Survey Research Handbook

Using Survey Research in the Operational Capability Requirements Study

2 September 2015

Office of Aerospace Studies

Headquarters, Air Force A5R-OAS
1655 1st Street SE
Kirtland AFB, NM 87117-5522
(505) 846-8322, DSN 246-8322

DISTRIBUTION A. Approved for public release: distribution unlimited.
377ABW-2015-0768

THIS PAGE INTENTIONALLY LEFT BLANK

Preface

In operational capability requirements studies, it is not uncommon for the analyst to collect some of the data using one or more forms of survey research. Given that the data is affected by the approach used to gather it, it is important that the analyst has a good understanding of the fundamental principles of survey research. Regardless of whether the survey is administered by an interviewer or self-administered, good questions and proper administration are essential to collecting meaningful data.

A survey is one of many analytic methods used in the world of operational capability requirements studies. Some problems and questions lend themselves well to surveys while other questions and problems are best answered through other methods. It is not the purpose of this handbook to advocate for or against survey research, but rather to describe the appropriate uses of survey research and provide insights into planning and conducting survey research in the Capabilities-Based Assessment (CBA), Analysis of Alternatives (AoA), and other types of operational capability requirements studies.

This handbook describes the fundamental principles of survey research that are necessary for ensuring questions are both reliable (i.e., provide consistent responses in comparable situations) and valid (i.e., answers correspond to what they are intended to measure). With expert elicitation being a special form of survey research, this handbook also presents an approach to conducting expert elicitation in operational capability requirements studies. Lastly, this handbook is designed to supplement other OAS handbooks by providing a comprehensive discussion of survey research principles.

OAS promotes an effective dialogue with the analyst community involved in planning and conducting operational capability requirements studies. We encourage you to contact us and ask questions, or to provide always-appreciated feedback. If you have questions regarding specific information in the handbook or you have suggestions for improvements, please contact OAS at (OAS.DIR@us.af.mil) or 505-846-8322 (DSN 246).

Jeff Erikson

Director, Office of Aerospace Studies

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1 Survey Research 1
1.1 Introduction 1
1.2 Types of Measurement and Data 2
1.3 Designing Good Questions 4
1.4 Question Sequence 9
1.5 Scale Considerations 10
1.5.1 Nominal and Ordinal Scales 11
1.5.2 Likert Scale 11
1.5.3 Neutral Midpoint 12
1.5.4 Not Applicable or Don't Know as a Selection 12
1.6 Designing the Survey 13
1.6.1 Interview-Administered Survey 13
1.6.2 Self-Administered Survey 14
1.7 Pretesting the Survey 15
1.8 Sampling 15
1.8.1 Nonprobability Sampling 16
1.8.2 Determining the Sample Size for Nonprobability Sampling 16
1.8.3 Probability Sampling 17
1.8.4 Determining the Sample Size for Probability Sampling 17
1.9 Survey Question Validity and Reliability 22
1.10 Administering the Survey 23
1.10.1 Interview-Administered Surveys 23
1.10.2 Self-Administered Surveys 24
1.11 Data Analysis 25
1.11.1 Central Tendency and Dispersion Measure Criteria 27
1.11.2 Ordinal Data Analysis 28
1.11.3 Analyzing Data with "Not Applicable" or "Don't Know" Selections 30
1.11.4 Sensitivity Analysis 32
2 Expert Elicitation 33
2.1 Introduction 33

2.2 What is an Expert? 33
2.3 Expert Judgment 34
2.4 An Expert Elicitation Approach for Capability Requirements Studies 34
2.4.1 Step 1: Identify the Need for Expert Elicitation 34
2.4.2 Step 2: Develop the Questions 35
2.4.3 Step 3: Select the Experts 36
2.4.4 Step 4: Prepare the Experts 37
2.4.5 Step 5: Conduct the Elicitation 38
2.4.6 Step 6: Aggregate the Data 42
2.4.7 Step 7: Report the Results 43
2.5 Summary 43
Appendix A: Acronyms A-1
Appendix B: Information Sources B-1
Appendix C: References C-1
Appendix D: Glossary D-1
Appendix E: Examples of Five, Six, and Seven Point Scale Descriptors E-1
Appendix F: Determining Sample Size Involving Means F-1

List of Figures

Figure 1-1: Bar Chart Example 26
Figure 1-2: Box and Whisker Plot Example 27
Figure 1-3: Bar Chart without “Don’t Know” Selections 31
Figure 1-4: Bar Chart with “Don’t Know” Selections 31

List of Tables

Table 1-1: Levels of Measurement 3
Table 1-2: Four Combinations of Measurement and Data 4
Table 1-3: Example of Closed and Open Versions of the Same Question 5
Table 1-4: Examples of Incomplete and Complete Questions 5
Table 1-5: Examples of Non-Specific and Specific Questions 6
Table 1-6: Example of a Flawed Assumption Underlying a Question 7
Table 1-7: Example of Lengthy and Brief Versions of the Same Question 8
Table 1-8: Examples of Negative, Positive, and Neutral Questions 8
Table 1-9: An Example of a Double-Barreled Question 9
Table 1-10: An Example of Leading and Loaded Questions 9
Table 1-11: Example of Previous Question Bias 10
Table 1-12: Nominal and Ordinal Based Scales 11
Table 1-13: Example of a Likert Scale 12
Table 1-14: Item with “Don’t Know” Included as a Selection 13
Table 1-15: Sampling Techniques Appropriate for Each Population Type 18
Table 1-16: Examples of Questions Designed to Elicit Attribute Data and Variable Data 18
Table 1-17: Confidence Ranges for Variability Attributes to Sampling 22
Table 1-18: Assigning Numerical Values to the Scale Descriptors 25
Table 1-19: Examples of Central Tendency and Dispersion Measure Criteria 28
Table 1-20: Inappropriate Analysis of Ordinal Data Example 29
Table 1-21: Appropriate Analysis of Ordinal Data Example 30
Table 2-1: Examples of Information Elicited from Experts 34
Table 2-2: Examples of Attributes 35
Table 2-3: Examples of Selection Criteria 37
Table 2-4: Common Forms of Bias 40
Table 2-5: Examples of Elicitation Methods for Group Interview Situations 41
Table 2-6: Examples of Information Elicited from Experts 42

1 Survey Research

After a brief introduction of survey research and the types of measurement and data, this chapter describes some fundamental principles of designing good questions. This is followed by a discussion of nominal and ordinal scales and guidelines for developing scales for questions. Finally, several principles are discussed that the analyst or interviewer[1] should consider when designing, pretesting, and administering a survey.

1.1 Introduction

Survey research is used to collect data and information of some aspect for analysis. All surveys, whether conducted through interviews (personal or group) or self-administered with questionnaires (electronic or paper), rely on using questions to collect information from respondents.

There are two main reasons for using survey research in operational capability requirements studies.[2] First, an analyst may want to collect someone's opinion or judgment regarding a specific question or issue. For example, an analyst may want to identify the risks associated with each of the alternatives in an AoA study. In this situation, the analyst could conduct a survey with a group of experts to identify the risks. Survey research is specially designed for collecting an individual's opinion or judgment about something. In other situations, the analyst may want to use survey data to supplement data collected through other methods. In this case, the survey data may be used to corroborate the results in a study.

The second main reason for using survey research is to collect data when it cannot be collected through other methods. For example, it may not be possible to estimate a system's reliability through modeling and simulation, parametric analysis, or comparative analysis. In this situation, survey research may be used to elicit reliability estimates from qualified experts. Survey data is often used to assess measures in the CBA, AoA, or other type of capability requirements study when other data collection methods are impractical or not possible.[3]

---

[1] The term "interviewer" refers to the individual who is conducting the survey. The term "analyst" refers to one or more individuals or members of a study team responsible for conducting all or some aspect of an analysis. The interviewer and analyst may be the same individual(s).

[2] Operational capability requirements studies include the Capabilities-Based Assessment, Analysis of Alternatives, and other types of requirements-related studies. Requirements refer to operational requirements that are typically identified during the requirements identification phase of the requirements development process. To annotate this meaning, the term "operational capability requirements" and a shortened version of the term "capability requirements" are used throughout this handbook.

[3] Some examples of measures that may be assessed using survey data include the measure of effectiveness (MOE), measure of suitability (MOS), and measure of performance (MOP). For more information about developing and analyzing measures for the CBA, AoA, or other types of capability requirements studies, please see *The Measures Handbook*, Office of Aerospace Studies, 6 August 2014.

1.2 Types of Measurement and Data

As shown in Table 1-1, there are four general levels of measurement: nominal, ordinal, interval, and ratio. The levels of measurement range in sophistication from low (nominal) to high (ratio). Nominal and ordinal levels of measurement are commonly referred to as qualitative measures. Interval and ratio levels of measurement are referred to as quantitative measures. Since the measurement characteristics are different at each level, there are particular statistics that are appropriate for each level.

When determining the levels of measurement that will be used, the analyst must consider various factors such as the attribute being measured, purpose of the measurement (e.g., counting objects in categories, attaining a rank order), and data collection requirements and constraints. The analyst should strive to use the highest possible levels of measurement that are suitable for the study. It is important that the analyst has an understanding of the levels of measurement to ensure the appropriate statistics are used.

There are two general types of data: objective and subjective. Objective data is collected with instrumentation or some other direct means. If collected without the use of personal judgment, data collected through direct observation is considered an objective data collection method (e.g., pilot report of airspeed read directly from a display). In contrast, subjective data is based on an individual's judgment or opinion about something. Table 1-2 shows the four possible combinations of measurement and data.

Table 1-1: Levels of Measurement

Category	Level	Description	Appropriate Statistics	
			Descriptive	Inferential
Qualitative	Nominal	Data are assigned the same value or symbol if they have the same attribute. Example: 1 – Male, 2 – Female	Mode, percentages, and frequencies.	Chi-square, binomial test, McNemar test, and Cochran Q test.
	Ordinal	Data are assigned numbers/symbols such that the order of the numbers/symbols reflects an order relation based on the attribute. Example: 1 – Bad, 2 – Medium, 3 – Good	All statistics permitted for nominal scales plus percentile (e.g., median (50th percentile), 80th percentile, and 95th percentile).	Mann-Whitney U-test, Kruskal Wallis test, Friedman two-way analysis of variance (ANOVA), and rank-order correlation.
Quantitative	Interval	Data are assigned numbers such that differences between numbers represent equivalent intervals. Example: Temperature in degrees Fahrenheit*	All statistics permitted for ordinal plus mean, standard deviation, and range.	Product-moment correlation, Z-test, T-test, F-test, factor analysis, and ANOVA.
	Ratio	Data are assigned numbers that have all the features of interval measurement as well as meaningful ratios between arbitrary pairs of numbers. There is a rational zero point for the scale which is necessary for the ratio statements to have meaning. Example: Length in feet; duration in seconds	All statistics permitted for interval scales plus geometric mean and harmonic mean.	Same as interval plus coefficient of variation.

* The Fahrenheit scale is an example of an interval scale because each degree represents an equivalent interval and the zero point (zero degrees) is not a true zero point (i.e., there is still heat at zero degrees). With these scale properties, it is not possible to state meaningful ratios between arbitrary pairs of numbers.

Sources: Kerlinger (1986); Leedy (1997); Tull and Hawkins (1980); Churchill (1979); Zikmund (1991)

Table 1-2: Four Combinations of Measurement and Data

Combination	Measure Example	Rationale
Quantitative Measure Objective Data	Target Location Error	The measure is quantitative because the measurement scale is ratio (i.e., distance measured in feet). The data is collected objectively using a tape measure.
Quantitative Measure Subjective Data	Operator estimate of the probability of survival	The measure is quantitative because the measurement scale is ratio (i.e., probability ranges from 0.0 to 1.0). The data is collected subjectively by the expert responding to questions.
Qualitative Measure Objective Data	Color of munitions	The measure is qualitative because the measurement scale is nominal (i.e., blue, red, or green). The data is collected objectively through direct observation of the color or measuring the wavelength of light.
Qualitative Measure Subjective Data	Operator rating of the display	The measure is qualitative because the measurement scale is ordinal (i.e., Completely Useless, Somewhat Useless, Somewhat Useful, and Completely Useful). The data is collected subjectively by operators responding to questionnaire items.

1.3 Designing Good Questions

Regardless of whether the survey is administered by an interviewer or self-administered, good questions and proper administration are essential to collecting meaningful data. Good questions are unmistakably clear, precise, and unambiguous and ensure the recorded responses align with what the analyst is trying to measure. Questions are specifically worded to avoid creating different interpretations of what is being asked. Differences in answers should be due to differences among respondents rather than from different interpretations of the question's wording. If respondents do not have the same understanding of what a question is asking, then error is likely to result. Good questions are both reliable (i.e., provide consistent responses in comparable situations) and valid (i.e., answers correspond to what they are intended to measure).

There are two general types of questions that should be considered in the survey: closed and open. It is important to note that the data that can be collected is intrinsically linked to whether the question is open or closed. Closed questions provide a list of acceptable responses to the respondent. For closed questions, the analyst must develop a measurement scale to record responses. Measurement scales, usually referred to as scales, enable the analyst to measure various aspects of interest in the study. With closed questions, respondents can answer questions more reliably since the responses are given and analysts can more reliably interpret the meaning of the answers (Fowler, 1993, p. 82).

In contrast, open questions do not provide a list of acceptable responses to the respondents. One benefit of using open questions is that the responses tend to describe more closely the real views of the

respondents and can elicit unanticipated responses. Open questions are appropriate for particular situations;

- When the list of possible responses is long, making it impractical to present to the respondents,
- When all reasonable responses cannot be anticipated by the interviewer or analyst.

Table 1-3 shows closed and open versions of the same question. With the closed version of the question, the responses are limited to the four that are shown. The respondent's focus is framed by providing the list of acceptable responses. With the open version of the question, the respondent may not only provide responses that are similar to those provided in the closed version of the question, but also other responses, perhaps unexpected, regarding the functions of the system.

Table 1-3: Example of Closed and Open Versions of the Same Question

Closed Version		
1. Which of the following functions can the system perform?		
Detect Threat	Yes	No
Identify Threat	Yes	No
Track Threat	Yes	No
Target Threat	Yes	No
Open Version		
2. What functions can the system perform?		

The wording of a question is critically important to ensuring the respondents' interpretations of the question are the same. The following are some recommendations for designing questions:

Use Complete Questions

Table 1-4 shows examples of both incomplete and complete versions of the same question. With incomplete questions, there is a risk that the respondents will interpret the meaning of the question differently. In question 3, age can be interpreted as being one's age as of this current time or age on one's last birthday. In question 4, one can interpret the reason to be anything and not just a medical problem or reason. The complete versions of the questions eliminate these potential interpretation problems.

Table 1-4: Examples of Incomplete and Complete Questions

Incomplete Question	*Complete Question*
3. Age?	3. What was your age on your last birthday?
4. Reason last saw doctor?	4. What was the medical problem or reason for which you most recently went to the doctor?
Source: Fowler (1993, p. 71)	

Use Specific Wording

Table 1-5 shows examples of non-specific and specific versions of the same question. In the non-specific version of question 5, the particular gun control legislation in question is not clear (i.e., banning the legal sale of certain types of guns, limiting the number and types of guns that can be owned, limiting people with certain backgrounds from owning or possessing guns). The specific version of the question fully describes the gun control legislation to ensure the respondents have a consistent interpretation.

In non-specific question 6, there are multiple appropriate responses to this question (e.g., When I was in the Army, When I was a child, When I was 4 years old, In 1985), requiring the respondent to guess what response is required. In survey research, the question must elicit comparable answers that can be organized and analyzed. The question must specify the focus of the answer. In the specific version of the question, it is explicit that the respondent's age is the required answer.

The nonspecific version of question 7 is a "why" question that can be answered in many different ways. For instance, the respondent may remark about the capability, cost, or risk associated with Alternative C as the main factor in selecting it. The specific version of the question explicitly asks what system capabilities were considered in the selection of the alternative. In this version of the question, the respondent is prompted to address the capabilities of the system rather than the cost and risks.

There are some ambiguous words that should be avoided such as often, occasionally, usually, regularly, and frequently. These words have different meanings depending on the respondent. For example, one respondent may think regularly means three times a week, while another may think it means every day.

Table 1-5: Examples of Non-Specific and Specific Questions

Non-Specific Question	*Specific Question*
5. Do you favor or oppose gun control legislation?	5. One proposal for the control of guns is that no person who ever had been convicted of a violent crime would be allowed to purchase or own a pistol, rifle, or shotgun. Would you oppose or support legislation like that?
6. When did you have the measles?	6. How old were you when you had the measles?
7. Why did you select Alternative C?	7. What capabilities of Alternative C led you to select it?
Source: Fowler (1993, pgs. 75, 78, 79)	

Maintain Relevance

Questions in an interview or questionnaire should have a purpose that is relevant to the study. There is a natural temptation to add questions, especially "nice to know" types of questions (O'Brien and Charlton 1996, p. 89). Questions that are not relevant to the study will impose an unnecessary burden on the respondents and data analysts and should be avoided.

Check the Assumption Underlying the Question

Though a question may be clear and straightforward, the assumption underlying the question may be flawed (Leedy 1997). In the question 8 example shown in Table 1-6, the underlying assumption is that a person smokes about the same number of cigarettes each day. For some smokers, the number of cigarettes smoked will vary depending on the circumstances. For example, an impulsive smoker may smoke more cigarettes in a pressure-packed workday than at home or on holiday. Checking the assumptions underlying the questions will help ensure the questions elicit the answers the analyst is seeking. A survey pretest (discussed in Section 1.7) can help identify questions that have flawed underlying assumptions.

If a question has a flawed underlying assumption, it may be possible to revise the question to eliminate the problem. In the example shown in Table 1-6, adding a preceding question (8a) that asks whether the smoker is a consistent smoker would ensure that only consistent smokers answer question 8b.

Table 1-6: Example of a Flawed Assumption Underlying a Question

Flawed Assumption Question				
8. How many cigarettes do you smoke each day? (Check one of the following)				
1 – 5	6 – 10	11 – 15	16 – 25	More than 25
Possible Solution				
8a. Are your daily smoking habits reasonably consistent; that is, do you smoke about the same number of cigarettes each day?				
Yes		No (If "No", skip the following question)		
8b. How many cigarettes do you smoke each day? (Check one of the following)				
1 – 5	6 – 10	11 – 15	16 – 25	More than 25
Source: Leedy (1997, p. 192)				

Avoid Difficult Vocabulary

The interviewer should avoid using jargon, acronyms, or overly technical terms. Choosing words that communicate thoughts fully, clearly, and accurately through plain discourse not only helps enhance readability and comprehensibility, but also avoids ambiguity.

Avoid Exceptionally Lengthy Questions

As a general rule, the longer it takes to say something, the weaker the communication. Question length usually increases complexity and diminishes clarity, resulting in greater opportunities for misunderstanding. Single sentence questions are usually best; however, the meaning of a question or item should not be sacrificed for brevity (DeVellis 1991). Table 1-7 shows lengthy and brief versions of the same question.

Table 1-7: Example of Lengthy and Brief Versions of the Same Question

Lengthy					
What is your level of agreement with the following statement: 9. It is fair to say that one of the things I seem to have a problem with much of the time is getting my point across to other people.					
Completely Disagree	Substantially Disagree	Slightly Disagree	Slightly Agree	Substantially Agree	Completely Agree
Brief					
What is your level of agreement with the following statement: 10. I often have difficulty making a point.					
Completely Disagree	Substantially Disagree	Slightly Disagree	Slightly Agree	Substantially Agree	Completely Agree
Source: DeVellis (1991, p. 57)					

Avoid Negative and Positive Questions

Agreement bias is the tendency to agree with a question or survey item irrespective of its content (DeVellis 1991).[4] This is particularly a concern when questions are positively or negatively worded. Table 1-8 shows examples of negative and positive questions. In the negative version of question 11, respondents would tend to agree that the system has no limitations. In addition, the use of "no" in the question can be missed by the respondents. The neutral version of question not only eliminates the potential agreement bias, but is more easily understood. Similarly, in the positive version of question 12, respondents would tend to agree that the system is adequate. The neutral version of question 12 eliminates the potential agreement bias.

Table 1-8: Examples of Negative, Positive, and Neutral Questions

Negative	*Neutral*
11. Do you agree that the system has no capability limitations?	11. Rate the capability limitations of the system
Positive	*Neutral*
12. Do you agree that the system is adequate?	12. Rate the adequacy of the system

[4] Agreement bias is also known as acquiescence or affirmation bias.

Avoid Double-Barreled Questions

Double-barreled questions pose two or more questions simultaneously. Table 1-9 shows an example of a double-barreled question. If the responsiveness and reliability of the system differ, how should the respondent answer? In this example, the problem is rectified by using two separate questions, one for responsiveness and one for reliability.

Table 1-9: An Example of a Double-Barreled Question

Double-Barreled	*Separate*
13. Rate the responsiveness and reliability of the system	13a. Rate the responsiveness of the system 13b. Rate the reliability of the system
Source: O'Brien and Charlton 1996, p. 88	

Avoid Leading and Loaded Questions

Leading questions presuppose some event or state and can bias the responses for a question (O'Brien and Charlton 1996). The leading question example in Table 1-10 presupposes that the system is unresponsive. By removing the "lack of" from the question as shown in the neutral version, the leading aspect is eliminated.

Similar to leading questions, loaded questions have emotional or sensitive content that can have carry-over effects on the entire interview or questionnaire.[5] In the loaded question example in Table 1-10, the respondent may get the impression that he or she lacks ability to perform his or her duties. In the neutral version of the question, there is no sensitive content (i.e., "your lack of ability") and the focus is directed to the system rather than the respondent.

Table 1-10: An Example of Leading and Loaded Questions

Leading	*Neutral*
14. Rate the lack of responsiveness of the system	14. Rate the responsiveness of the system
Loaded	*Neutral*
15. Rate your lack of ability with respect to the duties you carried out	15. Rate your ability to perform your duties using the system
Source: O'Brien and Charlton 1996, p. 89	

1.4 Question Sequence

Opening questions that are interesting, simple to understand, and easy to answer can help establish cooperation and maintain involvement in answering questions (Zikmund 1991, p. 419). These types of questions not only help build curiosity and confidence in the respondent, but also rapport between the respondent and interviewer. Though simple and easy to answer, demographic-type questions should

[5] Carry-over effects are produced when early questions influence or bias a respondent's answers to later questions (O-Brien and Charlton, 1996).

not be used as opening questions. Demographic-type questions tend to elicit more personal or sensitive information and should be asked later or at the end of the interview or questionnaire when a rapport between the interviewer and respondent has been established (Zikmund 1991).

The sequencing of questions and answers can create an order bias that can distort survey results (Zikmund 1991). For example, specific questions asked earlier in an interview or questionnaire can influence responses to more general questions asked later. To mitigate this bias, the analyst can use the funnel technique which entails asking more general questions to first understand the respondent's frame of reference before asking more specific questions.

Table 1-11 shows an example of how the position of questions can bias the results. In this example, the researchers found that the responses to the pollution problems were almost identical. The "air pollution from automobile exhausts" question biased the responses to the other air pollution questions. This is an example of the anchoring effect which occurs when the first concept measured becomes the comparison point from which subsequent responses are made. To mitigate the anchoring effect, the analyst can randomize the order of the questions. In Table 1-11, for example, the order of the problem statements would vary from respondent to respondent.

Another related problem the analyst must contend with is the tendency of respondents to select the first answer listed (Zikmund 1991). Randomizing the order of responses can help mitigate this bias.

Table 1-11: Example of Previous Question Bias

16. Circle the number that best expresses your feelings about the severity of each environmental problem:					
Problem:	Not a Problem				Very Severe Problem
Air pollution from automobile exhausts	1	2	3	4	5
Air pollution from open burning	1	2	3	4	5
Air pollution from industrial smoke	1	2	3	4	5
Air pollution from foul smoke	1	2	3	4	5
Source: Zikmund (1991, p. 421)					

1.5 Scale Considerations

While the question prompts the response, the scale determines the form of the response. The thought and deliberation that goes into crafting good questions applies as well to selecting the appropriate scales to use. A scale must align with the wording used in the question and the intent of the measure. A question asking if something is useful, for example, should have a scale that measures usefulness in some way. If the data will be used to assess a measure (e.g., MOE, MOS, or MOP), the scale should be linked to the measure to ensure the data can be used to make an assessment as to whether the measure criteria are met or not.

When determining what response alternatives to use, the analyst should consider the following:

- Response alternatives should retain the same directional order for all questions to avoid response errors (i.e., low to high, or high to low), unless there is a belief that the order will make a difference in the responses selected.
- Balanced scales such as the Likert scale (discussed in Section 1.5.2) have an equal number of positive and negative response alternatives and tend to produce distributions that are more nearly normal (O'Brien and Charlton, 1996, p. 84).
- Although greater discriminability can be obtained by adding more response alternatives, more than seven response alternatives increases response variability and lowers overall reliability (O'Brien and Charlton, 1996, p. 87).

The response alternatives (also known as scale descriptors) used in scales are chosen for consistency, discriminability, and comprehensibility. Response alternatives with these attributes can help avoid nonresponse and response bias. Appendix E shows examples of five, six, and seven point scales with these attributes.

1.5.1 Nominal and Ordinal Scales

Closed questions typically have nominal or ordinal based scales to measure a response. As shown in Table 1-12, the nominal scale uses categories (e.g., yes or no) that have no rank order relationship. In contrast, the ordinal scale uses a rank order relationship. Responses to questions are normally treated as nominal or ordinal data based on the scale used. It is important that the analyst know the type of data and the appropriate statistics that can be used (see Section 1.2 (Types of Measurement and Data) for more information).

Table 1-12: Nominal and Ordinal Based Scales

Nominal Scale	
The system will enable the nuclear enterprise to accomplish its mission.	
Yes	No

Ordinal Scale					
Rate the adequacy of the system in supporting the nuclear enterprise mission.					
Totally Inadequate	Very Inadequate	Somewhat Inadequate	Somewhat Adequate	Very Adequate	Totally Adequate

1.5.2 Likert Scale

The Likert scale, developed by sociologist Rensis Likert to measure psychological attitudes in a scientific way, is an ordinal based scale that is commonly used in studies to measure the level of agreement or disagreement. As shown in Table 1-13, the scale is bivalent (two-directional) and balanced (i.e., equal number of positive and negative response alternatives) with a neutral middle. The scale has verbal

labels that connote evenly spaced graduations of the response alternatives. Five-point response alternative scales are often used, though seven and nine point scales can be used as well.

Table 1-13: Example of a Likert Scale

A two-level maintenance concept can be used to maintain this system.				
Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree

1.5.3 Neutral Midpoint

Another aspect the interviewer must consider is whether to use a neutral midpoint in a scale such as the example shown in Table 1-13 above. Scales without a neutral midpoint force respondents to select a response that departs from true neutrality which can occasionally result in nonresponse. The drawbacks of forcing respondents to make a choice must be carefully weighed against the benefits of obtaining non-neutral responses.

1.5.4 Not Applicable or Don't Know as a Selection

The selection of respondents to participate in a survey not only requires careful planning and preparation, but also a thorough understanding of the respondents' qualifications to answer the survey questions. Despite the best efforts of the interviewer, there may be cases when respondents are asked questions concerning things about which they do not know. One approach to deal with such a possibility is to include "Not Applicable" or "Don't Know" as a selection separate from the response alternatives. A "Not Applicable" or "Don't Know" selection indicates the respondent did not have adequate knowledge or experience on which to base an answer. These selections are for administrative purposes and are separate from the responses of interest.

There are two other reasons for including "Not Applicable" or "Don't Know" as a selection. First, it will allow the analyst to better ascertain whether item nonresponse was intentional or unintentional since the likelihood that respondents will intentionally not answer the question is low. Second, a "Not Applicable" or "Don't Know" selection helps prevent respondents who do not have adequate knowledge or experience on which to base an answer from arbitrarily selecting a response. Table 1-14 shows an example of an item with "Don't Know" included as a selection.

When qualified respondents do select a "Not Applicable" or "Don't Know" response, the analyst must investigate the reason for the selection to ensure it was not accidental and that the respondent actually had no basis for an opinion. In the report, the analyst should clearly identify these occurrences, describe the reasons for them, provide justification for their removal from the sample population, and clearly document the actual sample size. Because neutral responses like "Neither Agree nor Disagree" or "Borderline" are not particularly informative, the analyst should also investigate and document the reasons for these response selections. Finally, the analyst should always investigate and document any response anomalies such as outliers (in either direction) and bimodal distributions. With graphical and numerical representations of the data, the analyst can discover patterns and anomalies and identify

potential causes.[6] The information helps facilitate understanding and interpretation of the data, enabling the analyst to describe and present results in more meaningful ways.

Table 1-14: Item with "Don't Know" Included as a Selection

How important is the airborne radiation survey reconnaissance map for the ground planning mission?					
Not Important at All	Not So Important	Neutral	Fairly Important	Very Important	Don't Know
Comments:					

1.6 Designing the Survey

Whether the survey is interview-administered or self-administered, there are several characteristics of a well-designed questionnaire. Though the questions may be worded properly and have the appropriate format and response scale, the questions should flow from the most general to the most specific, or from the most frequent/common events to the rare or unusual (O'Brien and Charlton 1996, p. 89). This is done to minimize carry-over effects, instances where the responses to earlier questions bias responses later in the questionnaire.

A well-designed questionnaire is brief and to-the-point (O'Brien and Charlton 1996). With overly long questionnaires, there is a risk of respondents choosing neutral or extreme responses out of expediency or fatigue. If the questionnaire takes more than 15 minutes, O'Brien and Charlton (1996) recommend that the analyst consider dividing the questions into two or more questionnaires that can be administered at different times during the study.

The remaining discussion in this section describes guidelines for designing the interview-administered survey and self-administered survey.

1.6.1 Interview-Administered Survey

In the interview-administered survey, an interviewer reads the questionnaire instructions and questions to the respondent and records the responses. Fowler (1993, p. 99) offers the following guidelines for designing a survey that will not only make the tasks of the interviewer and respondent as easy as possible, but also help ensure the information elicited from the respondents is what the interviewer is seeking:

[6] Examples of graphical representations include bar charts, histograms, and box and whisker plots. Numerical representations include measures of central tendency such as the mean, median, and mode and measures of variability such as range, variance, and standard deviation.

- Adopt a convention that differentiates between the words that the interviewer reads to respondents and words that are instructions. A common convention is to use uppercase letters for the instructions and lowercase letters for words to be read aloud.
- Establish a clear convention for handling instructions to skip questions that do not apply to a particular respondent. The instructions should be keyed to a particular response and tell the interviewer where to go to ask the next questions. Visual cues such as boxes and arrows are probably the most self-explanatory. Such visual cues, however, require some formatting at the typing and printing stages of the questionnaire. Another option is to use clearly written "skip" instructions. Whatever approach is used, it is worthwhile to be completely consistent so that the interviewer does not have to spend time thinking about which questions to ask.
- Put optional wording in parentheses. Conventions such as "(his/her)" or "(husband/wife)" are easy for interviewers to handle smoothly if they are alerted by the parentheses. A similar convention uses all caps (e.g., NAME) when the interviewer must supply a word that is not provided in the question itself.
- Check to make sure that all the words that an interviewer has to say are, in fact, written. This includes not only the phrasing of questions but transitions, instructions to questions, needed definitions, and explanations.

An additional interview technique to consider is the pitcher/catcher routine whereby one interviewer pitches a question while a second interviewer records the response. This technique allows the pitcher to focus on the response and generate follow-up questions if necessary. Additionally, the chances of missing a response or parts of a response are greatly reduced by having one interviewer devoted to recording the response.

1.6.2 Self-Administered Survey

In the self-administered survey, the respondent reads the questions in the questionnaire and provides responses, hence the name "self-administered". Similar to interview-administered surveys, self-administered surveys should have a questionnaire that is easy to use, although the format and design of the self-administered surveys is typically more important. Self-administered surveys should have a cover sheet that includes the title of the questionnaire, purpose of the questionnaire, information regarding the use of the data, tracking information (e.g., date, time, location), and if needed, clear and brief instructions and an assurance of confidentiality.

For self-administered surveys, Fowler (1993, p. 100) offers the following recommendations:

- A self-administered questionnaire should be self-explanatory. Written instructions should not be necessary because they will not be read consistently.
- Self-administered questionnaires should be restricted to closed questions. Checking a box or circling a number should be the only tasks required. When respondents are asked to answer in

their own words, the answers are usually incomplete, vague, and difficult to code, and therefore they are of only limited value as measurements.[7]

- The question forms in a self-administered questionnaire should be few in number. The more the questionnaire can be set up so that the respondent has the same kinds of tasks and questions to answer, the less likely it is that respondents will become confused; also, the easier the task will be for respondents.
- A questionnaire should be typed and laid out in a way that is clear and uncluttered. Photo-reduction (or other strategies for putting many questions on a page) actually reduces the response rate compared to when the same number of questions are spaced attractively over more pages.
- Question skip patterns (i.e.., how respondents skip questions that do not apply to them) should be kept to a minimum. If some respondents must skip some questions, arrows and boxes that communicate the skips without verbal instructions are best.

1.7 Pretesting the Survey

There are several approaches to pretest a survey. For interview-administered surveys, one approach entails identifying and selecting qualified individuals (i.e., subject matter experts or others familiar with the area of interest in the study) to serve as respondents for pre-testing the questions. The interviewer conducts trial interviews to observe whether the respondents have difficulty answering the questions and determine whether the responses align with expectations. For example, a respondent's request for additional information may indicate a problem with the question sequence or the question itself. The interviewer can test alternative wordings, order and number of questions, question format (e.g., open versus closed), and sequence to determine which approach is best suited for the respondents.

A similar approach can be used to pretest self-administered questionnaires. The interviewer first identifies and selects qualified individuals to serve as respondents. Once the respondents have been selected, the interviewer directs the respondents to fill out the questionnaire. Finally, the interviewer leads a discussion to determine whether there are any problems regarding the clarity and completeness of the instructions, expected responses (i.e., whether the responses align with expectations), and question wording, format, and order.

As part of pre-testing the survey, the interviewer determines whether the length of the interview or questionnaire is appropriate. The criteria for determining survey length include cost, the effect on response rate, and limits of the respondent's ability and willingness to answer questions (Fowler 1993, p. 103).

1.8 Sampling

There are two main types of sampling: nonprobability and probability. This section describes each type and provides recommendations for determining sample size.

[7] Some experts in survey design recommend providing a comment section with each closed question to provide an opportunity for the respondent to explain the rationale behind a response.

1.8.1 Nonprobability Sampling

Unlike probability sampling, nonprobability sampling is a procedure for selecting the sample elements that is not based on chance (i.e., the selection is not random). A special form of nonprobability sampling commonly used in operational capability requirements studies is judgment or purposive sampling.[8]

In purposive sampling, the analyst uses judgment in selecting the sample elements (i.e., those who the analyst intends to interview or will fill out a questionnaire). The analyst purposively selects the sample elements because they will serve a specific purpose and have some appropriate characteristic that is required. There are a variety of reasons the analyst can use to select the sample elements. For example, there may be a limited number of individuals who have the expertise in the area being studied, or the interest of the study is on a small group or in a specific field. Alternatively, the analyst may be attempting to draw a representative sample of a population, but uses judgment rather than a probabilistic approach in the selection procedure.

Though there are many advantages of nonprobability sampling, all nonprobability sampling techniques are subject to selection error (Tull and Hawkins 1980, p. 389). The level of selection error is dependent on the degree of expertise of the individual making the selection decision. As the sample size increases, judgment becomes less trustworthy compared to random sampling procedures used in probability sampling. In addition, given that the elements are not selected probabilistically, there are no appropriate statistical techniques for measuring sampling error (i.e., the degree to which a parameter (e.g., median, mean) of a sample represents the parameter of a population) and projecting the data beyond the sample (Zikmund 1991, p. 462).

1.8.2 Determining the Sample Size for Nonprobability Sampling

There is little theoretical basis for determining the sample size for nonprobability sampling. With the sample elements being selected based on judgment, some scholars think sample size may also be determined on the basis of judgment, though no heuristics or criteria have been prescribed (Zikmund 1991, p. 510). Others recommend using sample sizes that are similar to sample sizes used in previous studies.

Another approach known as the Bayesian method for determining sample size entails choosing a sample size based on the expected net gain of sampling. The expected net gain of sampling is the difference between the expected value of information and the cost of sampling. The objective is to choose the sample size that produces the greatest positive difference. Though it may be possible to calculate the cost of sampling, calculating the expected value of information is much more difficult and may require resources beyond those of a typical capability requirements study.

---

[8] Another type of nonprobability sampling is convenience sampling (selecting individuals that are most conveniently available). Convenience sampling is also known as accidental or haphazard sampling. Subcategories of purposive sampling include expert sampling (selecting individuals with known or demonstrable experience and expertise in some area) and quota sampling (selecting individuals non-randomly according to some fixed quota where the quota may be proportional or non-proportional).

A more practical approach is to calculate the sample size as if it were a probability sample. The procedure described in Section 1.8.4 (Determining the Sample Size for Probability Sampling) can be used to determine the sample size for nonprobability sampling.

1.8.3 Probability Sampling

In some situations, the analyst may not only want the sample elements to represent a population of interest, but also a selection procedure based on chance (i.e., the selection is random). For example, an analyst may want to first gather information about a particular population (e.g., F-16 aircraft pilots), and then select individuals from the population to participate in an expert elicitation panel to gather more specific information. In this example, probability sampling could be used to gather the information about the F-16 pilot population. In probability sampling, every individual in the population of interest has a chance to be selected as a sample element. The procedure for selecting the sample elements is random and is the basis for all probability sampling techniques.

There are several different types of probability sampling techniques. Simple random samplings, simple and proportional stratified sampling, and cluster sampling, are some examples of probability sampling techniques. As shown in Table 1-15, selecting the appropriate sampling technique requires an understanding of the population.

Simple random sampling is the least sophisticated of all the probability sampling techniques. The sample is derived by means of a simple randomization process. Random number tables or computer-generated numbers are used for the sample selection. More information about simple random sampling and the other types of probability sampling techniques can be found in the references listed in Appendix C.

1.8.4 Determining the Sample Size for Probability Sampling

With probability sampling, whatever statistics are calculated from the sample (e.g., mean, median, percentage), a population value is being estimated. The relationship between sample size and sampling error (i.e., the degree to which a parameter of a sample represents the parameter of a population) can be generally expressed as: the smaller the sample, the larger the error; and the larger the sample, the smaller the error (Kerlinger 1986). As a general rule, the analyst should use as large a sample as possible.

There are statistical equations that can be used to estimate sample size. The choice of which equation to use is dependent on whether the data being collected is attribute or variable data. Table 1-16 shows examples of survey questions designed to elicit attribute data and variable data. Attribute data is associated with the proportion or percentage of the population that has the same characteristic, feels a certain way, or views something the same way. As shown in the table, the attribute question is designed to elicit a maintainer's opinion about how easy or difficult it would be to maintain the new electronic warfare system. In this example, the analyst is interested in knowing what percentage of the respondents think the system would be somewhat easy or extremely easy to maintain.

Table 1-15: Sampling Techniques Appropriate for Each Population Type

Population Characteristic	Example	Appropriate Sampling Technique
Population is generally a homogeneous group.	A quantity of flower seeds of a particular variety from which random samples are selected for testing as to their germination quality.	Simple Random Sampling
Population consists of definite strata, each of which is distinctly different, but the units within the stratum are as homogeneous as possible. Each stratum is essentially equal in size.	A particular town whose total population consists of three types (strata) of citizens: European-American type; African-American type; and Mexican-Indian type.	Simple Stratified Sampling
Population consists of definite strata with differing characteristics, and each stratum has a proportionate ratio in terms of numbers of members to every other strata.	A community in which the total population consists of individuals whose religious affiliations are found to be Catholic, 25%; Protestant, 50%, Jewish, 15%; nonaffiliated, 10%.	Proportional Stratified Sampling
Population consists of clusters whose cluster characteristics are similar yet whose unit characteristics are as heterogeneous as possible.	A survey of the nation's 20 leading air terminals by soliciting reactions from travelers who use them. (All air terminals are similar in atmosphere, purpose, design, and so forth, yet the passengers who use them differ widely in individual characteristics: age, gender, national origin, philosophies, beliefs, and socioeconomic status)	Cluster Sampling
Source: Leedy 1997, p. 214		

Table 1-16: Examples of Questions Designed to Elicit Attribute Data and Variable Data

Attribute Data Question				
How easy or difficult would it be to maintain this new electronic warfare system?				
Extremely Difficult	Somewhat Difficult	Borderline	Somewhat Easy	Extremely Easy
Variable Data Question				
How long will it take in hours to perform routine scheduled maintenance on the new electronic warfare system?				
Write the number of hours here:				

In contrast, variable data is associated with an amount of something expressed as a number or value. The variable question example in the table is designed to elicit a value from the respondents. The analyst is interested in determining the average number of hours from the estimates provided by the respondents.

In capability requirements studies, the analyst is typically dealing with attribute data rather than variable data such as means and standard errors of a mean. For attribute data, the analyst would use equations designed to estimate sample size involving proportions. The remaining discussion in this section describes how the analyst would use these equations to determine the sample. Appendix F provides the equations for estimating sample size involving means. Though different equations are used, the approach is similar to what is described in this section.

According to Emory (1985, p. 287), the single most important factor in determining the size of a sample needed for estimating a parameter of a population is the magnitude of the population variance (commonly expressed as the standard deviation), not the size of the population. The greater the dispersion or variance of the population, the larger the sample must be to provide an estimate with a given level of accuracy and precision. For example, if a population held the same view about something (i.e., no variance), then only a sample of one would be needed to know the view held by the population. If there are many possible views, then a larger sample is needed to collect them and estimate their frequencies.

For the statistical equations that are used to estimate sample size, an estimate of the population variance is needed (Zikmund 1991, p. 506).[9] Information from prior studies or a pilot study is commonly used to estimate the variance. In capability requirements studies, it is likely that such information is not available or time and resource constraints make it impossible to collect the information through a pilot study. With attribute data, it is still possible to calculate a sample size.

The variance is measured in terms of *p*, the proportion of the population that has the given attribute. With *n* = size of the sample, the variance of the proportion is as follows:

$$\sigma_p^2 = \frac{p(1-p)}{n}$$

The standard error of the proportion is as follows:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

[9] See Zikmund (1991), Emory (1985), Tull and Hawkins (1980), and O'Brien and Charlton (1996) for variations of statistical equations for estimating the sample size for variable data (e.g., mean and standard errors of the mean) and attribute data.

Solving for *n*, the equation becomes:

$$n = \frac{p(1-p)}{\sigma_p^2}$$

Before the sample size can be estimated statistically, the analyst must determine (1) the magnitude of acceptable sampling error (precision); and (2) the confidence level. These specifications typically involve judgment and are based on how the data will be used. Questions such as "How much error in the estimate is acceptable?" and "How confident do you want to be that the error really isn't greater than that?" must be answered (Tull and Hawkins 1980, p. 413).

As noted earlier, sampling error is a measure of precision and is the degree to which a parameter (e.g., median, mean) of a sample represents the parameter of a population. Sampling error is also referred to as the confidence interval. The sampling error or confidence interval is a specified range of numbers within which a population parameter should lie. For example, a sampling error of plus or minus 10% means the sample parameter is within 10% of the population parameter.

Closely associated with the sampling error (or confidence interval), the confidence level is distinct and calculated differently. Expressed as a percentage or decimal value, the confidence level tells how confident an analyst can be about being correct (Zikmund 1991). For example, a confidence level of .95 means there is a .95 probability that the true population parameter is correctly estimated. Stated another way, there is a .05 probability that the true population parameter is incorrectly estimated.

Using judgment, the analyst selects the appropriate items or questions to be used for sample size calculations (Zikmund 1991, p. 510). In most studies, the desired degree of precision may be different across the items or questions selected by the analyst. The convention is to select the item or question that will produce the largest sample size and use it to determine the ultimate sample size. The analyst should consider the cost of data collection and exercise judgment regarding the importance of such information.

To facilitate an understanding of determining a sample size involving proportions, an example for simple random sampling[10] is described below. Appendix F provides the equation used to calculate the sample size involving means.

EXAMPLE

> Using a simple random sampling technique, the analyst wants to determine the size of the sample to estimate the true proportion in the population within plus or minus 10 percent. In addition, the analyst wants to be 95 percent confident that the population proportion is within plus or minus 10 percent of the sample proportion. Given that there is no estimate of the population variance, the analyst follows the established convention of using the largest

[10] Though the principles of determining the sample size for simple random sampling are applicable to all methods of probability sampling, see Tull and Hawkins (1980, p. 422 and Appendix B) for more information about determining sample size for non-simple random sampling techniques.

variance. Recall that with attribute data, that variance is measured in terms of *p*, the proportion of the population that has the given attribute. In the standard error of a proportion equation shown below, the standard error is largest when $p = 0.5$. For example, if $p = 0.5$, then the product of p (0.5) and $1 - p$ (0.5) is 0.25. If p or $1 - p$ is greater than 0.5, then the product will always be smaller than 0.25. If $p = 0.6$, for instance, then $p(1 - p) = 0.6 \times 0.4 = 0.24$. With this feature of the equation, the convention is to assume the largest variance by selecting $p = 0.5$ when an estimate of the variance is not known. Note that the cost of *p* being unknown is an increase in the sample size.

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Given that *n* is what the analyst is trying to calculate in the standard error equation above, the next step is to calculate the standard error (σ_p). As noted earlier, the analyst wants to determine the size of the sample to estimate the true proportion in the population within plus or minus 10 percent. This means the confidence interval will be ± 0.1. Recall that the analyst wants only a .05 probability that the true population proportion is incorrectly estimated.[11] Hence, the confidence interval must encompass a dispersion of ± 1.96 standard errors of the proportion. The 1.96 value is the z-value that corresponds to the 0.025 proportion of the area of the normal curve. One standard error is determined by dividing the confidence interval by the standard errors (0.1/1.96 = 0.051).

With the standard error calculated, the analyst can now calculate the sample size *n* as shown below:

$$n = \frac{p(1-p)}{\sigma_p^2}$$

$$n = \frac{.5 \times .5}{.051^2}$$

$$n = 96$$

Rather than calculating the sample size for each survey, the analyst can use a table such as the one shown in Table 1-17. There are several aspects of the table that are worth noting. First, the sampling error or confidence interval is greatest when the proportion is 50/50 and decreases as the proportion approaches 5/95. Second, as the sample size increases, the sampling error decreases. It is important to note that sampling error steadily decreases as the sample size increases from 35 to 200, after which point the reductions in sampling error become much less. Adding sample elements to a sample reduces sampling error more significantly for small samples than large samples. For instance, increasing the

[11] In a normal curve, this means that an incorrect estimate must fall in each tail of the normal curve where the proportion of the area is 0.025. With two tails in the normal curve, each with a proportion of area of 0.025, the total proportion of area becomes 0.05.

sample size from 35 to 75 decreases the confidence interval by 2 (from 7 to 5) for the 5/95 proportion, whereas increasing the sample from 300 to 350 would not yield a noticeable decrease in sampling error.

Fowler (1993, p.34) points out that it is unusual for an analyst to be able to specify a desired level of precision in more than a general way. It is only the exception, rather than the common situation, when a specific level of precision can be specified in advance. This means it would be unusual for the analyst to use the equations described above to determine the sample size. Given this, Fowler (1993) highlights the value of the table in determining the sample size. He recommends that the analyst start at the low end of the sample size continuum shown in the table rather than the high end. With the level of precision associated with each sample size, the analyst asks the question "Are 35 observations adequate?" If not, the analyst asks the same question for a larger sample size and so on. As noted earlier, sampling error steadily decreases as the sample size increases from 35 to 200, after which point the reductions in sampling error become much less.

Table 1-17: Confidence Ranges for Variability Attributes to Sampling

	Proportion of Sample with Characteristic[1,2]				
Sample Size	5/95	10/90	20/80	30/70	50/50
35	±7%	±10%	±14%	±15%	±17%
50	±6%	±8%	±11%	±13%	±14%
75	±5%	±7%	±9%	±11%	±12%
100	±4%	±6%	±8%	±9%	±10%
200	±3%	±4%	±6%	±6%	±7%
300	±3%	±3%	±5%	±5%	±6%
500	±2%	±3%	±4%	±4%	±4%
1,000	±1%	±2%	±3%	±3%	±3%
1,500	±1%	±2%	±2%	±2%	±2%

[1] Chances are 95 in 100 that the real population figure lies in the range defined by the ± number indicated in the table, given the proportion of sample reporting the characteristic and the number of sample cases on which the proportion is based.
[2] This table describes the variability attributable to sampling. Errors resulting from nonresponse or reporting errors are not reflected in this table. In addition, a simple random sampling is assumed. Estimates may be subject to more variability than this table indicates because of the sample design or the influence of interviewers on the answers they obtained; stratification might reduce the sampling errors below those indicated here.

Source: Fowler (1993, p. 31)

1.9 Survey Question Validity and Reliability

The approach to designing questions and surveys described in this handbook is in accordance with widely accepted practices in survey research. Practices such as the wording and length of questions, pretesting, and survey design and administration are intended to establish question reliability (provide consistent measures in comparable situations) and validity (answers correspond to what they are intended to measure).

There are statistical techniques to measure question reliability and validity, but using such techniques would likely exceed the resources of a typical capability requirements study and are beyond the scope of

this handbook. For more information regarding statistical techniques for assessing validity and reliability, please see DeVellis (1991), Scale Development: Theory and Applications.

1.10 Administering the Survey

Given that the approach used to administer the survey can influence the quality of the data collected, it is important that the interviewer use proper procedures in administering the survey. This section describes some guidelines for administering surveys.

1.10.1 Interview-Administered Surveys

According to Fowler (1993), the interviewer has three primary roles to play in the collection of interview-administered survey data:

- Locate and enlist the cooperation of selected respondents,
- Train and motivate respondents to do a good job of being a respondent,
- Ask questions, record answers, and probe incomplete answers to ensure that the answers meet the question objectives.

To enlist cooperation, the interviewer must first locate potential respondents and make contact. Though some individuals may readily agree to being interviewed, others may be reluctant to participate. Interviewers who are good at enlisting cooperation typically exhibit a confident assertiveness and personable style (Fowler 1993). They present the study in a way that the individual will want to participate. The tone and content of the conversation leaves no doubt that an interview will be conducted. The interviewer engages the individual personally by tailoring the interaction to be responsive to the individual's needs, concerns, and situation. It is also important that the interviewer is available to conduct an interview when a respondent wants to be interviewed.

Most respondents have little understanding of what they are expected to do and how they should perform (Fowler 1993). A critical part of the interviewer's job is to train and motivate the respondent's behavior. Studies have shown how the encouragement provided by the interviewer affects what respondents do and how well they report. For example, an interviewer who reads a question quickly encourages the respondent to answer the question quickly, but not necessarily accurately. In contrast, an interviewer who reads a question slowly and deliberately encourages the respondent to take the time to thoughtfully and accurately answer the question.

To ensure that differences in responses are attributed to differences in the views and experiences of the respondents rather than to differences in how the survey is administered, there are aspects of interviewer behavior that must be standardized. Fowler (1993, pgs. 107-108) offers the following:

- <u>Presenting the Study</u>. Respondents should have a common understanding of the purposes of the study. This sense of purpose may have a bearing on the way they answer questions. Assumptions about such things as confidentiality, the voluntary nature of a project, and who will use the results also can have some effect on answers. A good interviewer will give all

respondents a similar orientation to the study so that the context of the interview is as consistent as possible.

- Asking the Questions. Survey questions are supposed to be asked exactly the way they are written, with no variation or wording changes. Even small changes in the way questions are worded have been shown, in some instances, to have significant effects on the way questions are answered.
- Probing. If a respondent does not answer a question fully, the interviewer must ask some kind of follow-up question to elicit a more complete answer; this is called probing. Interviewers are supposed to probe incomplete answers in nondirective ways—ways that do not push the respondent and increase the likelihood of any one answer over another. A short list of standard probes, including repeating the question, asking "Anything else?", "Tell me more", and "How do you mean that?" will handle most situations if the survey is designed well.
- Recording the Answers. The recording of answers should be standardized so that no interviewer-induced variation occurs at that stage. When an open-ended question is asked, interviewers are expected to record answers verbatim; that is, exactly in the words that the respondent uses, without paraphrasing, summarizing, or leaving anything out. In closed-response questions, when respondents are given a choice of answers, interviewers are required only to record an answer when the respondent actually chooses one. There is potential for inconsistency if interviewers code respondent words into categories that the respondent did not choose.
- Interpersonal Relations. The interpersonal aspects of an interview are to be managed in a standardized way. Inevitably, an interviewer brings some obvious demographic characteristics into an interview, such as gender, age, and education. By emphasizing the professional aspects of the interaction and focusing on the task, the personal side of the relationship can be minimized. Interviewers generally are instructed not to tell stories about themselves or to express any views or opinions related to the subject matter of the interview. Interviewers are not to communicate any judgments on answers that respondents give. In short, behaviors that communicate the personal, idiosyncratic characteristics of the interviewer are to be avoided because they will vary across interviewers if more than one interviewer is used. To behave as a professional, and not as a friend, helps to standardize the relationship across interviewers and respondents. There is no evidence that having a friendly interpersonal style improves the accuracy of reporting; it probably tends to have a negative effect on accuracy.

1.10.2 Self-Administered Surveys

Proper administration of self-administered surveys includes providing the questionnaire to respondents, ensuring the respondents understand all questionnaire items, conducting a quality control check of the responses, and actively investigating reasons for certain responses (generally those that are ambiguous or unexpected). For capability requirements studies, questionnaires should never be simply handed to respondents who are then asked to "fill them out and return them whenever you can." A much more effective approach is to schedule a specific time and place for the respondents to gather and complete the questionnaires. The interviewer remains with the group to field questions and clarify items that may

be confusing or are being misinterpreted. As questionnaires are returned, the interviewer should carefully examine them to ensure:

- A response alternative (or "Not Applicable" or "Don't Know" selection) has been selected for all items,
- The respondent viewed the scale directions correctly (an indication that the scale direction may have been viewed incorrectly is when responses from a respondent are mostly opposite to those of other respondents), and
- Responses to open-ended questions and any other comments can be read and understood.

If there are issues with any questionnaire responses, the interviewer should review and resolve them with the respondent immediately. It is never good practice to put off addressing questionnaire problems to a later date as memories fade and people may become unreachable as they move on to other activities.

1.11 Data Analysis

As previously shown in Table 1-1, there are various descriptive and inferential statistics that can be used depending on the type of data (i.e., nominal and ordinal). To use these statistics, the analyst must first convert the data into a numerical form. To do this, the analyst assigns numerical values to the descriptors used in the measurement scale. Table 1-18 shows an example of a numerical value assignment for a five-point Likert scale. Using the scale and associated numerical values, the analyst scores the numerical value that corresponds to the respondent's answer. For example, an answer of "Strongly Agree" would be scored as a "5". An answer of "Somewhat Agree" would be scored as a "4", and so on.

Table 1-18: Assigning Numerical Values to the Scale Descriptors

Question:	A two-level maintenance concept can be used to maintain this system.				
Scale Descriptor:	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Assigned Value:	1	2	3	4	5

There are several graphical representations such as the bar chart, histogram, and box and whisker plot that the analyst can use that will help facilitate an understanding of the data.[12] Figures 1-1 and 1-2 show examples of the bar chart and box and whisker plot. In these examples, subject matter experts were asked to rate how easy or difficult it will be to maintain, transport, and deploy a specific system. The scale used with each question was a five-point Likert scale ranging from "Extremely Difficult" to "Extremely Easy". As shown in Figure 1-1, the bar chart is useful for showing the response frequencies

[12] There is a difference between the bar chart (or bar graph) and the histogram. The bar chart is used to present categorical-type data such as that associated with nominal and ordinal levels of measurement. The histogram is used to present continuous data such as that associated with interval and ratio levels of measurement.

of each rating and overall pattern of the data. These representations can reveal whether there are outlier responses and the nature of the distribution (e.g., bimodal). In this example, the single "Extremely Easy" response to the transportability question is a potential outlier.

The box and whisker plot is one way to illustrate both variability and central tendency of the data in a single format (see Figure 1.2). The top and bottom of each box represents the third quartile (Q3) and first quartile (Q1) respectively. The horizontal red line inside the box represents the median value. The whiskers can represent several possible values such the minimum and maximum values, the 2nd and 98th percentiles, and multiples of the interquartile range values.[13] In this example, the whiskers represent the minimum and maximum values of the data. As shown in the figure, the variability of the deployability responses (as indicated by the size of the box) is much less compared to the variability of the maintainability and transportability responses.

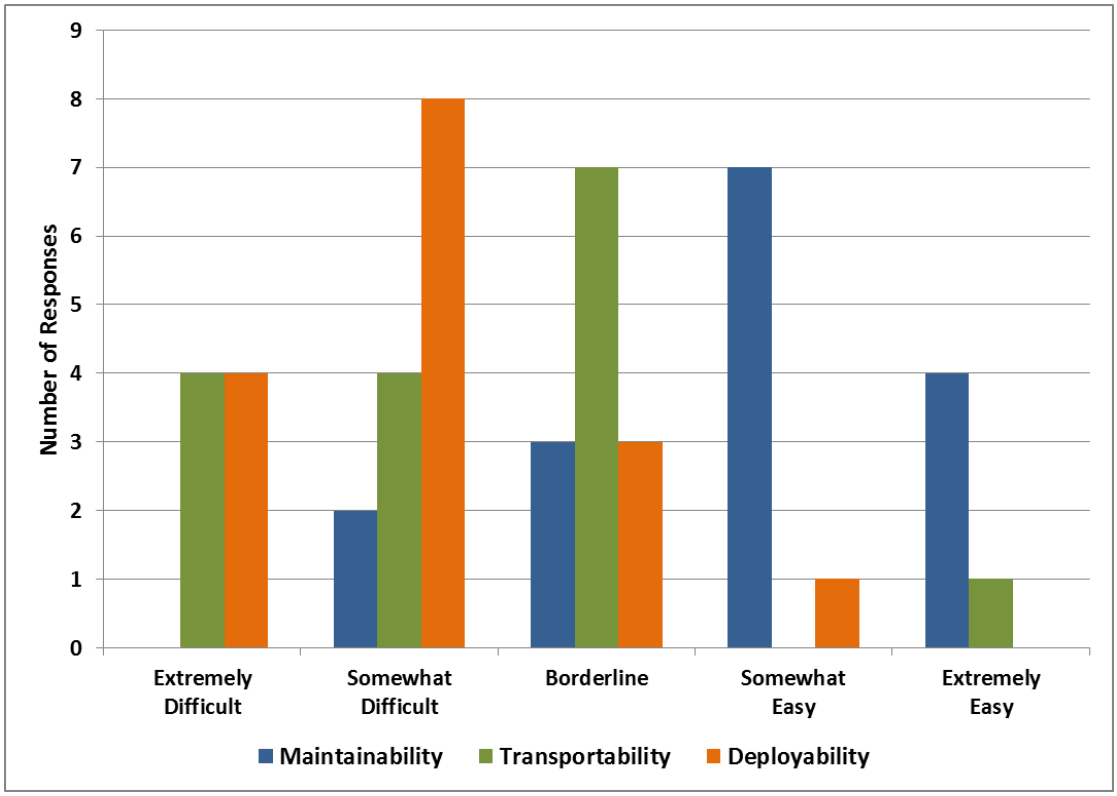

Figure 1-1: Bar Chart Example

[13] The interquartile range (IQR) is the difference between Q3 and Q1. The interquartile range is often used to find outliers in data. Outliers are observations that fall below Q1 - 1.5(IQR) or above Q3 + 1.5(IQR). In a box and whisker plot, the highest and lowest values are drawn as the bar of the whiskers, and the outliers as individual points.

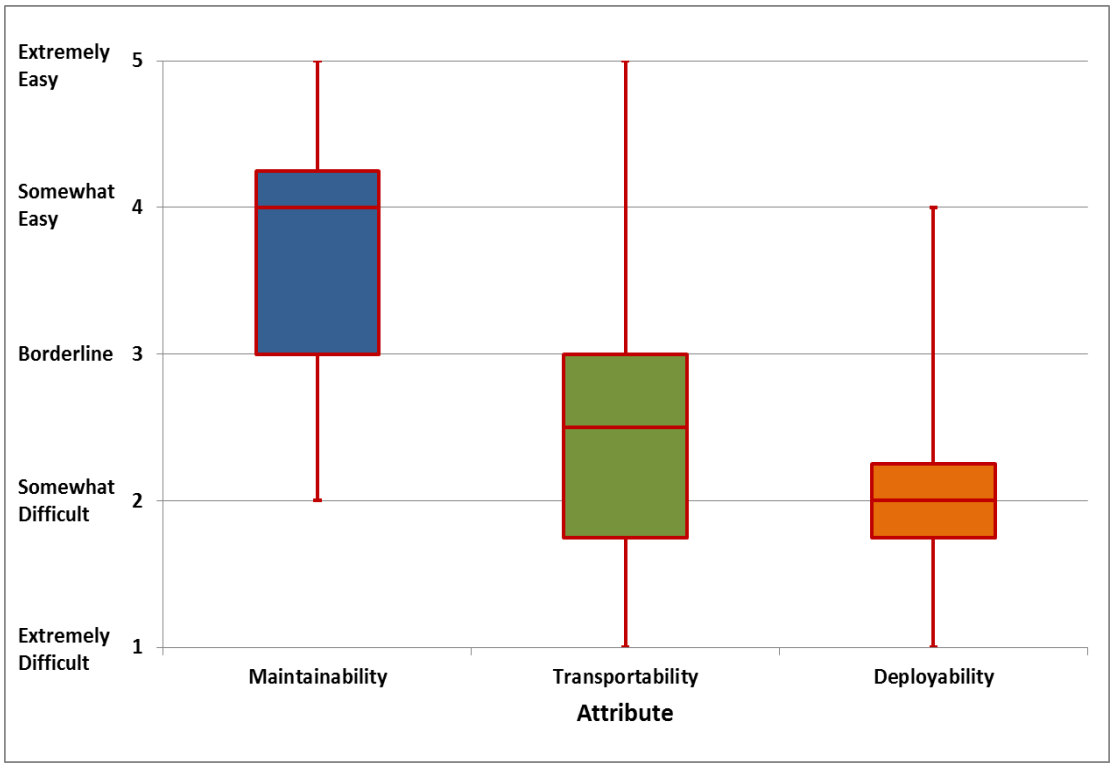

Figure 1-2: Box and Whisker Plot Example

1.11.1 Central Tendency and Dispersion Measure Criteria

The criteria used for a measure may require calculating the central tendency, dispersion, or both. Central tendency statistics such as the median or mode are used to identify the direction (positive or negative) and magnitude of the ratings. Dispersion statistics such as the variance or ratings corresponding to a percentile capture the level of agreement in the ratings.

Table 1-19 shows an example of a measure where the criteria requires the computation of a central tendency statistic and a dispersion statistic. The measure criteria states that the median rating must be greater than or equal to "Somewhat Agree" and the 80th percentile must be "Somewhat Agree" or better. In this example, the central tendency and dispersion criteria are met, indicating that not only a majority of the responses are favorable (86% in this example), but there is substantial agreement among the respondents.

Table 1-19: Examples of Central Tendency and Dispersion Measure Criteria

Measure Description

Task	Attribute	Measure	Metric	Criteria (Threshold)	Data
Maintain and Sustain System	Maintainability	Logistician rating of maintainability (MOS)	Median	≥ Somewhat Agree	Logistician responses to questionnaire item with 5-point Likert scale (see below)
			Percentile	≥ 80th Percentile must be Somewhat Agree or Better	

Questionnaire Item and Responses

Question:	A two-level maintenance concept can be used to maintain this system.				
Scale Descriptor:	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Assigned Value:	1	2	3	4	5
Number of Responses:	0	0	2	8	4
Results					
Median: 4 (Somewhat Agree)					
Percentile: 86th Percentile is Somewhat Agree or Strongly Agree					

1.11.2 Ordinal Data Analysis

One pitfall to avoid is the inappropriate analysis of ordinal data. Recall that ordinal data is one of the types of data that is usually collected through surveys.[14] In the example shown in Table 1-20, data for the measure was collected through a self-administered survey and exhibits the properties of ordinal data (i.e., data are assigned numbers such that the order of the numbers reflects an order relation based on the attribute). The analyst, however, incorrectly selected the mean as the metric. Despite six of the seven respondents agreeing that a two-level maintenance concept can be used for the system, the measure is rated "did not meet criteria" because 3.8 is not greater than or equal to the threshold value of 4 (Somewhat Agree).

[14] The most common usage of ordinal scales is in obtaining attitude or preference measurements. The analyst is typically interested in whether something has more or less of a characteristic or attribute. Unlike an interval scale, the ordinal scale represents a ranking of an attribute because the intervals are not exactly equal for all the points on the scale. This limits what statistics can be used, making it impossible to determine how much more or less of an attribute something has. The analyst can only determine whether something has more or less of an attribute or characteristic, but not to what extent.

Table 1-20: Inappropriate Analysis of Ordinal Data Example

Measure Description

Task	Attribute	Measure	Metric	Criteria (Threshold)	Data
Maintain and Sustain System	Maintainability	Logistician rating of maintainability (MOS)	Mean	≥ Somewhat Agree	Logistician responses to questionnaire item with 5-point Likert scale (see below)

Questionnaire Item and Responses

Question:	A two-level maintenance concept can be used to maintain this system.				
Scale Descriptor:	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Assigned Value:	1	2	3	4	5
Number of Responses:	0	0	1	6	0
Results					
Mean Response: 3.8 ((1 x 3) + ((6 x 4))/7					

To rectify this situation, the analyst selects the mode as the metric (Table 1-21). Using the mode as the metric, the analyst rates the measure as "met criteria" because the mode value of 4 corresponds to "Somewhat Agree." This example highlights how using inappropriate metrics can sometimes produce different results.

Table 1-21: Appropriate Analysis of Ordinal Data Example

Measure Description

Task	Attribute	Measure	Metric	Criteria (Threshold)	Data
Maintain and Sustain System	Maintainability	Logistician rating of maintainability (MOS)	Mode	≥ Somewhat Agree	Logistician responses to questionnaire item with 5-point Likert scale (see below)

Questionnaire Item and Responses

Question:	A two-level maintenance concept can be used to maintain this system.				
Scale Descriptor:	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Assigned Value:	1	2	3	4	5
Number of Responses:	0	0	1	6	0
Results					
Mode Response: 4 (Somewhat Agree)					

1.11.3 Analyzing Data with "Not Applicable" or "Don't Know" Selections

When analyzing and presenting data for an item, the analyst should exclude "Not Applicable" and "Don't Know" selections from the other responses. Including "Not Applicable" or "Don't Know" selections with the other responses can produce misleading results and lead to incorrect interpretations of the data. Figures 1-3 and 1-4 show how the response distributions change significantly when the "Don't Know" selections are inappropriately included as part of the data set.

In Figure 1-3 (without "Don't Know" selections), the majority of respondents (8 out of 10) who are qualified to answer the question think the reconnaissance map is fairly or very important for the ground planning mission. With the addition of the "Don' Know" selections in Figure 1-4, it appears there is no longer a majority of respondents who think the reconnaissance map is important for the ground planning mission. The visual image presented in the figure draws attention to the high number of respondents who did not have adequate knowledge or experience on which to base an answer. In this case, interpreting the results becomes more difficult and may lead to faulty conclusions. For instance, one may incorrectly conclude from the figure that less than half of the respondents think the airborne radiation survey reconnaissance map is important for the ground planning mission.

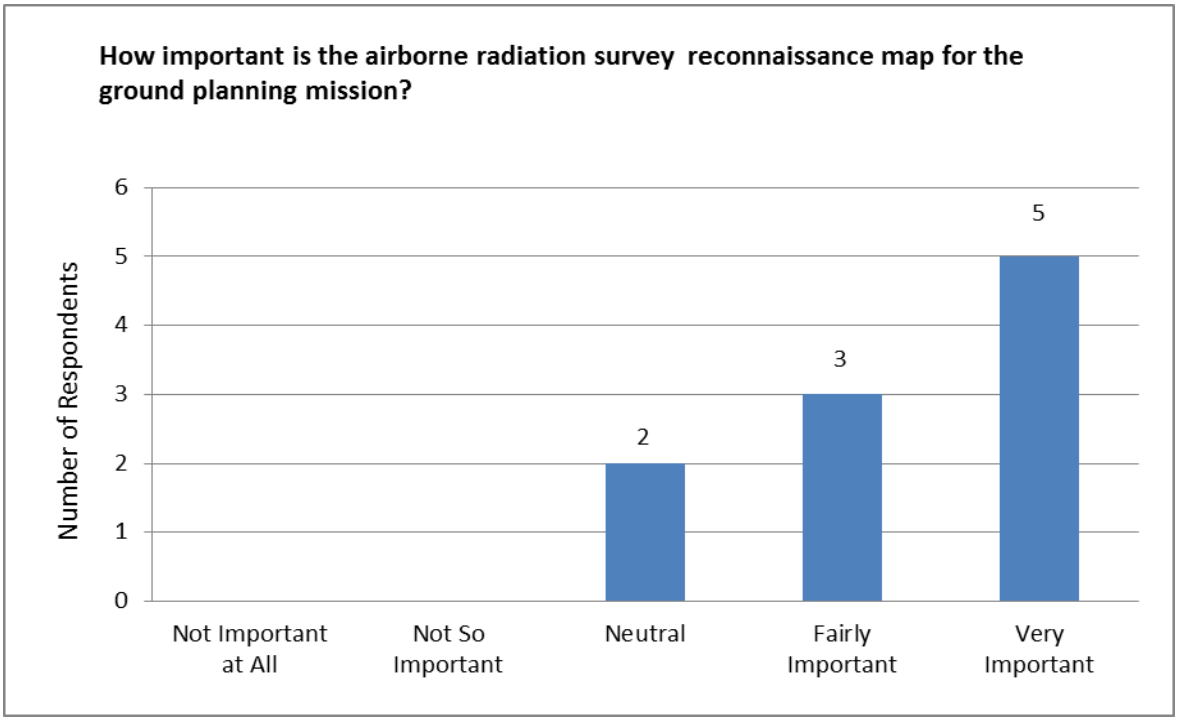

Figure 1-3: Bar Chart without "Don't Know" Selections

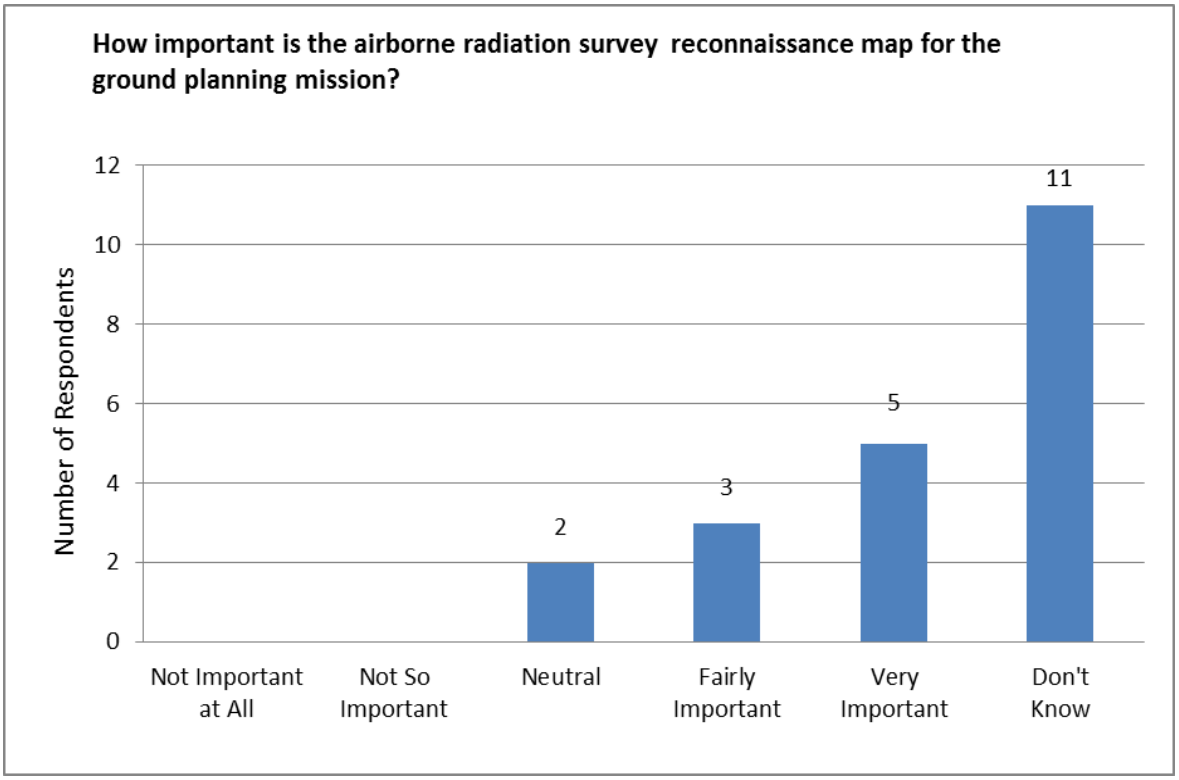

Figure 1-4: Bar Chart with "Don't Know" Selections

1.11.4 Sensitivity Analysis

Though there are numerous ways to analyze data and produce results, the sensitivity analysis is distinct in that it can yield new and meaningful insights that can profoundly influence the interpretation of the results. The main purpose of the sensitivity analysis is to highlight performance stability or robustness of the system, solution, or concept being assessed in the study. This is accomplished by varying performance parameters, operational conditions, scenarios, or assumptions to determine the resulting changes in performance. The sensitivity analysis not only enhances the credibility of the analysis, but also facilitates the identification of key performance tradeoffs. The results of this analysis often serve as the basis for study conclusions, recommendations, and decisions.

There are a number of ways data collected in surveys can be used in a sensitivity analysis. For example, if experts are asked to estimate a range of possible reliability values of a system, then the analyst could vary the reliability values within the specified range to determine whether the overall results significantly change. As another example, the analyst may want to identify which variables and associated point estimates collected through a survey significantly drive results. For these variables, the analyst could determine whether any reasonable changes in the point estimates produce very different results. The analyst could use such information to inform decision-makers of the potential risks associated with any decisions based on the results of the analysis.

2 Expert Elicitation

After an introduction of expert elicitation and judgment, this chapter presents an approach to conducting expert elicitation in operational capability requirements studies. It provides insights into the selection of experts, development of questions, and design and conduct of the elicitation process.

2.1 Introduction

As a special form of survey research, expert elicitation relies on surveys to collect information from subject matter experts. It is a structured method of gathering expert judgment and answering questions concerning issues or problems of interest in a study. The Delphi method, developed by the RAND Corporation in the 1950s, was one of the first recognized expert elicitation methods.[15] Over the years, many other elicitation methods have been developed and used by various organizations in both the private and public sectors. There are numerous examples of its use by federal agencies to include the United States Army Corps of Engineers, Nuclear Regulatory Commission, National Aeronautics and Space Administration, Department of Transportation, Department of Energy, Department of Agriculture, and the Environmental Protection Agency.

Since expert judgment is affected by the approach used to gather it, a specially designed process is required that includes procedures for developing questions, conducting the elicitation, and handling biases that may arise. The process is designed to facilitate thinking and encourage experts to state their true opinions. Through the elicitation process, experts derive judgments from the available body of evidence ranging from direct empirical data to theory. Although the process is formal and structured, it can differ in terms of the degree of interaction between experts, level of detail in information elicited, number of meetings, type of communication mode, and degree of structure in the elicitation process.

Expert elicitation is different from probability sampling methods since respondents are not considered to be representative of a population (Chan et al, 2010). Instead, respondents are viewed as representing a large body of knowledge. Expert elicitation seeks to reflect the range of credible opinions regarding a specific question or problem, so the foremost concern is the quality and diversity of the participating experts.

2.2 What is an Expert?

Meyer and Booker (2001) define an expert as “a person who has background in the subject area and is recognized by his or her peers or those conducting the study as qualified to answer questions.” It is natural to think of experts as professionals such as scientists, physicians, and engineers, but any person with sufficient knowledge of the subject matter can be considered an expert for the purpose of the study. Although an individual’s knowledge is important, other factors such as personality, experience,

[15] The Delphi method entails a group of experts who anonymously reply to questionnaires and subsequently receive feedback in the form of a statistical representation of the "group response," after which the process repeats itself. The goal is to reduce the range of responses and arrive at something closer to expert consensus.

and expertise in organizing and using his or her knowledge are critical to the success of the elicitation (O'Hagan et al, 2006, p. 27). Achieving a balanced and broad spectrum of viewpoints may require eliciting judgments from individuals with various backgrounds and degrees of expertise.

2.3 Expert Judgment

There is a variety of terms used to describe expert judgment such as expert opinion, subject matter expert assessment, subject matter expert analysis, subjective judgment, and expert knowledge. Whatever it is called, expert judgment is the data given by an expert in response to a question and represents an expression of opinion based on knowledge and experience. Judgment is shaped by the expert's state of knowledge at the time of the response to the question, and because experts have different experiences and knowledge, their judgments can differ and change over time as new information is learned.

Expert judgment is commonly expressed in quantitative terms, although it is possible to obtain expert judgment in a variety of other non-numeric or qualitative forms. Some examples of information elicited from experts are shown in Table 2-1.

Table 2-1: Examples of Information Elicited from Experts

Quantitative	*Qualitative*
Probability of an occurrence of an event	Impact of a change
Probability of failure of a system	Risks and consequence of a decision
Estimates of ranges of uncertainty	Variables, assumptions, and data used in an analysis
Likelihood of a causal relationship	Elements needed for decision making
Allocation of funding	Failure causes, potential failures, potential solutions
Rating of the performance of a model	Methods to optimize performance

2.4 An Expert Elicitation Approach for Capability Requirements Studies

It is necessary to follow a formal and structured process to ensure the information elicited from experts is suitable for analysis. The following describes a seven-step approach to conducting expert elicitation in the CBA, AoA, or other type of capability requirements study. It provides guidelines for the selection and preparation of experts, development of questions, design and conduct of the elicitation process, and analysis and reporting of data.

2.4.1 Step 1: Identify the Need for Expert Elicitation

In conducting a capability requirements study, the analyst typically deals with many unknowns associated with new and complex concepts. Choosing the appropriate research methods to collect and analyze data is a primary concern. Study objectives, data accessibility, time and resource constraints, and available tools and techniques are some important factors that the analyst must consider when determining which research methods to use.

In most studies, there is a need to gather subjective information for various reasons. Expert elicitation is one of several techniques that can be used to gather such information. Some examples of using expert elicitation for this purpose include the following:

- Determining the priority or rank of capability gaps,
- Establishing study ground rules, constraints, and assumptions,
- Determining the scope, purpose, and objectives of a study,
- Identifying criteria (threshold and objective values) of measures,
- Determining the screening criteria for alternatives,
- Identifying and rating risks, probability of risk, and risk consequences,
- Identifying and prioritizing threats and targets,
- Selecting scenarios to be used in a study,
- Identifying potential materiel and non-materiel solutions,
- Determining the most viable alternatives.

In some cases, expert elicitation can be used when other research methods (e.g., modeling and simulation, parametric analysis) are not feasible or data is insufficient, unattainable, or too costly or impractical to collect. In these cases, the data and information elicited from experts is typically used in capability requirements studies to compute a metric for a measure of effectiveness, suitability, or performance.[16] For example, experts may be asked to provide estimates of a system's probability of kill performance against specific targets under certain operational conditions. The mean, median, or other appropriate metric for the measure is used to compute a value from the estimates. The analyst uses the value to determine whether the measure criteria are met or not. The data and information elicited from experts can address various attributes of interest in the study such as those shown in Table 2-2.

Table 2-2: Examples of Attributes

Accessibility	Accuracy	Adaptability	Adequacy
Availability	Capacity	Compatibility	Deployability
Flexibility	Interoperability	Latency	Maintainability
Mobility	Networkability	Persistence	Precision
Reliability	Scalability	Security	Simplicity
Sufficiency	Survivability	Sustainability	Time
Timeliness	Transportability	Utility	Vulnerability

2.4.2 Step 2: Develop the Questions

Expert elicitation relies on surveys to collect data of some aspect for analysis. Expert judgment is primarily elicited through face-to-face interviews. The choice of whether to use personal interviews (i.e., interview one expert at a time) or group interviews (i.e., interview experts in a group) will depend

[16] For more information on metrics and measures, see *The Measures Handbook*, Office of Aerospace Studies, 6 August 2014.

on various factors such as time constraints and the availability of experts. Whatever method is chosen, using good questions is an essential part of the survey process.

Crafting good questions requires careful forethought and a sound approach. The process entails drafting a set of initial questions and using a small group of experts to design the final questions. Subject matter experts who are not among the experts in the panel can assist in developing the questions as well as any assumptions, definitions, or other supporting information associated with the questions. Expert insights gleaned during the question development process will help ensure the questions are eliciting the information of interest in the study. Capability requirements studies typically require many different types of experts (e.g., aircraft operators, logisticians, intelligence experts), so it is critical to have the right ones participating at the right time.

Pre-testing the questions with several other experts can help refine the questions and identify problems such as unclear wording or misreading that must be addressed prior to using the questions in the elicitation. Feedback from experts will be helpful in determining how specific questions should be worded, order and number of questions, and question format. See Section 1.3 (Designing Good Questions) for additional information regarding question development.

2.4.3 Step 3: Select the Experts

Selection criteria define the set of individuals that have a chance of being selected to participate as expert panel members in the study. With the selection of experts being a critical step in the process, it is important to establish selection criteria through careful deliberation. Given that the expert panel selection is not random, there is a risk of researcher bias when the researcher makes selections based on inappropriate criteria. Selection error present in an expert panel depends on the degree of expertise of the person making the selection decision. It is advantageous to consider a range of possible criteria by drawing from the expertise of the study director, study team members, study advisory group, and other appropriate individuals and groups.

A “good” expert has technical knowledge, experience, and intuition as well as an ability to integrate information and draw conclusions. Criteria such as level of training, type of skill, and years of experience can be used to ensure the panel consists of experts with the proper knowledge and expertise. Ultimately, selection criteria will depend on the objectives of the study. Table 2-3 provides some examples of criteria that can be used to identify experts for participation in a study.

Like other studies, the number of experts used in a capability requirements study will be driven mostly by resources and time available to conduct the study as well as the number and availability of individuals who have the expertise in the area being studied. A major challenge for the analyst is not only identifying experts, but also getting them to participate. In some cases, there may be no experts. For example, if the area of interest involves break-through technologies or existing systems will be used differently from how they were designed, then there may well be no individuals with the appropriate experience and expertise to be deemed “experts”. In these cases, the analyst may still find the “most knowledgeable” individuals and elicit information from them, but the uncertainty of the information may be problematic for the study. Other research methods may be required in these cases.

In other cases, experts may exist, but they may not be available to participate in a study because their time is in very high demand. Other individuals may be recommended, but they may not have the requisite experience and expertise in the area of interest. Similar to the situation described above, these cases may require the use of other research methods.

Table 2-3: Examples of Selection Criteria

Criteria	Description
Knowledge of Area of Interest	Understanding of the area of interest, reputation as a technical authority, awards received, membership in organizations or groups in the area of interest.
Background and Experience	Years of experience, level and diversity of experience, type and number of past positions held.
Education and Training	Specialized training, type of advanced academic degree(s), special certification(s) and qualifications.
Published Work	Number and quality of publications in the area of interest.
Personal Skills	Interpersonal skills, communication skills, flexibility, impartiality, ability to generalize and simplify.
Economic or Personal Stake	Lack of economic or personal stake in the potential findings.
Availability and Willingness	Availability and willingness to commit the necessary time and effort to participate in the study, willingness to prepare for discussions and provide opinions.

Although there are no absolute rules regarding the number of experts, large panels increase the likelihood that all possible expert views are represented. While all are knowledgeable of the area of interest, experts have different experiences and perspectives that will shape their responses. Large panels can often produce insights that may not be possible with small panels.

Despite the lack of definitive approaches to determining the appropriate number of experts, a panel of practitioners in expert elicitation recommends at least six experts should be included and that the benefit of including additional experts beyond 12 begins to diminish (Cooke and Probst, 2006, p. 16). Using panels with less than six members will likely reduce the chances of collecting a diversity of information.

2.4.4 Step 4: Prepare the Experts

Once the experts have been identified and selected, the next step entails preparing them for the elicitation by providing relevant information about the study. Experts must have a thorough understanding of the issues before they are ready to answer questions. Issue familiarization is the process used to help the experts understand the issues of interest in the study, purpose of their participation, expectations, study objectives, elicitation process, list of questions, terminology, and key assumptions and definitions. Depending on the objectives of the elicitation, information about the technical aspects of the baseline capabilities, potential solutions, study methodology, and measures (e.g., MOEs, MOSs, and MOPs) may be required as well.

Whether done in a group or individually, it is important to present the same information to ensure a common understanding of the issues. Presentations, briefing books, and other documents should be assembled to provide the relevant information.

2.4.5 Step 5: Conduct the Elicitation

The approaches used to elicit judgments vary widely and will rely to a large degree on the objectives of the study. The amount of time required for the elicitation may range from a few hours to as much as a week depending on the size and complexity of the study. The analyst should consider a number of factors in designing the elicitation:

- Time and resources available for the study,
- Type of information to be elicited,
- Number of experts,
- Amount of time experts will need to provide judgments,
- Degree of interaction among the experts,
- Number and type of questions,
- Format for the answers,
- Mode(s) of communication,
- Type of interview.

Expert judgment is elicited through personal interviews, group interviews, or a combination of both. Personal interviews are usually done in private and in person and allow the interviewer to gather in-depth data from the experts without distraction or influence by other experts. Group interviews are conducted in person through a structured approach that defines when and how experts express and discuss their opinions.

Although personal interviews can be used, convening an in-person group meeting to conduct the elicitation has several advantages in the CBA, AoA, and other capability requirements study. Most importantly, it provides an opportunity to introduce the issue, review the relevant information, and describe the elicitation purpose and process. It can serve as a forum to answer questions, share information, discuss expectations, describe how the results will be used, and gain feedback on any issues that require further clarification or additional information. The major drawback to group elicitation is the undesirable effects of dominant or vocal participants, something that is avoided by eliciting experts individually through personal interviews (Cooke and Probst, 2006, p. 16).

In group elicitations, there are greater demands of time and effort on the interviewer to structure and facilitate the discussions and interactions amongst the experts. The interviewer is responsible for ensuring the integrity of the elicitation process and its implementation by initiating and maintaining effective discussions. Ayyub (2001, p. 18) recommends using a facilitator or moderator to help create an environment that ensures equity in presenting views and a successful elicitation of opinions and information from each expert.

In capability requirements studies, gaining insights into the underlying reasoning or rationale of an expert's response may be as important as the response itself. There are several techniques described by Meyer and Booker (2001) that can be used to interview experts and learn the rationale for a response:

- The verbal report involves instructing the expert to think aloud when answering a question and resembles someone talking to oneself. The technique can be time consuming since it is used on one expert at a time. It is important to note that not all experts are capable of verbalizing all their thoughts for various reasons (e.g., too difficult to articulate, thoughts are automatic or unconscious).
- The verbal probe entails phrasing questions in a way to minimize influencing the expert's thinking. The technique is a quick means of obtaining information and is suitable for both personal and group interviews. Some examples include repeating the question, asking "Anything else?", "Tell me more", and "How do you mean that?"
- The ethnographic technique involves transposing the expert's words into questions. Because the questions are based on the expert's own words, it is a non-biasing form of questioning. The technique can be time consuming and is not suitable for group interviews.

In structuring the elicitation, it is important to understand and anticipate bias that may occur. Bias is a skewing that arises from our personal perceptions and understanding. There are various forms of bias and methods for dealing with them. Table 2-4 provides a brief description of seven common forms of bias and when they are likely to occur.

Several steps can be taken in designing the elicitation process to help mitigate anticipated bias. For example, to reduce social pressure from the data gatherer, the interviewer can use the verbal report, verbal probe, and/or ethnographic phrasing of questions instead of direct questions that may lead the experts. If complicated response forms such as probability and uncertainty estimates are being elicited, prepare the experts for the elicitation by conducting a training session that describes the fundamental principles of the response form. The training will help eliminate the potential of confusion and underestimation and give the experts an opportunity to rehearse providing responses to sample questions in the appropriate form. Finally, as part of the preparation for the elicitation, it is important to make the experts aware of the forms of bias and why they happen. Although bias cannot be completely eliminated, experts will not be able to control their own tendencies toward bias without first having a good understanding of it.

While much can be done to design the elicitation to help mitigate bias, the interviewer must still be alert to the occurrences of bias during the elicitation process and make the appropriate adjustments to counter it. For example, if there are inconsistencies in responses, the interviewer should ask the experts to reconsider their responses. If fatigue is a factor, the interviewer can shorten the elicitation sessions or schedule breaks to help preclude potential inconsistencies in responses. In group situations, the interviewer should suspect group think is occurring when no one in the group voices a difference of opinion or the group consistently defers to one expert or a subgroup of experts in the group.

Table 2-4: Common Forms of Bias

Bias	Description
Social Pressure (Data Gatherer)	Individuals consciously or unconsciously alter the descriptions of their thoughts to gain acceptance and to be seen in the most positive light possible. Data gatherers can intentionally or unintentionally influence the individual through body language, facial expression, intonation, and word choice. More pronounced in cases when the interviewer uses leading questions.
Social Pressure (Group Think)	Social pressure from others in a group induces individuals to alter their responses or silently acquiesce to what they believe will be acceptable to the group. More pronounced when individuals in a group desire to remain as members, are satisfied with the group, and view the group as cohesive.
Wishful Thinking	Individuals' hopes influence their judgment—what individuals think should happen will influence what they think will happen. More pronounced when individuals do not have to explain their reasoning and when individuals are personally involved or would gain from their answers.
Inconsistency	Individuals are inconsistent in solving problems—their current thoughts or answers may contradict those expressed earlier. More pronounced when: 1. Elicitation sessions are long and individuals forget instructions, definitions, and assumptions, 2. Complicated response forms such as probability distributions and percentiles are causing confusion, 3. Experts are asked to consider too many things and become confused and inconsistent.
Underestimation of Uncertainty	Individuals underestimate the uncertainty in the answers they provide. More pronounced when response forms are probabilities and other quantitative estimates.
Anchoring	Individuals receive additional information but do not adjust from their first impression in answering the question. More pronounced when experts have described their positions orally or in writing and fear losing face if they change their response.
Availability	Individuals do not mention more than one or two considerations in giving their responses which can mean the experts are drawing from data that is easier to recall. More pronounced when the expert does not receive any information from others that could help trigger less accessible data when formulating a response.
Source: Adapted from Meyer and Booker, 2001, p. 133	

There are many different approaches to interview experts that would be appropriate in these studies. In group situations, one approach commonly used involves interviewing each expert separately, reviewing the answers in a group, and then providing an opportunity for the experts to revise their responses. Depending on the objectives of the study, the analyst may only be interested in collecting responses to questions, whereas in other cases, the rationale for the response may be required as well. Table 2-5 shows several examples of elicitation methods for group interview situations.

Table 2-5: Examples of Elicitation Methods for Group Interview Situations

Example A
Each expert in a panel is asked individually to provide a response to a question as well as rationale for his or her response that includes identification of issues that significantly influenced the response. After providing responses, the panel of experts is given an opportunity to review the results. During the review, each expert discusses the rationale for his or her response while the other panel members are encouraged to ask questions and contribute information. Following the review, the experts are given an opportunity to revise their responses and provide rationale in light of what was learned during the discussion. With the submission of the revised responses, the question is closed and the elicitation process resumes with the next question.
Example B
Each expert is asked individually to provide an initial response to a question. To avoid social pressure, the individual responses are then displayed anonymously to the panel of experts through an on-screen graphical presentation. The experts are given an opportunity to discuss the results of the presentation. Following the discussion, each expert is asked individually to provide a final response. With the submission of the final response, the question is closed and the elicitation resumes with the next question.
Example C
Questions with associated background information are provided to the panel of experts. To encourage knowledge sharing, the experts are given an opportunity to discuss the questions and information as a group. The interviewer monitors the discussion and responds to any questions from the panel members. If necessary, the interviewer provides additional information to help the panel in understanding the issues. The information may be requested by the panel, or the interviewer, through observation, deems the information is needed to facilitate the discussion. When the panel discussion is complete, each expert is asked individually to provide a response to each question. With the submission of the responses, the questions are closed and the elicitation resumes with the next set of questions.

In personal interview situations, experts are interviewed separately in face-to-face meetings or by telephone. If the response requires clarification or there is a desire to collect the rationale for the response, the analyst can use the verbal report, verbal probe, or ethnocentric technique described earlier to gather the information. For example, an analyst can instruct the experts to explain in detail their thinking process as they respond to the questions (verbal report). The verbal probe and ethnographic technique can be used to clarify responses and/or gain more insights into the rationale for the responses.

The questions used in the elicitation will depend on the objectives of the capability requirements study. Questions can be designed to elicit opinions in a variety of forms such as quantities, uncertainties, relationships, parameters, or events. Table 2-6 shows several examples of information that can be elicited.

Table 2-6: Examples of Information Elicited from Experts

Example A
In determining the probability of a system failure, experts are asked to provide a best estimate as well as a degree of uncertainty. The best estimate is expressed as a percentage, although the decimal or ratio can be used as well. This estimate is viewed as the median value where there is a 50% chance that the "true" value will be higher, and a 50% chance the "true" value will be lower. Next, the experts are asked to estimate an upper bound where there is a strong likelihood (95% chance) that the "true" value will be lower than the estimate, and only 5% chance that the "true" value will be higher. In the analysis, these estimates are used as the 50th and 95th percentile values.
Example B
After reviewing technical information of a system, the experts are asked to rate how easily the system can be configured for transport. Each expert is asked to answer a series of questions with five-point Likert scales ranging from "strongly disagree" to "strongly agree" and provide written rationale for his or her response. In the analysis, the mode statistic is determined for each question and the rationale used by the experts is highlighted in the discussion of the results.
Example C
Experts are given an opportunity to review five models used for predicting performance of a system. Each expert is asked to rate the plausibility of each model using a seven-point scale ranging from "Least Plausible" to "Most Plausible" and provide written rationale for his or her response. In the analysis, the responses are shown graphically along with the mode statistic for each model. The results provide a discussion of the mode statistic and rationale used by the experts in rating the models.

2.4.6 Step 6: Aggregate the Data

In capability requirements studies, there is typically a requirement to report a single value by combining responses. Whether judgments are elicited from experts separately or in a group in some instances, the analyst can mathematically aggregate the responses using simple algorithms such as the mean and median. For example, if experts were asked to provide an estimate of a system's reliability (a probability value), the analyst can use the mean, median, or other simple algorithms to aggregate the estimates. More complex weighted means can be used to give more weight to experts who are viewed as having more expertise, although the prevailing recommendation among practitioners in expert elicitation is to use equal weights since it is a simple and robust method for aggregating expert judgments (O'Hagan, 2006, p. 222; Meyer and Booker, 2001, p. 329). Measurement scales such as the Likert scale produce ordinal data, so it is important to use appropriate statistics such as the mode or median.

If the judgments are elicited from experts in a group, another option is to use a behavioral aggregation that requires a convergence or consensus of opinion among the experts through discussion and interaction. A major risk of this approach is the undue influence of dominant participants.

2.4.7 Step 7: Report the Results

Since there is both potential value and danger of using expert judgment, some guidelines are necessary when reporting results derived from expert judgment. Traditional scientific research does not explicitly accommodate the use of opinions as scientific data. It is important to ensure the distinction between empirical data and expert judgment data is maintained by clearly identifying which analyses are based on empirical data and which are based on expert judgment data. Cooke (1991) recommends that sufficient information should be provided about the data and calculations so that the results can be reproduced by others.

Another important consideration is the generalizability of results. Unlike probability sampling, expert elicitation is unlikely to produce results that are representative of a population since all individuals in the population do not have equal chances of being selected. This means the analyst should not make statistical inferences about a population from the expert judgment data. Expert elicitation does not entail randomly selecting individuals with the intent of making inferences about a population, but rather, individuals are selected based on their knowledge and experience with the intent of drawing conclusions about the existing knowledge base. Lastly, the analyst should provide the names and background information of the experts used in the study in the final report. This will help readers ascertain the credibility of the experts.

2.5 Summary

Expert elicitation can be a useful technique for gathering various types of data for analysis in the CBA, AoA, and other capability requirements study. Expert elicitation is a formal and structured process that entails the selection of experts, conduct of the elicitation, and analysis of data. The approach described in this chapter will help ensure the information elicited from experts is properly collected and suitable for analysis. It provides guidelines for the selection and preparation of experts, development of questions, design and conduct of the elicitation process, and analysis and reporting of data.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A: Acronyms

ANOVA	Analysis of Variance
AoA	Analysis of Alternatives
CBA	Capabilities-Based Assessment
IQR	Interquartile Range
JCIDS	Joint Capabilities Integration and Development System
JS	Joint Staff
MOE	Measure of Effectiveness
MOP	Measure of Performance
MOS	Measure of Suitability
OAS	Office of Aerospace Studies

Appendix B: Information Sources

AFI 10-601, *Operational Capability Requirements Development*, 6 November 2013

AFI 10-604, *Capabilities-Based Planning*, 10 May 2006

AFI 61-101, *Management of Science and Technology*, 14 March 2013

AFOTECMAN 99-101, *Operational Test Processes and Procedures*, 6 August 2010

Analysis of Alternatives Handbook: A Practical Guide to the Analysis of Alternatives, Office of Aerospace Studies, 10 June 2013

Capabilities-Based Assessment (CBA) Handbook: A Practical Guide to the Capabilities-Based Assessment, Office of Aerospace Studies, 10 March 2014

CJCSI 3170.01H, *Joint Capabilities Integration and Development System*, 10 January 2012

Defense Acquisition University (DAU) Glossary, Glossary of Defense Acquisition Acronyms and Terms, 13th Edition, November 2009.

DoDD 5000.01, *The Defense Acquisition System*, 8 December 2008

DoDI 5000.02, *Operation of the Defense Acquisition System*, 8 December 2008

Joint Publication (JP) 1-02, *Department of Defense Dictionary of Military and Associated Terms*, 8 November 2010

JS/J8, *Manual for the Operation of the Joint Capabilities Integration and Development System*, 12 February 2015 (JCIDS Manual)

The Measures Handbook: A Practical Guide for Developing and Analyzing Measures in the Capabilities-Based Assessment, pre-Materiel Development Decision Analysis, and Analysis of Alternatives, Office of Aerospace Studies, 6 August 2014.

Appendix C: References

Ayyub, Bilal M. (2001). *A Practical Guide on Conducting Expert-Opinion Elicitation of Probabilities and Consequences for the Corps Facilities*. Alexandria, VA: Institute for Water Resources Report 01-R-01, United States Army Corps of Engineers.

Chan, Gabriel, Laura D. Anadon, Melissa Chan, Audrey Lee. (2010). Expert Elicitation of Cost, Performance, and R&D Budgets for Coal Power with Carbon Capture and Sequestration. *Energy Procedia*.

Churchill, Jr., Gilbert A. (1979). *Marketing Research: Methodological Foundations, Second Edition*. Hinsdale, Illinois: The Dryden Press.

Cooke, Roger M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science.* New York, NY: Oxford University Press.

Cooke, Roger M. and Katherine N. Probst. (2006). *Highlights of the Expert Judgment Policy Symposium and Technical Workshop*. Washington DC.

DeVellis, Robert F. (1991). *Scale Development: Theory and Applications*. Newbury Park, CA: Sage Publications.

Emory, C. William. (1985). *Business Research Methods, Third Edition*. Homewood, IL: Richard D. Irwin, Inc.

Fink, Arlene. (2003). *The Survey Handbook, Second Edition*. Thousand Oaks, CA: Sage Publications, Inc.

Fowler, Jr., John J. (1993). *Survey Research Methods, Second Edition*. Newbury Park, CA: Sage Publications, Inc.

Kerlinger, Fred N. (1986). *Foundations of Behavioral Research, Third Edition*. Fort Worth, Texas: Harcourt Brace Jovanovich College Publishers.

Leedy, Paul D. (1997). *Practical Research: Planning and Design, Sixth Edition*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Meyer, M. A. and J.M. Booker. (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics, American Statistical Association.

O'Brien, Thomas G. and Samuel G. Charlton (editors). (1996). *Handbook of Human Factors Testing and Evaluation*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

O'Hagan, Anthony, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow. (2001). *Uncertain Judgements: Eliciting Experts Probabilities*. West Sussex, England: John Wiley and Sons, Ltd.

Tull, Donald S. and Del I. Hawkins. (1980). *Marketing Research: Measurement and Method*. New York, New York: Macmillan Publishing Co., Inc.

Zikmund, William G. (1991). *Exploring Marketing Research, Fourth Edition*. Hinsdale, Illinois: The Dryden Press.

Appendix D: Glossary

Attribute – a quality or feature of something. Attributes of tasks (e.g., survivability, persistence, availability, accuracy, etc.) form the basis for identifying and drafting measures. (AFOTECMAN 99-101)

Capability – the ability to complete a task or execute a course of action under specified conditions and level of performance. (JCIDS Manual 2015)

Capability Gap (or Gap) – the inability to meet or exceed a capability requirement, resulting in an associated risk until closed or mitigated. The gap may be the result of no fielded capability, lack of proficiency or sufficiency in a fielded capability solution, or the need to replace a fielded capability solution to prevent a future gap. (JCIDS Manual 2015)

Capability Requirement – a capability which is required to meet an organization's roles, functions, and missions in current or future operations. To the greatest extent possible, capability requirements are described in relation to tasks, standards, and conditions in accordance with the Universal Joint Task List or equivalent DOD Component Task List. If a capability requirement is not satisfied by a capability solution, then there is also an associated capability gap which carries a certain amount of risk until eliminated. A requirement is considered to be "draft" or "proposed" until validated by the appropriate authority. (JCIDS Manual 2015)

Confidence Interval – a specified range of numbers within which a population parameter should lie; an estimate of the population parameter based on the knowledge that it will equate the sample statistic plus or minus a sampling error. (Zikmund 1991)

Confidence Level – a percentage or decimal value that tells how confident a researcher can be about being correct. It states the long-run percentage of confidence intervals, including the true population mean. (Zikmund 1991)

Criteria (also referred to as Standards) – define the acceptable levels or standards of performance for a metric and are often expressed as a minimum acceptable level of performance (threshold) and desired acceptable level of performance (objective). (AFOTECMAN 99-101)

Data – individual measurements that are used to compute the metric for a measure. (AFOTECMAN 99-101)

Measure – a device designed to convey information about an entity being addressed. It is the dimensions, capacity, or amount of an attribute an entity possesses. (AFOTECMAN 99-101)

Measure of Effectiveness – a measure designed to correspond to accomplishment of mission objectives and achievement of desired results. (DAU Glossary)

Measure of Performance – a measure of a system's performance expressed as speed, payload, range, time on station, frequency, or other distinctly quantifiable performance features. (DAU Glossary)

Measure of Suitability – a measure of an item's ability to be supported in its intended operational environment. (DAU Glossary)

Metric – a unit of measure that coincides with a specific method, procedure, or analysis (e.g., function or algorithm). Examples include: mean, median, mode, percentage, and percentile. (AFOTECMAN 99-101)

Objective – an operationally significant increment above the threshold. An objective value may be the same as the threshold value when an operationally significant increment above the threshold is not identifiable. (AFI 10-601)

Sample – a portion or subset of a larger group called a population. (Fink 2003)

Survey – a system for collecting information from or about people in order to describe, compare, or explain their knowledge, attitudes, and behavior. (Fink 2003)

Threshold – a minimum acceptable operational value of a system capability or characteristic below which the utility of the system becomes questionable. (AFI 10-601)

Appendix E: Examples of Five, Six, and Seven Point Scale Descriptors

Five Point Scales				
Totally Inadequate	Somewhat Inadequate	Borderline	Somewhat Adequate	Totally Adequate
Very Inadequate	Slightly Inadequate	Borderline	Slightly Adequate	Very Adequate
Completely Unacceptable	Somewhat Unacceptable	Borderline	Somewhat Acceptable	Completely Acceptable
Largely Unacceptable	Barely Unacceptable	Borderline	Barely Acceptable	Largely Acceptable
Completely Ineffective	Somewhat Ineffective	Borderline	Somewhat Effective	Completely Effective
Very Ineffective	Ineffective	Borderline	Effective	Very Effective
Extremely Difficult	Somewhat Difficult	Borderline	Somewhat Easy	Extremely Easy
Completely Disagree	Substantially Disagree	Borderline	Substantially Agree	Completely Agree
Extremely Unimportant	Moderately Unimportant	Borderline	Moderately Important	Extremely Important
Completely Useless	Somewhat Useless	Borderline	Somewhat Useful	Completely Useful
Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Undoubtedly Worse	Moderately Worse	The Same	Moderately Better	Undoubtedly Better
Undoubtedly Worst	Noticeably Worse	Borderline	Moderately Better	Undoubtedly Best
Strongly Dislike	Dislike	Neutral	Like	Strongly Like
Never	Rarely	Now and Then	Often	Always
Never	Seldom	Now and Then	Frequently	Always
Poor	Fair	Good	Very Good	Excellent
Not at all Important	Not so Important	Neutral	Fairly Important	Very Important
Very Unimportant	Not Important	Borderline	Important	Very Important
Completely Dissatisfied	Somewhat Dissatisfied	Neither Satisfied Nor Dissatisfied	Somewhat Satisfied	Completely Satisfied
Hardly ever	Sometimes	Often	Very Often	All the Time
Never	Rarely	Sometimes	Often	Very Often
Not at all Unique	Slightly Unique	Somewhat Unique	Very Unique	Extremely Unique

Six Point Scales					
Totally Inadequate	Very Inadequate	Somewhat Inadequate	Somewhat Adequate	Very Adequate	Totally Adequate
Completely Unacceptable	Largely Unacceptable	Somewhat Unacceptable	Somewhat Acceptable	Largely Acceptable	Completely Acceptable
Completely Ineffective	Largely Ineffective	Somewhat Ineffective	Somewhat Effective	Largely Effective	Completely Effective
Extremely Difficult	Moderately Difficult	Somewhat Difficult	Somewhat Easy	Moderately Easy	Extremely Easy
Extremely Unimportant	Moderately Unimportant	Barely Unimportant	Barely Important	Moderately Important	Extremely Important
Completely Useless	Largely Useless	Somewhat Useless	Somewhat Useful	Largely Useful	Completely Useful
Completely Disagree	Substantially Disagree	Slightly Disagree	Slightly Agree	Substantially Agree	Completely Agree
Undoubtedly Worse	Moderately Worse	Slightly Worse	Slightly Better	Moderately Better	Undoubtedly Better
Never	Very Rarely	Somewhat Rarely	Somewhat Often	Very Often	Always

Seven Point Scales						
Totally Inadequate	Very Inadequate	Somewhat Inadequate	Borderline	Somewhat Adequate	Very Adequate	Totally Adequate
Totally Inadequate	Very Inadequate	Barely Inadequate	Borderline	Barely Adequate	Very Adequate	Totally Adequate
Completely Unacceptable	Largely Unacceptable	Somewhat Unacceptable	Borderline	Somewhat Acceptable	Largely Acceptable	Completely Acceptable
Completely Unacceptable	Moderately Unacceptable	Barely Unacceptable	Borderline	Barely Acceptable	Reasonably Acceptable	Completely Acceptable
Completely Ineffective	Largely Ineffective	Somewhat Ineffective	Borderline	Somewhat Effective	Largely Effective	Completely Effective
Completely Ineffective	Very Ineffective	Ineffective	Borderline	Effective	Very Effective	Completely Effective
Undoubtedly Worst	Conspicuously Worse	Moderately Worse	Alike	Moderately Better	Conspicuously Better	Undoubtedly Best
Extremely Difficult	Moderately Difficult	Somewhat Difficult	Borderline	Somewhat Easy	Moderately Easy	Extremely Easy
Completely Disagree	Substantially Disagree	Slightly Disagree	Borderline	Slightly Agree	Substantially Agree	Completely Agree
Extremely Unimportant	Moderately Unimportant	Barely Unimportant	Borderline	Barely Important	Moderately Important	Extremely Important
Completely Useless	Largely Useless	Somewhat Useless	Borderline	Somewhat Useful	Largely Useful	Completely Useful
Undoubtedly Worse	Moderately Worse	Slightly Worse	The Same	Slightly Better	Moderately Better	Undoubtedly Better
Never	Very Rarely	Somewhat Rarely	Borderline	Somewhat Often	Very Often	Always

Appendix F: Determining Sample Size Involving Means

This appendix provides two versions of the equation that would be used in determining the sample size involving means. Version 1 is the equation without a finite population adjustment factor, whereas Version 2 is the equation with a finite population adjustment factor. When the size of a sample exceeds 5 percent of the population, it is acceptable to recognize that the finite limits of the population put a constraint on the size of the sample needed (Emory 1985, p. 296).

Unlike the equation used to determine sample size involving proportions described in Section 1.84 (Determining the Sample Size for Probability Sampling), these equations require a variance estimate (either the sample standard deviation or estimate of the population standard deviation). Previous studies or a pilot study may be used to estimate the standard deviation. If information is not available to estimate the standard deviation, a rule of thumb is to expect the standard deviation to be one-sixth of the range (Zikmund 1991, p. 507). For example, if the time required to perform routine maintenance on a new electronic warfare system is expected to range from 1 to 7 hours, then the rule of thumb estimate of the standard deviation would be 1 hour.

The standard error of the mean ($\sigma_{\bar{x}}$) is calculated in the same way as the standard error of the proportion (σ_p) that is described in Section 1.84. The analyst specifies a confidence interval (e.g., ± 0.1) and confidence level (e.g., .95). The z-value associated with a .95 confidence level is 1.96 standard errors. One standard error of the mean ($\sigma_{\bar{x}}$) is determined by dividing the confidence interval by the standard errors (0.1/1.96 = 0.051).

With the standard error of the mean ($\sigma_{\bar{x}}$) and standard deviation (s) determined, the analyst simply solves the equation for n to determine the sample size.

Version 1

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Solving for n, the equation becomes:

$$n = \frac{s^2}{\sigma_{\bar{x}}^2}$$

Where:

$\sigma_{\bar{x}} = \text{standard error of the mean}$

$s = \text{sample standard deviation or estimate of the population standard deviation}$

$n = \text{size of the sample}$

Version 2

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Solving for *n*, the equation becomes:

$$n = \frac{N\left(\frac{s}{\sigma_{\bar{x}} \times \sqrt{N-1}}\right)^2}{1 + \left(\frac{s}{\sigma_{\bar{x}} \times \sqrt{N-1}}\right)^2}$$

Where:

$\sigma_{\bar{x}} =$ standard error of the mean

$s =$ sample standard deviation or estimate of the population standard deviation

$n =$ size of the sample

$N =$ size of the population