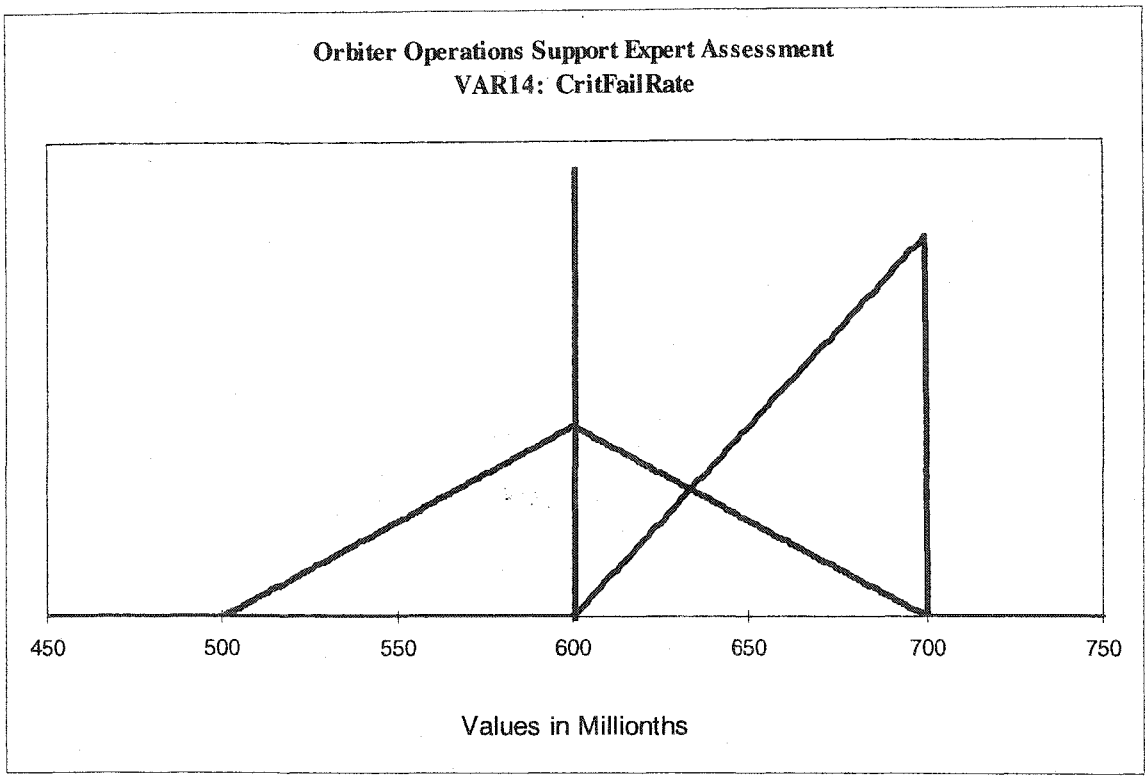
Operations Support Booster: VAR11 Calibrated Expert Assessments



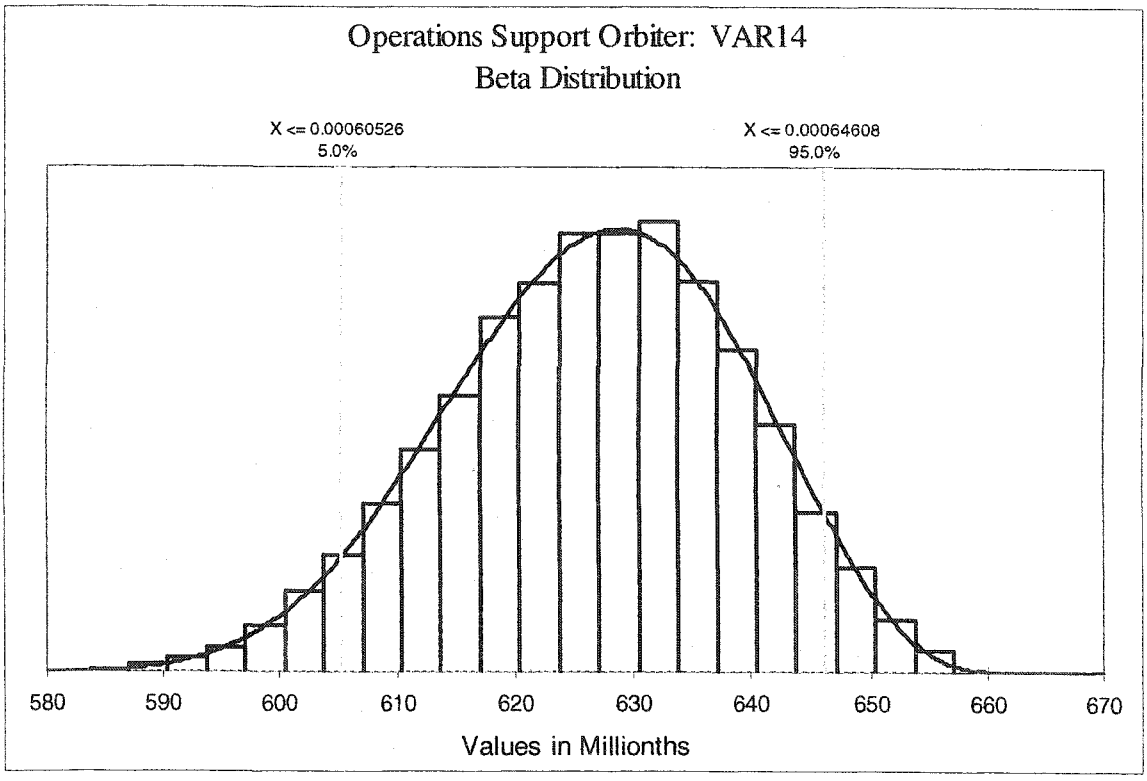
Operations Support Orbiter: VAR11 Aggregated Response

Fit	Beta
α1	10.05353371
α2	9.943891453
Minimum	258.756
Maximum	291.53
Mean	275.233
Mode	275.243
Median	275.236
Std. Deviation	3.5761
Variance	12.788
Skewness	-0.0046
Kurtosis	2.7391

	Chi-Sq	A-D	K-S
Test Value	79.67	0.2768	0.005328
P Value	0.6989	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A



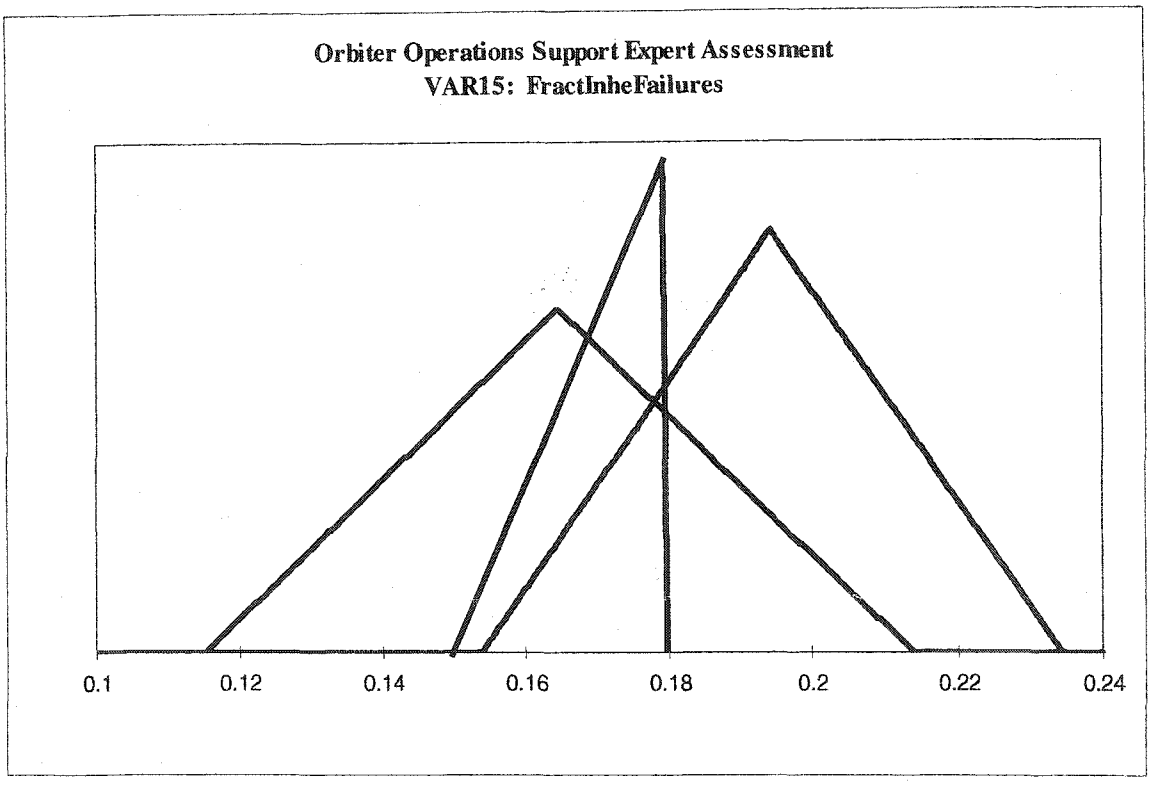
Operations Support Booster: VAR14 Calibrated Expert Assessments



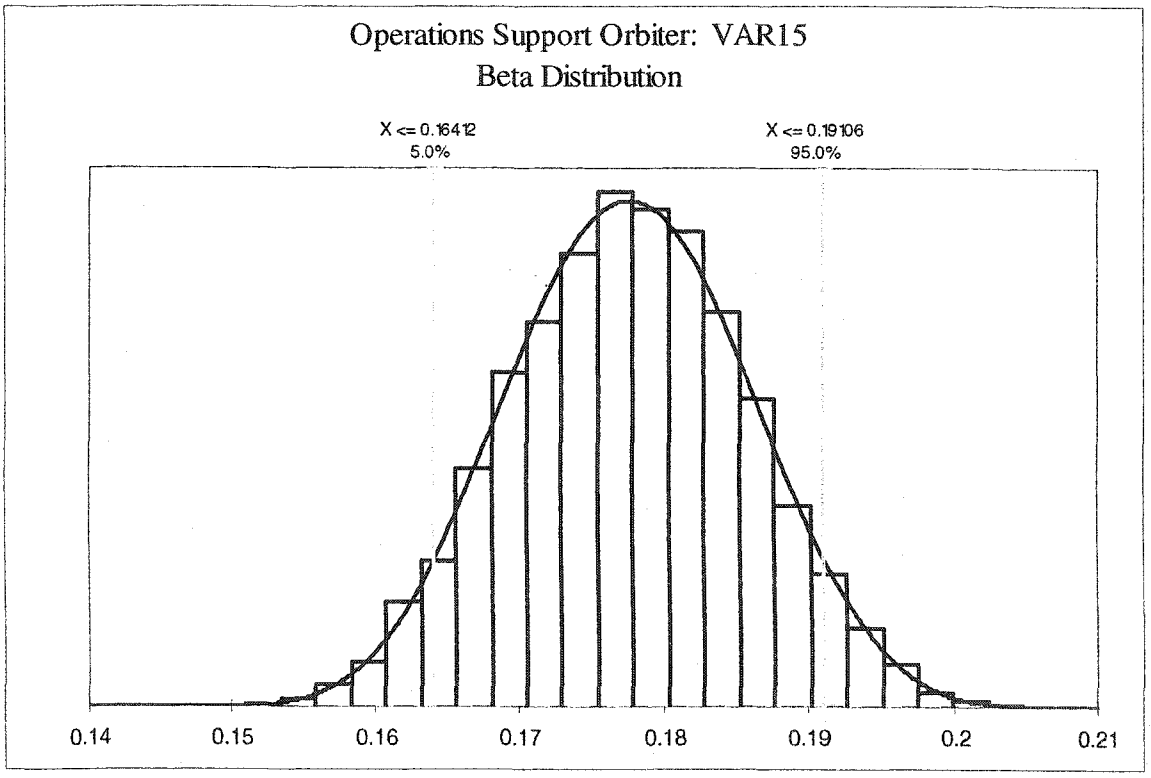
Operations Support Orbiter: VAR14 Aggregated Response

Fit	Beta
α1	8.095481556
α2	5.046064715
Minimum	0.000567606
Maximum	0.000663481
Mean	0.000626667
Mode	0.000628664
Median	0.000627247
Std. Deviation	1.24E-05
Variance	1.54E-10
Skewness	-0.237
Kurtosis	2.7073

	Chi-Sq	A-D	K-S
Test Value	69.92	0.3875	0.005147
P Value	0.5805	N/A	N/A
Rank	1	1	1
# Bins	74	N/A	N/A



Operations Support Booster: VAR15 Calibrated Expert Assessments



Operations Support Orbiter: VAR15 Aggregated Response

Fit	Beta
α1	9.143904816
α2	9.026762083
Minimum	0.141613
Maximum	0.213138
Mean	0.177606
Mode	0.177634
Median	0.177614
Std. Deviation	0.0081677
Variance	6.67E-05
Skewness	-0.0056
Kurtosis	2.7166

	Chi-Sq	A-D	K-S
Test Value	104.9	0.7071	0.007456
P Value	0.0928	N/A	N/A
Rank	1	1	1
# Bins	88	N/A	N/A