TRAC-M-TR-14-036
30 SEP 14

Applying Subject Matter Expert (SME) Elicitation Techniques to TRAC Studies

TRADOC Analysis Center -
Monterey
700 Dyer Road
Monterey, California 93943-0692

This study cost the
Department of Defense approximately
$33,000 expended by TRAC in
Fiscal Year 13-14.
Prepared on 20141006
TRAC Project Code # 060099.

This page intentionally left blank.

TRAC-M-TR-14-036
30 SEP 14

Applying Subject Matter Expert (SME) Elicitation Techniques to TRAC Studies

MAJ Michael Teter

TRADOC Analysis Center - Monterey
700 Dyer Road
Monterey, California 93943-0692

This study cost the Department of Defense approximately $33,000 expended by TRAC in Fiscal Year 13-14.
Prepared on 20141006
TRAC Project Code # 060099.

This page intentionally left blank.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
30-09-2014	Technical Report	15 Aug 13 - 30 Sep 14

4. TITLE AND SUBTITLE

Applying Subject Matter Expert (SME) elicitation techniques to TRAC studies

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Major Michael D. Teter

5d. PROJECT NUMBER

TRAC Project Number 060099

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

TRADOC Analysis Center - Monterey
700 Dyer Rd,
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

TRAC-M-TR-14-036

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

TRADOC Analysis Center - Military Research Office

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We are interested in quantifying measures which are qualitative and reside as knowledge held by a Subject Matter Expert (SME). This knowledge must be elicited from the SME and quantified in a logical manner to be used in analysis. Difficulties arise when trying to convey results in a mathematically rigorous way. To add to the complexity, we often combine the qualitative responses through estimating or averaging a single value that encompasses all information. We examine the applicable decision theory which supports mathematical representation of agreement among SMEs when they give ordinal responses. We pose a process in a logical and methodical way that draws from this theory. We follow this by demonstrating the application of these measures through three examples of SME collections in support of the Capability Portfolio Review (CPR) process, the Tactical Wheeled Vehicle (TWV) reduction study V, and the Army's selection of a portfolio of future technologies in which to invest.

15. SUBJECT TERMS

SME Elicitation, Measure of Consensus, Subject Matter Expert, Measure of Agreement, SME Reliability

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			MAJ Michael Teter
Unclassified	Unclassified	Unclassified	Unclassified	116	**19b. TELEPHONE NUMBER** *(include area code)* 831-656-7580

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

Table of Contents

List of Figures

List of Figures

This page intentionally left blank.

List of Tables

List of Tables

This page intentionally left blank.

Applying Subject Matter Expert (SME) Elicitation Techniques to TRAC Studies
Chapter 1
Introduction

Background

Often, we are interested in quantifying measures which are qualitative and reside as knowledge held by a Subject Matter Expert (SME). This knowledge must be elicited from the SME and quantified in a logical manner to be used in analysis. Difficulties arise when trying to convey results in a mathematically rigorous way. To add to the complexity, we often combine the qualitative responses through estimating or averaging a single value that encompasses all information.

In this paper, we do not attempt to exhaustively cover all topics within SME elicitation but contribute to the body of knowledge on the subject. For a thorough literature review of the range of topics and previous studies conducted by the TRADOC analysis centers see Marks, Smead, and Alt;[1]

Following a brief introduction, we examine the applicable decision theory in Chapter 2 which supports mathematical representation of agreement among SMEs when they give ordinal responses. We pose a process in a logical and methodical way that draws from this theory.

We follow this by demonstrating the application of these measures through three examples of SME collections in support of the Capability Portfolio Review (CPR) process, the Tactical Wheeled Vehicle (TWV) reduction study V, and the Army's selection of a portfolio of future technologies in which to invest.

In the appendixes, we include a the thorough literature review conducted previously (C), a white paper on conducting interviews written by an analyst with more than 20 years experience (D) and a review of another tool set available for quantifying elicited data (E).

[1]MAJ Christopher Marks, Ms. Kristen Smead, and LTC Jonathon Alt; *Enhancing Subject Matter Expert Elicitation Techniques*. Tech. rep. TRAC-M-TR-13-048. 700 Dyer Road Monterey, California 93943: TRADOC Analysis Center - Monterey, 2013.

Problem Statement

To expand the past project on gathering and employing input from subject matter experts (SMEs) in order to extend in greater depth SME elicitation practices within TRAC.

Issue 1: What are the best methods to design an SME elicitation?

Issue 2: What are methods to execute an SME eliciation plan?

Issue 3: How do current use-cases drive the plan for an SME elicitation design?

Constraints, Limitations, & Assumptions

- Constraints[2]
  - This effort will be complete NLT 30 September 2014.
- Limitations[3]
  - The scope of available SME elicitation planning and executing processes are limited to current TRAC studies.
- Assumptions[4]
  - - The SME elicitation plan for current studies provides detailed use-cases.

Approach

The approach for this project followed these steps:

- Review of supporting decision analysis theory that is applicable to TRAC studies.
- Three different use-case applications:
  - The Tactical Wheeled Vehicle (TWV) 5 study
  - The Science and Technology (S & T) portfolio investment
  - An application to Capability Portfolio Reviews (CPR)
- Additional materials collected in the appendixes include:

[2]Constraints limit the project team's options to conduct the research.
[3]Limitations are a project team's inabilities to investigate issues within the sponsor's bounds.
[4]Assumptions are research-specific statements that are taken as true in the absence of facts.

– Interviewing techniques presented in the form of a white paper
– Thorough Literature Review conducted in the previous project
– A review of the SHELF R-code and technique.

This page intentionally left blank.

Chapter 2
Applicable decision theory

Decision theory has proposed many different ways in which information can be quantified from qualitative information. There are three theories which we mention in this chapter: *Multi-Criteria Decision Analysis*, *Shannon Entropy*, and *Fuzzy Set Theory*.

Multi-Criteria Decision Analysis

There are multiple, competing objectives in Chapter §5 which are modeled. For the approach, we apply Multi-Criteria Decision Analysis (MCDA).

This field, which includes the seminal work of Keeney regarding Value Focused Thinking (VFT) Keeney,[5] is further explored with specific examples and implementation strategies in Kirkwood.[6] The contributions of Keeney and Kirkwood's research are the realizations of multiple, and often times competing, objectives that have to be considered when quantifying subjective inputs. It is imperative to develop appropriate value functions to use with available information when converting from qualitative data, e.g., high risk, to quantitative data, e.g., the number four. For an example of how this method can be used in data transformation, see Phillips and Costa.[7] For another example, see Ewing Jr, Tarantino, and Parnell;[8] in which this approach is used to determine the military value of an installation.

Shannon Entropy

In the field of signal processing, there is uncertainty surrounding the possibility of a signal received was not the intended signal sent. Different ways have been introduced to mitigate this difficulty.

Claude Shannon introduced Shannon Entropy in his seminal paper, "A Mathematical

[5] Ralph L. Keeney. *Value-Focused Thinking: A path to creative decisionmaking*. Harvard University Press, 2009.

[6] Craig W Kirkwood. *Strategic Decision Making*. Duxbury Press Belmont, CA, 1997.

[7] Lawrence D. Phillips and Carlos A. Bana e Costa. "Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing". English. In: *Annals of Operations Research* 154.1 (2007), pp. 51–68. ISSN: 0254-5330.

[8] Paul L Ewing Jr, William Tarantino, and Gregory S Parnell. "Use of decision analysis in the army Base Realignment And Closure (BRAC) 2005 military value analysis". In: *Decision Analysis* 3.1 (2006), pp. 33–49.

Theory of Communication" Shannon.[9] He applied logarithmic functions to random variables to produce an expected value of a binary signal received being the actual signal sent. He proposed a method to account for the uncertainty of receiving the wrong signal. His methodology forms the basis for fuzzy set theory discussed in Section §2. Figure 1 portrays the entropy of a binary response. The measure of uncertainty is represented by H, while p_1 is the probability of outcome 1. Both outcomes are equally likely, such as in the case of a coin flip ($p_1 = .5$), then the uncertainty of the outcome is the highest, i.e., $H = 1$. If the outcome is known, then the uncertainty is zero.

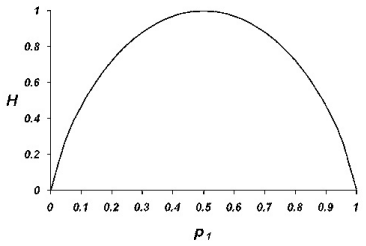


Figure 1. Shannon Entropy for binary response.

Shannon Entropy can be applied to information theory Lin,[10] and to genetics Jost,[11] and medical sciences De Araujo et al.,[12] among other fields. For our application, the relationship between signal processing in communications and signals sent by SME responses is clear, and both can be unintentionally misinterpreted from the original intent.

Fuzzy Set Theory

An extension of examining single, binary signals is expanding this to account for multiple possibilities of responses. This extension is necessary for our application because most SME responses are not binary.

When multiple responses are considered, we use fuzzy set theory to portray the results. Fuzzy set theory was first introduced in 1965 by Zadeh Zadeh.[13] Deschrijver and Kerre[14] covers the relationship between the original theory and some extensions, while Zimmermann[15] highlights both theory and applications. This method is appropriate to qualitative responses such as "high" and "low," which are not directly measurable, unlike

[9]Claude Elwood Shannon. "A mathematical theory of communication". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001), pp. 3–55.

[10]Jianhua Lin. "Divergence measures based on the Shannon Entropy". In: *Information Theory, IEEE Transactions on* 37.1 (1991), pp. 145–151.

[11]Lou Jost. "Entropy and diversity". In: *Oikos* 113.2 (2006), pp. 363–375.

[12]DB De Araujo et al. "Shannon Entropy applied to the analysis of event-related fMRI time series". In: *NeuroImage* 20.1 (2003), pp. 311–317.

[13]Lotfi A Zadeh. "Fuzzy sets". In: *Information and control* 8.3 (1965), pp. 338–353.

[14]Glad Deschrijver and Etienne E Kerre. "On the relationship between some extensions of fuzzy set theory". In: *Fuzzy sets and systems* 133.2 (2003), pp. 227–235.

[15]HJ Zimmermann. *Fuzzy Set Theory and Its Applications Second, Revised Edition*. Springer, 1992.

temperature or speed. The responses must be translated into a numerical form before computations are done. In Figure 2, a probability of a categorical response given a specific temperature can be calculated using Fuzzy Set Theory. The theory accounts for the possibility that the transitions between the categories are not mutually exclusive. For example, when it is 12 degrees Celsius, different respondents might consider it categorically "cold" or "warm."

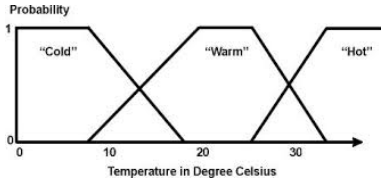


Figure 2. Fuzzy set example.

Often, categories such as "high" will be assigned an ordinal number, e.g., high = 3. In this way, all of the responses are easily converted to numerical representations. These numbers are treated as quantitative results to which analysis is applied.

One difficulty which may arise using a classical approach, as opposed to fuzzy set theory, is the underlying assumption that the ordinal responses are equidistant from each other on a scale, e.g., on a three-point ordinal scale, the magnitude between 1 and 2 and between 2 and 3 are equal. Another difficulty is mapping the statistics back to the original categorical answers. Consider the following mapping represented in Table 1. What conclusions can be drawn from a mean of 2.2? Does this represent a moderate response or a high response? There is no possible moderate-and-a-half response.

Response	Ordinal Value
Low	1
Moderate	2
High	3

Table 1. Sample mapping of categorical responses.

Measuring the strength of agreement among the responses is difficult. Commonly, the standard deviation is used on ordinal data mapped to the categories. This can be misleading as discussed in the previous paragraph and the standard deviation is also susceptible to the magnitude of the scale used to portray the data. Another measure addresses these difficulties as demonstrated in Chapter 5 and presented in the following section.

Using measures of consensus, dissension, and agreement

The theories mentioned in Sections 2 and 2 combine to form the basis for computing Tastle-Wierman measures Tastle and Wierman,[16] Tastle and Wierman,[17] Tastle and Wierman,[18] Tastle, Russell, and Wierman.[19] Since the input from SME elicitation is usually categorical in nature, it can be misleading to represent multiple responses to a single question with a simple point estimate of the mean and the standard deviation. The point estimate does not use all available information from the elicitation, such as proximity of an SME's response to that of another SME. An SME response in a category on an ordinal scale is more likely to agree with categories adjacent. Therefore, Tastle-Wierman derived three measures that are based on fuzzy set theory which do account for proximity.

To explain the notation in the equations, we will assume we have an elicitation with four possible responses on an ordinal scale from 1 through 4. The value we are trying to derive is $\hat{Z}$. The random variable Z_i is an ordinal value an SME assigns to a response, e.g., a number between 1 and 4. The subscript i denotes all possible values the random variable Z_i can assume, in our example, $i = \{1, 2, 3, 4\}$. The parameter $d_{\tilde{z}}$ represents the distance between the lowest possible score of Z_i (=1) and the highest possible of Z_i (=4). In our example, $d_{\tilde{z}} = 3$. The parameter τ represents the targeted response among the possible values the random variable can assume, e.g., 1-4. The parameters p_i and $\mu_{\tilde{z}}$ are estimated using standard statistical calculations for a discrete random variable with a binomial distribution since the category is either selected or not.

- Consensus is a measure of dispersion from the mean. The result is a score between zero and one. The closer to one, the more homogeneous the SME responses are around the selected ordinal value; the closer to zero, the more disagreement there is amongst the SME responses.

$$\text{Consensus}(\hat{Z}) = 1 + \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|Z_i - \mu_{\tilde{z}}|}{d_{\tilde{z}}}\right) \tag{2.1}$$

- The dissent measure is the complement of the consensus. For example, if the consensus measure is .75 then the dissent measure is .25.

[16] William J Tastle and Mark J Wierman. "Consensus and dissention: A measure of ordinal dispersion". In: *International Journal of Approximate Reasoning* 45.3 (2007), pp. 531–545.

[17] William J Tastle and Mark J Wierman. "Using consensus to measure weighted targeted agreement". In: *Annual Meeting of the North American Fuzzy Information Processing Society, 2007.* IEEE. 2007, pp. 31–35.

[18] William J Tastle and Mark J Wierman. "Consensus and dissention: a new measure of agreement". In: *Annual Meeting of the North American Fuzzy Information Processing Society, 2005.* IEEE. 2005, pp. 385–388.

[19] William J Tastle, Jack Russell, and Mark J Wierman. "A new measure to analyze student performance using the Likert scale". In: *Information Systems Education Conference: Proceedings of ISECON.* 26 (2005), p. 2007.

$$\text{Dissent}(\hat{Z}) = -\sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|Z_i - \mu_{\tilde{z}}|}{d_{\tilde{z}}}\right) \tag{2.2}$$

- The measure of agreement is a normalized measure of dispersion around a targeted response, rather than the mean. This measure has value when we are concerned with reaching a certain threshold minimum response on an ordinal scale. The distance parameter $d_{\tilde{z}}$ is multiplied by two since we have two degrees of freedom, $\hat{Z}$ and τ.

$$\text{Agreement}(\hat{Z}, \tau) = 1 + \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|Z_i - \tau|}{2d_{\tilde{z}}}\right) \tag{2.3}$$

Use Cases

We now examine three different studies in which these measures could or were employed. Each of these approaches use the measures in different ways.

In the first case, CPR Analisys, we use the measures as way to derive effectiveness and uncertainty. In the second approach, Tactical Wheeled Vehicle (TWV) V study4, we use the measure of agreement to see if consensus is reached around a targeted response. In the last approach presented, Science & technology subject matter expert elicitation, the measure is used to judge consensus of the experts in there assessment of future technologies capabilities.

This page intentionally left blank.

Chapter 3
Capability Portfolio Review (CPR) Analysis

The US Army uses a specific process called a Capability Needs Analysis (CNA) to assess perceived shortfalls in its ability to perform certain missions given a set of tasks. The shortfalls are classified as capability gaps while the missions are considered requirements. Once these gaps have been identified, the Army explores means by which to overcome these capability gaps through the implementation of solutions, e.g., changes to doctrine, organizational structure, training, equipment purchases, personnel or facilities. Each solution is then assessed as to how successfully it closes a capability gap, i.e., as to its effectiveness.

Because this process addresses future outcomes, there is associated uncertainty with each solution applied towards each capability gap. The US Army uses a measurement of uncertainty to calculate risk.

Deriving uncertainty and effectiveness

The measures in Section 2 are applied to the SME responses to derive the uncertainty and effectiveness. These parameters are used as inputs to other models to examine tradespace . Here, we cover the computation of the measures and how they apply.

Elicitation process

It is necessary to discuss a possible SME elicitation process that produce the results we use in this section to derive the measures. The SMEs are separated into groups to consider a single solution with respect to its ability to meet 27 different capability gaps. There are six groups of SMEs, one for each solution. The number of SMEs in each group is not the same. The smallest group has seven SMEs and the largest had 27. For each solution, an SME individually assigns an effectiveness value categorized on an ordinal scale mapped to a number between 1 and 4. Their individual values are aggregated for each capability gap per solution, resulting in 27 values for each of six solutions.

Deriving uncertainty among the SMEs

To derive uncertainty among the SME responses, the measure of dissent shown in Equation (2.2) is applied to the SMEs' response on how effective a solution is in addressing a capability gap. The measure of dissent uses the dispersion around the mean of the SME

responses to calculate a normalized number between zero and one. The closer to one the measure then the more disagreement there is among the SMEs.

We use the measure for uncertainty because we tie it directly to the uncertainty of the solution's effectiveness among the SMEs. Disagreements about how effective (or not effective) a solution may be in addressing a capability gap is measured as the uncertainty. The more SMEs agree on the effectiveness, the lower the variability.

Capability Gap (g)	Low	Mod	High	Extreme	Consensus	Dissent
1	27	0	0	0	1	0
2	13	0	0	13	0	1
3	3	9	0	0	0.802	0.198
4	1	11	0	0	0.919	0.081
5	15	5	7	0	0.546	0.454
6	5	5	1	2	0.486	0.514
7	12	6	9	0	0.54	0.46

Table 2. Example of measure of dissent of one solution against seven gaps

As an example of this method applied to SME responses regarding a single solution's effectiveness in addressing a capability gap, refer to Table 2. The two extremes are demonstrated in the results of capability gaps 1 and 2. Gap 1 is an example of complete agreement among SMEs; 27 agree that a solution has a low effect on a capability gap. The consensus has a value of 1 while the dissent has a value of 0. Gap 2 displays complete disagreement among the group; the response of the SMEs is bi-polar in that 13 selected low effectiveness and 13 selected extreme. Since this group is evenly divided at the extremes, it is measured with a consensus value of 0 and dissent value of 1.

The next two, gaps 3 and 4, show lower uncertainty amongst the SMEs than gap 2 exhibits. They are lower due to the proximity of their responses on the ordinal scale and accounted for in Equation (2.2).

For gaps 5, 6, and 7, it may appear that the SME responses generally agreed towards the low end of effectiveness since the majority of responses were low or moderate. The dissention among the SMEs is actually high because the responses are spread across multiple answers rather than concentrated, as in gaps 3 and 4.

Deriving effectiveness

For the effectiveness coefficient of a solution, we use the product of the measure of agreement (2.3) and the ordinal number associated with the SME response and τ. The calculation for a four-category-ordinal scale is represented in Table 3.

To quantify the effectiveness of each solution in addressing a given capability gap, we aggregate the values and normalize them to a single number between zero and one. The

Response	Ordinal Value	Effectiveness
Low	1	Agreement($\hat{Z}$,1) $\times$ 1
Moderate	2	Agreement($\hat{Z}$,2) $\times$ 2
High	3	Agreement($\hat{Z}$,3) $\times$ 3
Extreme	4	Agreement($\hat{Z}$,4) $\times$ 4

Table 3. Four-category-ordinal scale effectiveness calculation

generalized effectiveness score, using notation from Equation (2.1), is calculated for each value; see Equation (3.1).

$$\text{Effectiveness}(Solution) = \frac{\sum_{i=1}^{n} \left(\text{Agreement}(\hat{Z}, \tau) \cdot i\right)}{\sum_{i=1}^{n} i} \tag{3.1}$$

Displayed in Table 4 are sample results from the data collection from an SME elicitation described in Section 3, along with the calculated effectiveness, consensus and dissent values for one solution compared against six capability gap.

Comparing the results formulated using Equation (3.1) in Table 4 to a calculated mean of the SME responses provides insight to observed differences. For example, gap 6 has eleven responses in the high and extreme effectiveness while gap 4 had eleven in the same two categories. The mean score of gap 6 and 4, 3.14 vice 2.86, respectively, suggests gap 6 is more effectively addressed by the solution. Using the effectiveness equation, results in gap 4 as the most effective in address the gap by the solution.

Capability gap (g)	Low	Moderate	High	Extreme	Effectivness	Consensus	Dissent
1	5	5	1	2	0.5657	0.486	0.514
2	3	11	0	0	0.5989	0.822	0.178
3	3	10	1	0	0.6077	0.796	0.204
4	0	3	10	1	0.7517	0.796	0.204
5	0	4	10	0	0.7486	0.785	0.215
6	1	2	5	6	0.7027	0.539	0.461

Table 4. Example of the effectiveness score for one solution against six gaps

Note: All the measures presented in this section were coded and solved using R Venables, Smith, and R Core Team,[20] Version 3.0.3.

[20]W. N. Venables, D. M. Smith, and the R Core Team. *An Introduction to R.* Version 3.0.3. CRAN. 2014.

This page intentionally left blank.

Chapter 4
Tactical Wheeled Vehicle (TWV) V study

Background

The Army directed TRADOC to conduct a study to reduce the amount of TWV's in the ground fleet. To achieve the appropriate analysis, the study objective was defined "To provide a recommendation to the CSA that reduces TWV requirements in TOE units to an acceptable level such that the gains to the Army (Cost Savings, Greater Deployability and Agility) outweigh the losses (Capabilities, Mission Accomplishment, and Flexibility)."

TRAC-LEE was asked to support the study, lead by the Army's Capability Integration Center (ARCIC), through designing an SME elicitation workshop. The SME's would be asked to evaluate acceptable levels of risk for various courses of action. The workshop was a three day event that took place from 13-15 January, 2014.

The Survey Construction

A workshop even was hosted in which there was live SME interaction. Surveys were conducted using individual workstations through the FactPro software. Free text fields were limited and not used for this portion although they were used in the study analysis. All surveys were conducted using a Likert-Type ordinal scale with the following mapping:

Response	Ordinal Value
Low	1
Moderate	2
High	3
Extremely High	4

Table 5. TWV mapping of responses.

For all of the survey's, there was a targeted response of interest (3-High) and a desire to measure the consensus of the group around the response given. The measures of consensus and agreement described in Chapter 2 are appropriate for this analysis and the results are shown in Section 4.

The SMEs

The selection of SME's was conducted through ARCIC with most of them being either from TRADOC (ARCIC or the appropriate COE) or an operational unit within the WfF. The SME's were divided into six groups by Warfighting Functions (WfF):

Note: the number of SMEs in each WfF is in parenthesis.

- SUSTAINMENT (17)
- MOVEMENT & MANUEVER (26)
- FIRES (11)
- PROTECTION (14)
- MISSION COMMAND (12)
- INTELLIGENCE (7)

They were given the opportunity to create and discuss three plausible courses of action that varied the amount of TWVs reduced. The combined results were then briefed to all SMEs present. The SMEs were then asked to evaluate the impact to the operational risk (RC) and the commander's freedom of action (FoA) given different echelons of forces (Belts) along with the overall impacts and acceptability of the course of action.

The WfF surveys

The first group of surveys asked the SMEs to consider only the units for which their WfF was responsible within the different belts and COAs. The second grouping of surveys all SMEs were asked to evaluate every COA from the Army perspective rather than their own WfF in terms of risk, freedom of action and concurrence or non-concur with implementation.

Results

The initial results display two different measures, the level of consensus and target level of agreement (high or extremely high). The level consensus is in the column labeled "Cons". The blue highlighted cells shows consensus was not reached at the 60% level among the SMEs. This measure is the dispersion amongst the mean. The other aspect of measuring the level of agreement around the possible responses is reflected in the column "Target > 80%". Columns highlighted red reached either high or extremely high at greater than 80% agreement. Any of the four possible responses that reached 80% is listed in words in the column.

Figure 3 are the results of an SME elicitation in which the SMEs were asked to use professional military judgment to evaluate the level of risk and restriction to freedom of maneuver placed on a field commander if a certain course of action is chosen. The categorical responses available to the SMEs were low, moderate, high and extremely high. There results were then analyzed using the measure of consensus mentioned in Chapter 2.

	COA 1		COA A		COA NA	
	Cons	Target > 80%	Cons	Target > 80%	Cons	Target > 80%
Belt 1 RC	100%	Low	74%	Low, Mod	79%	High, Ext
Belt 1 FoA	74%	Low, Mod	87%	Mod	100%	High
Belt 2 RC	87%	Low	74%	Low, Mod	72%	High
Belt 2 FoA	79%	Low	83%	Mod	87%	High
Belt 3 RC	87%	Low, Mod	79%	Low, Mod	44%	NONE
Belt 3 FoA	79%	Low, Mod	72%	Mod	63%	Low,Mod
Overall Risk	74%	Low, Mod	74%	Low, Mod	70%	High
Overall FoA	79%	Low, Mod	83%	Mod	100%	High
Overall Acpt	74%	Low, Mod	79%	Low, Mod	79%	High, Ext

Figure 3. WfF Measure results.

Of interest in the results is the additional information the measure of agreement provides. For example in the Overall Risk of COA NA, the consensus was only measured at 70% yet the level of agreement around the response of High was over 80%. This allowed the study to focus on the SME responses in which the level was high or extremely high.

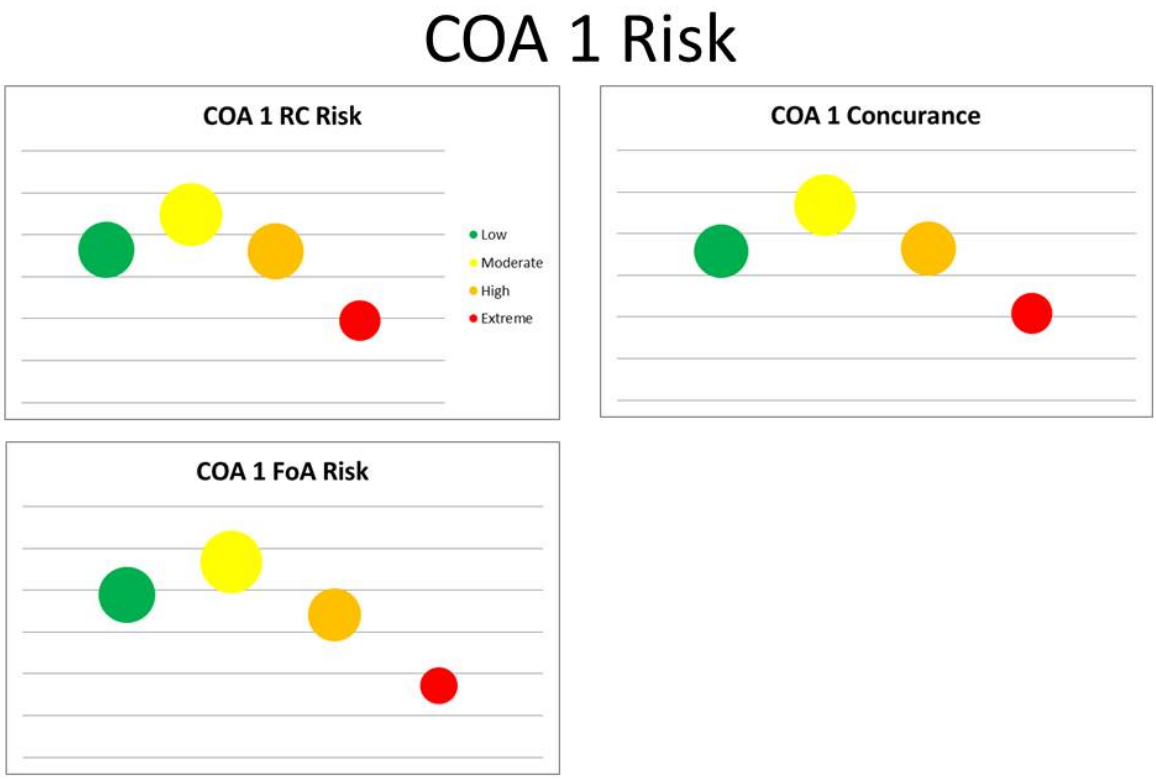


Figure 4. Consolidated Measure results.

During the results of the overall survey, we visualized the results in Figure 4 to show the trend of agreement among the SMEs given the four possible responses. The closer to the top of the chart and the larger the circle, the higher the level of agreement among the SMEs. This technique is a way to compare the results among all possible outcomes.

The measure of consensus was useful in identifying the SME surveys in which there was not agreement within the group and would have to be examined further. For all results please see Appendix A.

Chapter 5
Science & Technology (S & T) SME elicitation

In this chapter, we describe an example to demonstrate the methods presented in the previous chapter. We begin by giving background information for motivation, followed by a description of the SME data collection event. During this event, the responses were analyzed to present initial results.

Background

The chief of staff of the Army, GEN Odierno, is planning for the "Force 2025 and Beyond" and has ordered the Army staff to begin to determine the form the US Army should assume by 2025. In support of this effort, TRADOC is directed to recommend a portfolio of technologies in which to invest in support of Force 2025.

TRADOC limited these candidate technologies to a list of 254. This list was further reduced to 60 technologies through input from 11 Centers of Excellence (functional commands within TRADOC, e.g., sustainment, aviation) and the Army Capabilities and Integration Center (ARCIC). The subset of 60 technologies was examined closely by representatives from 39 organizations within the Army during a Science and Technology workshop held in Suffolk, VA from the 28th through the 30th of May 2014. This event produced scores for the 60 technologies that could be used as inputs to an optimization model with the objective of recommending the top 25 technologies for investment.

Method

There were 39 organizations acting as Subject Matter Experts (SMEs), each of which held one vote for the scoring event. The SMEs received a three-minute brief on a technology, immediately followed by a three-minute window in which they assigned, on a sliding scale between 1 and 100, the expected effects the technology would contribute to the following three capability traits: *efficiency, expeditionary, and dominance*, defined as follows:

- *Efficiency*: Army efficiency is the ability to provide greater or equal capabilities from a smaller pool of available resources. The Army must execute all of its assigned roles, functions, and missions using the least possible personnel, materials, funds, time, and/or energy.
- *Dominance*: Dominance is the ability to prevail in all domains and to control the operational environment by overmatching enemy capabilities at the critical time and

place across the full range of military operations and phases of the operation (Phase 0-5) to establish conditions to break the enemy's will to resist and achieve strategic objectives.

- *Expeditionary*: Expeditionary capability is the ability to promptly deploy combined arms forces worldwide into any area of operations and conduct operations upon arrival. Expeditionary operations require the ability to deploy quickly with little notice, rapidly shape conditions in the operational area, and operate immediately on arrival, exploiting success and consolidating tactical and operational gains. Expeditionary capabilities are more than physical attributes; they begin with a mindset that pervades the force.

Sixty technologies for which the SMEs assessed the capability traits resulted in a matrix of dimension approximately 39 by 180. Some peculiarities in the voting were: (i) switching voters during the scoring and (ii) not every voter completing every survey, resulting in some technologies only having 38, rather than 39, scores. These peculiarities were evident because of the unique identifiers assigned to voters.

Analysis of results during workshop

During the execution of the scoring event, each SME capability score assigned to a technology was analyzed as an independent survey, resulting in 180 different surveys.

- Inputs: voters (≈ 39), technologies (60), categories (3).
- Outputs: SME assessment score for each technology in each category (≈ 7020).
- Analyst team calculated:
  - Descriptive statistics (mean, SD, median, range, etc.).
  - Coefficient of Variation.
  - Measure of Consensus (see Chapter 2).
  - Scatter and Box-plot.
  - T-test, ANOVA, and Kruskal-Wallis to test for voting blocks.
  - Shapiro-Wilkes test for normality.
- Analyst team examined:
  - Scatter-plot and Box-plot visual inspection.
  - High compliment of Coefficient of Variation (1− Coefficient of Variance > 0.4).
  - High Measure of Consensus (Consensus > 0.6).

For an exemplar of the results see Appendix B

Appendix A
TWV Compiled Results

The following tables represent the results of various surveys taken during the TWV study workshop (Survey title in the first columns). Each column is described in the following table A–1.

Table A–1. Column Descriptions

Column 1	Survey Title in Bold Face
Column 2	Course of Action 1 Consensus measure
Column 3	Course of Action 1 Agrrement measure above .8 for any of the four possible responses
Column 4	Course of Action A Consensus measure
Column 5	Course of Action A Agrrement measure above .8 for any of the four possible responses
Column 6	Course of Action NA Consensus measure
Column 7	Course of Action NA Agrrement measure above .8 for any of the four possible responses

The four possible survey responses and color explanations are are listed in Table A–2.

Table A–2. Possible Responses and Colored Cells

Possible Survey Response	
Low Moderate High Extremely High	
Color Descriptions	
	Measure of Consensus Below .6
	Measure of Agreement above .8 for High or Extremely High

Table A–3. Protection WfF

	COA 1		**COA A**		**COA NA**	
	Cons	**Target > 80%**	**Cons**	**Target > 80%**	**Cons**	**Target > 80%**
Belt 1 RC	49%	NONE	72%	Low, Mod	70%	Mod, High
Belt 1 FoA	70%	Mod	72%	Mod	76%	Mod, High
Belt 2 RC	59%	Mod	76%	Low, Mod	72%	Mod, High
Belt 2 FoA	64%	Low, Mod	68%	Mod	74%	Mod, High
Belt 3 RC	63%	Mod	72%	Mod	93%	High
Belt 3 FoA	70%	Mod	72%	Mod	72%	Mod, High
Overall Risk	59%	Mod	82%	Mod	80%	High
Overall FoA	71%	Low, Mod	80%	Mod	79%	Mod, High
Overall Acpt	64%	Mod	93%	Mod	54%	High

Table A–4. Protection WfF

	COA 1		COA A		COA NA	
	Cons	**Target > 80%**	**Cons**	**Target > 80%**	**Cons**	**Target > 80%**
Belt 1 RC	49%	NONE	72%	Low, Mod	70%	Mod, High
Belt 1 FoA	70%	Mod	72%	Mod	76%	Mod, High
Belt 2 RC	59%	Mod	76%	Low, Mod	72%	Mod, High
Belt 2 FoA	64%	Low, Mod	68%	Mod	74%	Mod, High
Belt 3 RC	63%	Mod	72%	Mod	93%	High
Belt 3 FoA	70%	Mod	72%	Mod	72%	Mod, High
Overall Risk	59%	Mod	82%	Mod	80%	High
Overall FoA	71%	Low, Mod	80%	Mod	79%	Mod, High
Overall Acpt	64%	Mod	93%	Mod	54%	High

Table A–5. Intelligence WfF

	COA 1		COA A		COA NA	
	Cons	**Target > 80%**	**Cons**	**Target > 80%**	**Cons**	**Target > 80%**
Belt 1 RC	100%	Low	74%	Low, Mod	79%	High, Ext
Belt 1 FoA	74%	Low, Mod	87%	Mod	100%	High
Belt 2 RC	87%	Low	74%	Low, Mod	72%	High
Belt 2 FoA	79%	Low	83%	Mod	87%	High
Belt 3 RC	87%	Low, Mod	79%	Low, Mod	44%	NONE
Belt 3 FoA	79%	Low, Mod	72%	Mod	63%	Low,Mod
Overall Risk	74%	Low, Mod	74%	Low, Mod	70%	High
Overall FoA	79%	Low, Mod	83%	Mod	100%	High
Overall Acpt	74%	Low, Mod	79%	Low, Mod	79%	High, Ext

Table A–6. Mission Command WfF

	COA 1		COA A		COA NA	
	Cons	**Target > 80%**	**Cons**	**Target > 80%**	**Cons**	**Target > 80%**
Belt 1 RC	54%	Low	92%	Mod	72%	High
Belt 1 FoA	68%	Low, Mod	85%	Mod	83%	High
Belt 2 RC	74%	Low	85%	Mod	74%	High
Belt 2 FoA	68%	Low, Mod	85%	Mod	80%	Mod, High
Belt 3 RC	91%	Low	77%	Low, Mod	72%	High
Belt 3 FoA	79%	Low, Mod	80%	Low, Mod	68%	Mod, High
Overall Risk	74%	Low	80%	Low, Mod	90%	High
Overall FoA	76%	Low, Mod	80%	Low, Mod	100%	High
Overall Acpt	67%	Low, Mod	92%	Mod	74%	High, Ext [t]

Table A–7. Movement & Maneuver WfF

	COA 1		COA A		COA NA	
	Cons	**Target > 80%**	**Cons**	**Target > 80%**	**Cons**	**Target > 80%**
Belt 1 RC	58%	Mod	74%	Mod	62%	High
Belt 1 FoA	66%	Mod	76%	Mod	66%	High
Belt 2 RC	58%	Mod	67%	Mod	79%	High
Belt 2 FoA	67%	Mod	82%	Mod	73%	Mod, High
Belt 3 RC	55%	Low	67%	Mod	57%	Mod
Belt 3 FoA	56%	Mod	69%	Mod	59%	Mod
Overall Risk	54%	NONE	75%	Mod	68%	High
Overall FoA	62%	Mod	83%	Mod	67%	High
Overall Acpt	53%	Mod	78%	Mod	59%	High [t]

Table A–8. Fires WfF

	COA 1		COA A		COA NA	
	Cons	**Target > 80%**	**Cons**	**Target > 80%**	**Cons**	**Target > 80%**
Belt 1 RC	84%	Mod	100%	Mod	74%	High, Ext
Belt 1 FoA	84%	Mod	91%	Mod	91%	High
Belt 2 RC	79%	Low, Mod	91%	Mod	74%	High, Ext
Belt 2 FoA	89%	Mod	91%	Mod	89%	High
Belt 3 RC	74%	Low, Mod	91%	Mod	62%	High
Belt 3 FoA	91%	Mod	89%	Mod	75%	High
Overall Risk	79%	Low, Mod	100%	Mod	71%	High, Ext
Overall FoA	91%	Mod	91%	Mod	91%	High
Overall Acpt	100%	Mod	100%	Mod	84%	Ext [t]

Table A–9. Sustainment WfF

	COA 1		COA A		COA NA	
	Cons	**Target > 80%**	**Cons**	**Target > 80%**	**Cons**	**Target > 80%**
Belt 1 RC	74%	Low, Mod	81%	Mod	79%	High
Belt 1 FoA	70%	Low, Mod	86%	Mod	83%	High
Belt 2 RC	64%	Low, Mod	88%	Mod	73%	High
Belt 2 FoA	66%	Low, Mod	78%	Mod	83%	High
Belt 3 RC	70%	Mod	86%	Mod	77%	High
Belt 3 FoA	66%	Low, Mod	78%	Mod	86%	High
Overall Risk	68%	Low, Mod	88%	Mod	75%	High
Overall FoA	70%	Low, Mod	83%	Mod	88%	High
Overall Acpt	75%	Low, Mod	79%	Mod	43%	Ext [t]

This page intentionally left blank.

Appendix B
S & T Sample Results

In this appendix one technology scored in three qualities is displayed for exemplar purposes only. For the actual results of the 60 technologies scored contact Mr. Saul Solis, study lead, at TRAC-WSMR The TRAC study title is *Force 2025 and Beyond Science and Technology*. There were three qualities under consideration for each technology in which the SMEs scored as described in Chapter 5; Dominance, efficiency and expeditionary.

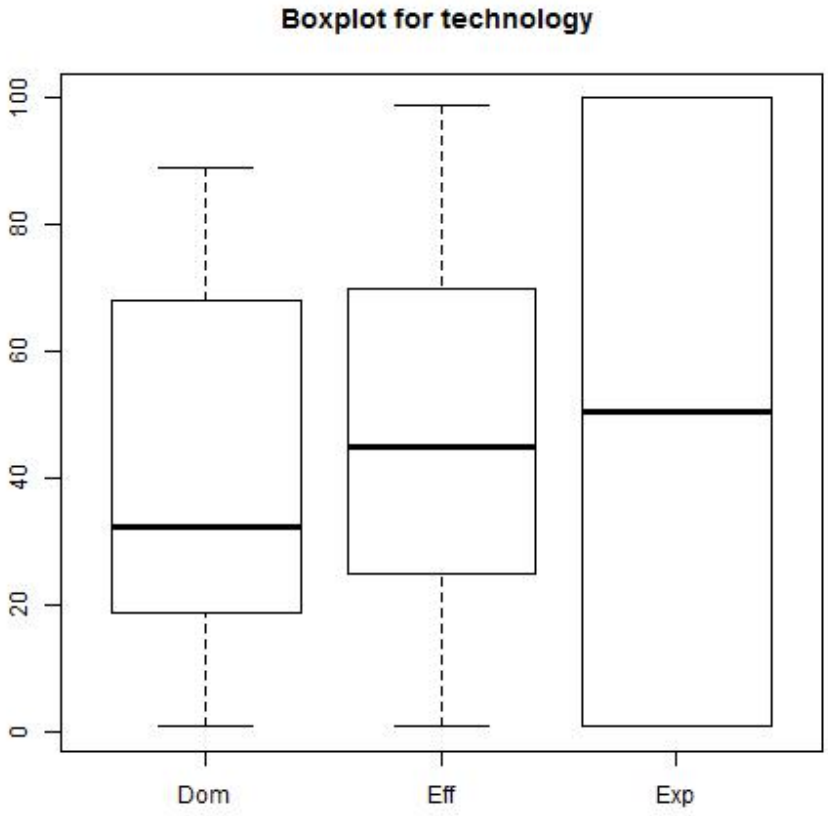
In Table B–1, the numeric results of the tests described in Chapter 5 are displayed.

Table B–1. Sample Technology result

	MU	median	SD	CV	con	SW
Dom	41.80952	32.5	28.92989	0.308055	0.548	0.004419
Eff	48.64286	45	28.09984	0.422323	0.576	0.060055
Exp	50.5	50.5	50.10002	0.00792	0	5.97E-09

The three qualities were compared in box-plots as represented in figure B–1.

Figure B–1. Box-Plots



For inspection, the data points were also displayed in histograms (See Figure B–2) and scatter-plots (See Figure B–3) although more weight was given to the numeric results of the scores assigned by the SMEs.

Figure B–2. Histogram

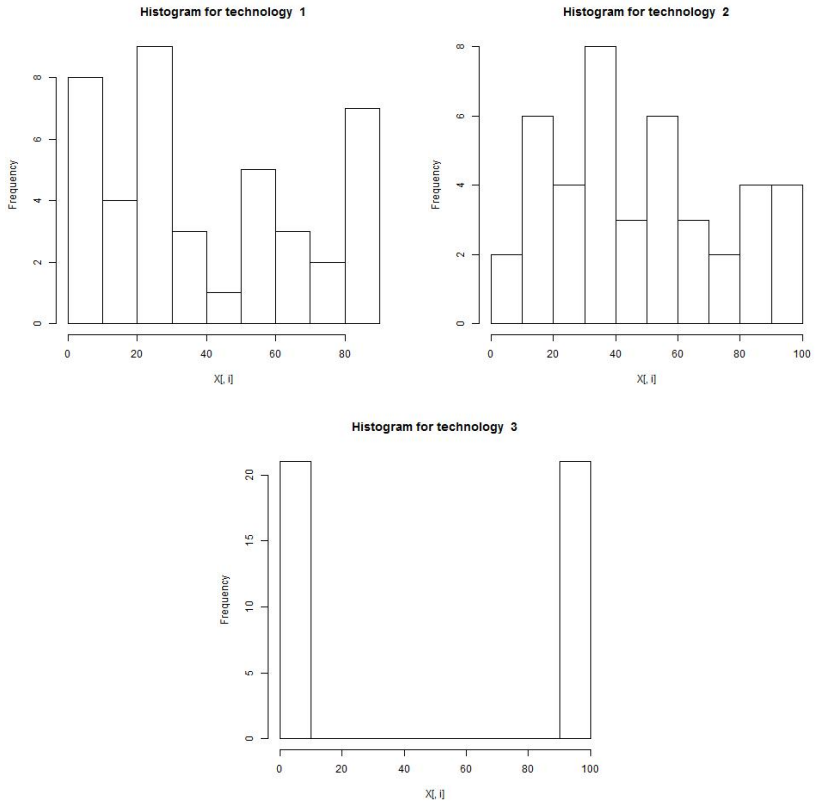
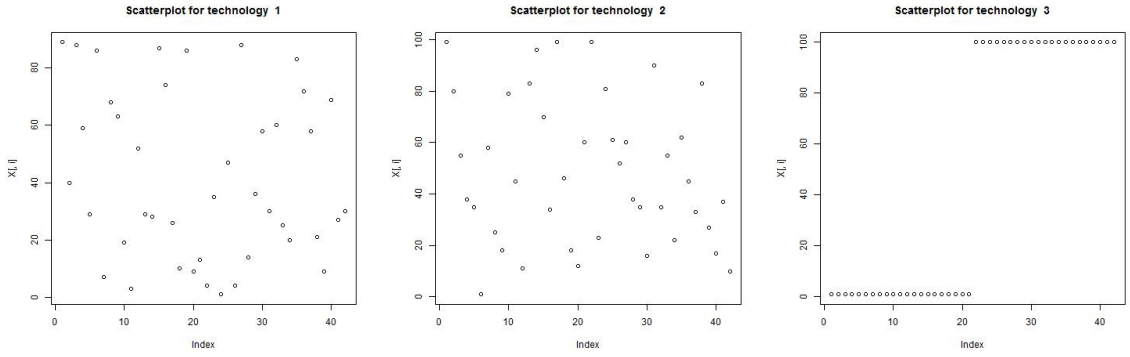


Figure B–3. Scatter-Plots



This page intentionally left blank.

Appendix C
Background & Review of Literature[1]

This appendix provides a review of recent literature documenting procedures and methods for accurately gathering and analyzing expert judgment. Many studies and analyses from a variety of fields rely on SME elicitation as a source of data, typically because objective data that meets the requirements of the study either does not exist or is too costly to obtain. Elicitation data, however, is generally subjective in nature. There are two potential sources of error, or *bias*, when using SME elicited data to inform a study or analysis Anderson and Johnson and Meyer and Booker:[2]

(1) Elicitation bias. Elicitation bias occurs with the data collected during elicitation does not accurately conform with the real world phenomena the analyst is attempting to quantify through elicitation. While expert judgment cannot always be externally calibrated, there are actions the analysts can take to minimize this type of bias, which are discussed in section C. We consider two types of elicitation bias:

   – Motivational bias, when an expert does not report his true judgment in his responses.

   – Cognitive bias, when an expert's judgment does not accurately represent the quantities of interest to the analysts. Because experts cannot be perfectly externally calibrated, i.e., they will always be wrong sometimes, cognitive bias can never be completely eliminated. However, there are some sources of cognitive bias that can be controlled through good elicitation practices, especially good question structuring, which we discuss in section C.

(2) Modeling bias. This type of bias occurs when the analysts use flawed or otherwise inappropriate methods of employing the SME-elicited data into the analysis. In section C we present numerous techniques for using SME-elicited data in analyses, providing some advantages and disadvantages of each. When choosing a method of applying or

[1] This is a reproduction of the literature review completed in a previous TRAC technical report Marks, Smead, and Alt; (MAJ Christopher Marks, Ms. Kristen Smead, and LTC Jonathon Alt; *Enhancing Subject Matter Expert Elicitation Techniques*. Tech. rep. TRAC-M-TR-13-048. 700 Dyer Road Monterey, California 93943: TRADOC Analysis Center - Monterey, 2013)

[2] Dr. Michael R. Anderson and Mr. Eric E. Johnson. *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*. TRADOC Analysis Center Methods and Research Office. Powerpoint presentation. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center Methods and Research Office, 2009; Mary A. Meyer and Jane M. Booker. *Eliciting and Analyzing Expert Judgement, A Practical Guide*. 3600 University City Science Center, Philadelphia, PA 19104-2688: Society for Industrial and Applied Mathematics, 2001. ISBN: 0-89871-474-5.

analyzing SME-elicited data, it is very important for analysts to justify all assumptions related to that method and be aware of exactly what the results mean, as well as what the don't mean. practicalguide give two types of modeling bias:

- Tool bias. This occurs when analysts use an inappropriate statistical method for incorporating the SME data into the analysis.
- Training bias. This occurs when the analysts misrepresent or inaccurately characterize when they encode the expert responses into data. Section C discusses some methods for encoding and conducting initial analysis on the expert response data.

An elicitation is successful if it minimizes both sources of error to the extent possible, i.e., its output is an accurate depiction of the expert's true judgment on the quantities of interest that serves to inform research objectives Garthwaite, Kadane, and O'Hagan.[3] However, even a perfect elicitation can yield inaccurate data, because there is always the possibility that the expert judgment is wrong. Analysts relying on SME-elicited data should be mindful of this potential source of error, which cannot be mitigated. Due to its subjective nature and the difficulties involved in accurately obtaining it, analysts should avoid assigning too much weight to SME-elicited data and should appropriately caveat the analysis Garthwaite, Kadane, and O'Hagan; O'Hagan; and Ayyub.[4]

Review of Literature on Elicitation

practicalguide provide the primary reference for the SME elicitation literature review. The remainder of this chapter roughly follows the outline of this reference, and we present many of their ideas in this paper Meyer and Booker.[5] While we do not rely as heavily on it, o2006uncertain provide another good resource for SME elicitation O'Hagan et al.[6]

What is SME Elicitation?

SME elicitation provides a means for obtaining expert judgment and expressing it in a statistically useful form Garthwaite, Kadane, and O'Hagan.[7] practicalguide give several alternative descriptions of "expert judgment." They are Meyer and Booker:[8]

[3]Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. "Statistical methods for eliciting probability distributions". In: *Journal of the American Statistical Association* 100.470 (2005), pp. 680–701.

[4]Idem, "Statistical methods for eliciting probability distributions", op. cit.; Tony O'Hagan. "Elicitation". In: *Significance* 2.2 (2005), pp. 84–86; Bilal M Ayyub. "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities". In: *Institute for Water Resources, Alexandria, VA, USA* (2001).

[5]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[6]A. O'Hagan et al. *Uncertain Judgements: Eliciting Experts' Probabilities*. Statistics in Practice. Wiley, 2006. ISBN: 9780470033302.

[7]Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.

[8]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

- Expert opinion.
- Subjective judgment.
- Expert forecast.
- Best estimate.
- Educated guess.
- Expert knowledge.

"Stakeholder analyses," a part of the problem definition step in the systems design process and a process familiar to TRAC studies, typically involves SME elicitation Parnell, Driscoll, and Henderson and Anderson and Johnson.[9]

People seek input from experts in solving a variety of problems, including Meyer and Booker:[10]

- To provide estimates on new, rare, complex, or otherwise poorly understood phenomenon.
- To forecast future events.
- To integrate or interpret existing data.
- To learn an expert's problem-solving process or a group's decision making processes.
- To determine what is currently known, what is not known, and what is worth learning in a field of knowledge.

TRAC often conducts SME elicitations in order to obtain estimates on the relative utility of various courses of action measured against specific criteria for which data is not available Anderson and Johnson.[11]

[9]Gregery S. Parnell, Patrick J. Driscoll, and Dale L. Henderson, eds. *Decision Making in Systems Engineering and Management*. Second. Hoboken, NJ: John Wiley & Sons, Inc., 2011; Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

[10]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[11]Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

General guidelines

An elicitation should be designed to fit the experts rather than forcing the experts to adapt to the analyst's methods while generating data required to address specific study objectives and EEA. It should control, to the extent possible, for factors that can enter into the elicitation and influence the experts' responses (see section C). Good elicitation requires a degree of understanding in two fields: psychology and statistics Garthwaite, Kadane, and O'Hagan.[12]

Additionally, it is often useful to gather as much information as possible about the experts' problem solving processes, in addition to the experts' judgments Meyer and Booker.[13]

When conducting an elicitation, there are several common pitfalls to avoid Meyer and Booker:[14]

- Modeling bias.
- Overloading the experts.
- Inappropriate elicitation granularity.
- Unintended conditioning.
- Treating SME-elicited data as objective fact.

Interviewers, knowledge engineers, and analysts can introduce modeling bias into the elicitation and the analysis. This bias could be "training bias," i.e., misrepresenting the expert response data, or "tool bias," i.e., using familiar analytical tools as opposed to the appropriate ones. Another common pitfall is overloading the experts with information or questions, failing to account for the limits to the number of things the SMEs can simultaneously process and the inherent difficulty of the task the analysts are asking the SMEs to do; this mistake can introduce cognitive bias in the results. Eliciting data at the incorrect level of detail to support the analysis is another potential error in planning and conducting elicitations. Also, the "conditioning effect," whereby experts condition their responses on some events that are unintended in the study, is a common elicitation pitfall Meyer and Booker.[15] For example, an expert might make specific, unidentified assumptions, based on his basic branch or operational background, that affect his responses pertaining to a question about a given operational scenario.

A final pitfall, mentioned in the introduction to this chapter, is to fail to recognize the inherent subjectivity of SME-elicited data. Analysts should always point out to study stakeholders and sponsors the points in their analyses that rely on SME input. To the

[12]Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.
[13]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.
[14]Ibid.
[15]Ibid.

extent possible, the study team should conduct sensitivity analyses, verify the validity of the response data, and consider the risk involved if somehow the experts' beliefs, the elicited data, or the statistical inference were erroneous O'Hagan and Ayyub.[16]

Procedure

The general sequence of events for SME elicitation is Meyer and Booker:[17]

Planning

- Identify the data required to address the issues for analysis that can only be obtained from SMEs.
- Select the question areas and particular questions to elicit this data.
- Refine the questions.
- Select and motivate the experts.
- Select the method and components of the elicitation.
- Design and tailor the components of elicitation to fit the application.

Preparation

- Practice the elicitation and train the in-house personnel.

Execution

- Elicit and document expert judgments.

Elicitation Planning

In this section we discuss selecting and refining the elicitation questions, selecting the experts, and and selecting the components of the elicitation. Good planning is the most important step in analysis that relies on acquiring and employing SME-elicited data.

Identify the objectives and data requirements

The planning starts with the problem definition and measurement space work discussed in section **??**. Conducting a measurement space workshop is well-documented in TRAC literature Leath and Bauman et al.[18] Data requirements that require SME elicitation will

[16] O'Hagan, "Elicitation", op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[17] Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[18] Ms. Lynn Leath. *Study Directors' Guide; A Practical Handbook for Planning, Preparing, and Executing a Study.* Tech. rep. TRAC-F-TM-09-023. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2013; Michael F. Bauman et al. *Measurement Space Code of Best Practice (CoBP).* tech. rep. TRAC-H-TM-12-034. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2012.

be identified during the measurement space development for the study or analysis project. The measurement space also provides the team with the context and scenario that provide the background information needed to elicit the data. Finally, the measurement space enables the team to develop the methodology for employing the SME-elicited data in the analysis in order to arrive at meaningful results Bauman et al.[19]

Knowing how the elicited data is going to support the analysis guides the elicitation planning, preparation, and execution. Therefore, it is very important to identify not only *what* data is required, but also *how* the SME-elicited data is going to support the subsequent analyses and ultimately affect the outcome. These determinations form part of the study methodology and should be complete before attempting to plan the specific question areas for the elicitation event.

In summary, a good measurement space identification is essential to successful SME elicitation in support of TRAC studies and analyses. An incomplete measurement space identification all but guarantees problems in both acquiring the correct data through SME elicitation and applying it in the analysis in a meaningful way.

Elicitation questions

Any research or analysis team preparing to conduct a SME elicitation must carefully consider the items or questions they will present to the experts in order to elicit response data that meets the requirements of the study. For the purpose of this literature review, we use the term *question* to refer to any set of material presented to an expert with the intention of eliciting a specific data response. This definition includes the background information, operational scenario, context, and framing information, in addition to the question prompts, provided to the experts in eliciting their feedback.

In order to develop the questions, the team must first identify and define the project's objectives and determine how the elicitation supports those objectives. Precise, unambiguous objectives are key to a well-conceptualized and methodologically sound instrument Brenda M. Wenzel and Kathy L. Nau.[20] Following this problem definition effort, the team can decide on the general question areas and then, finally, identify the individual questions for the elicitation Meyer and Booker.[21]

The questions identified must support the purpose and goals of the project, providing data to answer the study issues and essential elements of analysis (EEA). The questions should be formed so that they either gather experts' answers or gather their problem-solving processes, according to study requirements. They must include the appropriate degree of technical complexity and be designed to elicit the desired quality, quantity and level of

[19]Idem, *Measurement Space Code of Best Practice (CoBP)*, op. cit.

[20]Ph.D. Brenda M. Wenzel and Ph.D. Kathy L. Nau. *Code of Best Practices for Survey Efforts.* Tech. rep. TRAC-W-TM-12-001. Martin Luther King Drive, White Sands Missile Range, NM 88002-5502: TRADOC Analysis Center, White Sands Missile Range, 2011.

[21]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

precision required. Finally, all questions should fit within the problem scope (i.e., constraints, limitations, and assumptions) and should not require an extreme effort by the experts or the analysts in obtaining the data Meyer and Booker.[22]

When feasible, the study team should involve one or more experts from the targeted domain in the elicitation question development or technical review Jenkinson.[23] There are different levels of this involvement. One option is to not have experts participate at all in elicitation planning and question development. Alternatively, experts can do most of the development, determining the question areas and questions prior to the elicitation. Another option is for the analysts to develop the question areas and then have the experts develop the actual questions. Finally, experts can simply provide feedback on the questions and question areas determined by the analysis team to ensure that SMEs from the domain understand the intent of the question and can provide data useful to the study team through that question. Early expert involvement comes with advantages and disadvantages. surveycobp recommend that any experts involved in technical review of a survey instrument not participate in the actual elicitation event Brenda M. Wenzel and Kathy L. Nau.[24] In planning an elicitation and developing elicitation questions, experts can provide input and feedback on Meyer and Booker:[25]

- Potential question areas.
- Approximate number of experts, based on the sample size required to inform the issues for analysis.
- Ideas on what would motivate external experts' participation.
- Question specification.
- Question definition.
- How much diversity of opinion might exist.
- Questions of proprietary rights.
- Feasibility of the elicitation plan (i.e., questions).

If it is not feasible or not practical to consult with experts in the field related to the study issues, the team can alternatively consult with an expert or knowledgeable peer on SME elicitation and survey techniques who, while not necessarily knowledgeable on topics of the study, can still provide general insights into conducting the elicitation and framing the questions Brenda M. Wenzel and Kathy L. Nau.[26] This is a useful step that should be

[22]Ibid.

[23]David Jenkinson. "The elicitation of probabilities-a review of the statistical literature". In: *Bayesian Elicitation of Experts' Probabilities (BEEP) working paper* (2005).

[24]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

[25]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[26]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

pursued if possible even if domain specific experts are available. The *Code of Best Practices for Survey Efforts* published by TRAC also provides an invaluable resource for developing surveys and survey questions Brenda M. Wenzel and Kathy L. Nau.[27]

The elicitation questions should be decomposed, to the extent possible, into small tasks that are distinct from each other Kynn.[28] This involves careful structuring, including the organization of the questions and the control of their presentation to the experts. Some objectives of questions structuring are Meyer and Booker:[29]

- Lessening the cognitive burden of solving the questions by making them easier to understand and process.
- Delimiting the questions so that the experts do not interpret them differently.
- Making the questions more acceptable to the experts because they encompass their views and use their terminology.

The steps involved in structuring questions include determining the types of information the experts will need (background, scenario, assumptions, definitions), determining the optimal order, and determining the roles of the project personnel and the experts during the elicitation Meyer and Booker and Ayyub.[30] Additional considerations include representation and phrasing of each question, with the goal of ensuring that each expert can easily read, digest, and understand the question material and response options Brenda M. Wenzel and Kathy L. Nau.[31] In other words, experts should easily understand exactly what they are being asked to do. It is generally not desirable to attempt to assess multiple issues with a single elicitation question or item, as the result often leads to data that provide a poor explanation of any of the issues the analysts are attempting to assess Brenda M. Wenzel and Kathy L. Nau.[32]

Analysts should also be careful not to ask questions that provide numbers or scenarios for the experts to use as an anchor or otherwise lead experts in the direction of a certain answer Kynn and Ayyub.[33] For example, a TRAC study might involve investigating an upgrade to a certain component. The study team might elicit failure probabilities from SME's by providing sufficient background information and context, and then asking the following questions:

---

[27]Ibid.

[28]Mary Kynn. "The 'heuristics and biases' bias in expert elicitation". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171.1 (2007), pp. 239–264.

[29]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[30]Idem, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[31]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

[32]Ibid.

[33]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

1) What is the probability of failure of the original component in this scenario?

2) What is the probability of failure of the improved component in this scenario?

Asking these questions as depicted presents two (and arguably more) problems. First, presenting these questions sequentially gives the expert the opportunity to anchor his answer to the question 2 on his answer to question 1. Second, describing the upgraded component as "improved" is a way of leading the expert toward a lower probability of failure. It would be better to separate these questions to prevent the anchoring and to use more neutral language in relating them. Some authors would argue that direct elicitation or probabilities, as shown in these two questions, is also problematic (see section C).

Often elicitation goals include estimating uncertain quantities. In these cases, experts might be more comfortable giving a range of values rather than having to provide a single value. In at least some cases, experts will provide a range even though they were asked to provide a single value Meyer and Booker.[34] Elicitations of probabilities, ranges, and distributions are well documented in the literature O'Hagan; Shephard and Kirkwood; Chesley; Savage; Jenkinson; Garthwaite, Kadane, and O'Hagan; and DeGroot.[35] In these cases, questions must be carefully structured to best obtain the data required by the analysts, as humans in general are prone to many sources of error and bias in these types of judgments (see section C) Kynn.[36] Analysts should ultimately be equally concerned with *what* they ask and *how* they ask it Kynn.[37] Section C expands on this topic.

As a final note on developing questions we again turn to the TRAC literature. The *Code of Best Practices for Survey Efforts*, which provides detailed considerations for analysts, especially those with little survey experience, can assist in developing surveys and survey questions. Although many parallel ideas appear in the *Code of Best Practices*, we do not intend to reproduce the *Code of Best Practices for Survey Efforts* here. TRAC study teams that choose to employ surveys for the purpose of gathering SME judgment or knowledge should thoroughly review the *Code of Best Practices* Brenda M. Wenzel and Kathy L. Nau.[38] Another TRAC resource is the *Capability Gap Assessment* technical report, which provides detailed guidelines on planning and executing elicitations aimed at identifying and quantifying capability gaps Rubemeyer et al.[39] This reference includes

[34]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[35]O'Hagan, "Elicitation", op. cit.; Glenn G Shephard and Craig W Kirkwood. "Managing the judgmental probability elicitation process: a case study of analyst/manager interaction". In: *Engineering Management, IEEE Transactions on* 41.4 (1994), pp. 414–425; GR Chesley. "Elicitation of subjective probabilities: a review". In: *The Accounting Review* 50.2 (1975), pp. 325–337; Leonard J Savage. "Elicitation of personal probabilities and expectations". In: *Journal of the American Statistical Association* 66.336 (1971), pp. 783–801; Jenkinson, "The elicitation of probabilities-a review of the statistical literature", op. cit.; Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.; Morris H DeGroot. "Reaching a consensus". In: *Journal of the American Statistical Association* 69.345 (1974), pp. 118–121.

[36]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[37]Ibid.

[38]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

[39]Mr. Adam Rubemeyer et al. *Capability Gap Assessment; Blending Warfighter Experience with Science.*

measurement space identification, expert selection, and question, response scale, and survey development. It also gives some techniques for displaying and otherwise employing the resulting data in the overall study.

Choosing experts

The role of experts.. During an elicitation, the interviewer or analyst asks the experts to solve problems. The difficulty of this task is often underestimated by the interviewer. Essentially, this problem-solving breaks down into four major tasks Meyer and Booker:[40]

- Understand the questions and the context.
- Remember relevant information.
- Make judgments.
- Formulate and report the answer.

The difficulty involved opens many opportunities for misinterpretation, confusion, and bias to enter this process. Experts will often disagree. These disagreements follow from interpreting the question differently, applying a different problem-solving method, using different data, or using the same data but interpreting it differently. Another reason for disagreement is the lack of clearly defined standards pertaining to the subject of interest, which is often also the reason for conducting the elicitation Meyer and Booker.[41] It has been shown that, while their calibration can be improved, experts cannot be perfectly, externally calibrated and often fail to adequately adjust their judgment based on new information or are overconfident when faced with uncertainty Meyer and Booker and Kynn.[42] However, it is important to note that consensus is not required in most cases and that we are typically asking them to express their opinion based on their individual experience. A lack of consensus from SMEs, especially on controversial topics should not be surprising. The goal of an elicitation is to obtain data to address a study issue not to force consensus if it does not exist.

Because of the human nature of experts in SME elicitation discussed in this section, data gathered in SME elicitation events is not reproducible. This limitation constitutes a drawback to using SME elicitation to obtain data Meyer and Booker.[43] A well constructed elicitation will seek to control for noise in the response by minimizing these extraneous

Tech. rep. TRAC-F-TR-13-022. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2013.

[40]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[41]Ibid.

[42]Idem, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[43]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

sources of error so that variance in the response is based on true differences of opinion related to the study issues Kynn.[44]

Selecting experts. Statistics provide the mathematical rigor for analysis of data, including expert data. Therefore, when conducting SME elicitation it is important for analysts to consider the statistical nature of the analysis from the outset. One of the first things the analysts must do is determine the *population* of interest, for which the expert data will provide a sample. The issues for analysis should assist in identifying the population; knowing the characteristics of the population should contribute to answering the issues for analysis.

The statistical population can be an abstract, infinite population or a physical, finite population. In some cases, the statistical population can be the pool of experts. If the statistical population, driven by the data requirements, is not the expert population, the team must determine who is qualified to provide accurate sample data from the population Brenda M. Wenzel and Kathy L. Nau.[45] These people are the *experts*; they are the individuals whose domain knowledge, once elicited, will inform our analysis Garthwaite, Kadane, and O'Hagan and Jenkinson.[46] They must be among those best qualified to provide accurate estimates of the real world quantities of interest. Sometimes it might be best to elicit data from multiple groups of experts in order to capture judgments from different perspectives Anderson and Johnson.[47] literaturereview provides some desirable traits in identifying potential experts for elicitation Jenkinson.[48] They are:

- Tangible evidence of expertise.
- Reputation.
- Availability and willingness to participate.
- Understanding of the general problem area.
- Impartiality.
- Lack of economic or personal stake in the potential findings.

Once the pool of experts is identified, the team must select a subset of them, a *sample*, to participate in the elicitation. The analysts *must* consider the modeling assumptions that support the methods they plan to use. Statistical methods generally assume a *random*

[44]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[45]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

[46]Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.; Jenkinson, "The elicitation of probabilities-a review of the statistical literature", op. cit.

[47]Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

[48]Jenkinson, "The elicitation of probabilities-a review of the statistical literature", op. cit.

sample, although there are some variations to random sampling. For example, TRAC studies often involve sampling a minimum number of experts from each of several ranks. This sampling method is referred to as stratified sampling and is well developed in the literature. However, stratified sampling still assumes samples within each strata are random Rice.[49]

Another method of sampling is to attempt to select experts in order to capture the range of "typical" views. Sometimes TRAC analysts imply this method by assuming a "representative" sample of from the population. While appealing, there is no mathematical way to guarantee estimates coming from these samples are free from error or unbiased Rice.[50] It is important to note, however, a "representative" sample is an implied assumption when using bootstrapping methods (see section C). In general, we recommend analysts assume random sampling in TRAC studies and projects because this assumption is inherent in almost all statistical methods. To the extent possible, TRAC analysts should choose experts to participate in SME elicitation events in a manner that allows them to assume that the sample has been randomly generated.

For example, both the TRAC FD/FM study and the TRAC E-MIB workshop involved elicitation events in which experts were selected to represent different demographic aspects of the population, but the subsequent analyses assumed random samples Dabkowski et al. and Deavens et al.[51] These two analysis efforts are covered in more detail in chapters **??** and **??**. Each of these studies explicitly stated and justified the assumption of a random sample, but each could have benefited from more detailed analyses to examine the potential effects of the sampling strata.

The number of experts. In any statistical analysis, a larger sample enables more powerful inference. However, larger samples cost more. In deciding on the number of experts required for an elicitation, the analysts must consider the objectives of the study, the issues for analysis, the data the SMEs will provide, the size of the pool of experts, the analytic tools and models they will use, and total cost to the Army in conducting the elicitation. Some statistical analyses require samples that are "sufficiently large" in order to meet certain, theoretical conditions. Many references cite a sample size of $n \geq 30$ in order to apply the normal approximation to the distribution of the sample mean as specified by the central limit theorem Rice and Rubemeyer et al.[52] Analysts should be wary of using this number as a benchmark because it unnecessarily constrains thinking to applications of the central limit theorem and can lead them to accepting input from people not qualified

[49] J.A. Rice. *Mathematical Statistics And Data Analysis*. Third. Advanced series. Brooks/Cole CENGAGE Learning, 2007. ISBN: 9780534399429.

[50] Ibid.

[51] MAJ Matther Dabkowski et al. *Force Design/Force Mix: Building the Best Army Possible with Reduced End-Strength*. Tech. rep. TRAC-F-TR-11-020. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2011; MAJ Tom Deavens et al. *Support for the Expeditionary Military Intelligence Brigade Commanders' Assessment Workshop*. Tech. rep. TRAC-M-TR-13-029. 700 Dyer Road, Monterey, CA 93943: TRADOC Analysis Center, Monterey, 2013.

[52] Rice, *Mathematical Statistics And Data Analysis*, op. cit.; Rubemeyer et al., *Capability Gap Assessment; Blending Warfighter Experience with Science*, op. cit.

as experts. It is generally better to get a few data points from well-qualified SMEs than to get many data points from less qualified individuals Anderson and Johnson.[53]

On the other hand, if the statistical population is finite and small, the sample size required to make inference about the population might also be small. This condition might be true if the statistical population is the pool of experts, such as in the E-MIB Workshop analysis project (see chapter **??**) Deavens et al.[54]

The study team should be able to discern an approximate number of experts, n, required by observing the effect of n on the power of the statistical methods they plan to use. For each possible value of n, they should consider the question: "will we be able to answer the issues for analysis with a data set of this size?" Once the team has determined how many experts they need, they should rehearse their analysis with generated data and ensure the statistical results are precise enough to meet the objectives of the study.

Elicitation components and methods

An elicitation consists of the following components Meyer and Booker:[55]

- Elicitation Method. The method provides the setting in which the elicitation of expert judgment takes place. This can an individual interview, an interactive group, a Delphi event, a web-based survey, a meeting, or something else Parnell, Driscoll, and Henderson and Meyer and Booker.[56] We provide a brief discussion of choosing the elicitation method below.
- Mode of communication. This is the means by which the data gatherer and the expert communicate. It can be face-to-face, by mail, e-mail, online system, telephone, or some other form of communication. More and more, TRAC analyses are turning to online or other "distributed" methods of elicitation communication.
- Elicitation technique. This is the means by which experts are led to describe aspects of their knowledge. It can be a verbal report, responses to probing questions, or responses to survey items. A *verbal report* is simply an expert's response, including an explanation. For example, an analyst might ask an expert to describe the best features of the new ground combat vehicle. The expert then provides his response. This elicitation technique can be expanded to include written reporting as well. *Probing questions* are questioning techniques that, based on and expert's responses, are designed to make the expert re-think his or her response or provide a more detailed explanation. For

[53]Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

[54]Deavens et al., *Support for the Expeditionary Military Intelligence Brigade Commanders' Assessment Workshop*, op. cit.

[55]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[56]Parnell, Driscoll, and Henderson, *Decision Making in Systems Engineering and Management*, op. cit.; Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

example, after listening to an expert describe the best features on the new ground combat vehicle, an interviewer might ask why the expert considered turning radius but not maximum incline. Finally, *responses to survey items* are responses to simple questions, provided orally, in writing, or electronically. Typically, in responding to survey items, experts are not expected to elaborate beyond providing responses in the format requested in the survey.

- Response mode. This is the form by which experts encode their judgment. There are many response modes. Some examples are estimates of quantity, probability, probability distribution, ranks, pairwise comparisons, ranges, percentiles, and standard deviations. When using surveys, the TRAC *Code of Best Practices for Survey Efforts* provides an array of standardized question prompts and response scales Brenda M. Wenzel and Kathy L. Nau.[57] Likert scale responses are common within TRAC elicitations Brenda M. Wenzel and Kathy L. Nau; Wolfe; Rubemeyer et al.; and Anderson and Johnson.[58] gapassessment also provide standardized response scales for gap assessments based on outcome probability and mission impact severity as defined in Army FM 5-19 (Composite Risk Management) Rubemeyer et al. and Headquarters, Department of the Army.[59] wolferisk provides similar scales for assessing probability and severity, with the objective of eliciting a triangular distribution to model each (see section C) Wolfe.[60] madm provide a value incremental approach that is useful when eliciting a value function for use in multi-attribute decision making Anderson and Johnson.[61]

- Aggregation scheme. This is the process employed to obtain a single datum from multiple, different data points. This can mean behavioral aggregations, i.e., forcing the experts to reach a consensus, or mathematical aggregations, e.g., finding the mean response. wolferisk uses the mean response from the experts in identifying a minimum standard of performance in her example elicitation scenario Wolfe.[62]

- Documentation. This is how the elicitation data is recorded. It can be documented answers only, problem solving steps and answers, summary documentation, verbatim documentation, or planned, structured documentation. Final documentation can include the questions in their final form, the identities of the experts, the methods used

[57]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

[58]Idem, *Code of Best Practices for Survey Efforts*, op. cit.; Ms. Michele Wolfe. *Operational Risk Analysis, A New Approach*. Tech. rep. TRAC-F-TR-13-026. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2013; Rubemeyer et al., *Capability Gap Assessment; Blending Warfighter Experience with Science*, op. cit.; Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

[59]Rubemeyer et al., *Capability Gap Assessment; Blending Warfighter Experience with Science*, op. cit.; Headquarters, Department of the Army. *FM 5-19: Composite Risk Management*. Government Printing Office.

[60]Wolfe, *Operational Risk Analysis, A New Approach*, op. cit.

[61]Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

[62]Wolfe, *Operational Risk Analysis, A New Approach*, op. cit.

to obtain the data, and the experts' responses. Procedures for documentation include media (audio or video) recordings, human recorder note-taking, and submission of responses by experts.

Note that not every elicitation requires all components, and that each method has advantages and disadvantages, making it more appropriate for some elicitations than others.

There are many factors to consider when choosing an elicitation method. Analysts must decide what type of information they require to support their analysis, and what form that data will take in the experts' answers. They must consider the number of experts that will be available to participate in the elicitation (as well as the minimum number required), whether they desire interaction among the experts, and the amount of time the experts will need in order to provide the judgments and responses required. Finally, the analysts must consider the project scope (i.e., constraints, limitations, and assumptions), the methodology, and the level of difficulty involved in preparing the problems or questions for the experts Meyer and Booker.[63]

In some cases the study team might want to aggregate the experts' responses for use within the study. This aggregation can be accomplished during the elicitation by encouraging the experts to interact and eventually agree upon a single response (*behavioral* aggregation), or after the elicitation using statistical methods (*mathematical* aggregation) Meyer and Booker; O'Leary et al.; and O'Hagan.[64] One popular method of encouraging experts to come to an agreement on a particular topic is to use the "Delphi" elicitation method, which seeks to minimize the biasing effects of dominant individuals, irrelevant or misleading information, and group pressure Parnell, Driscoll, and Henderson and Anderson and Johnson.[65] This method, developed by RAND for the U.S. Air Force in the 1950's, involves presenting the questions to the experts and recording their initial feedback anonymously Ayyub.[66] This feedback is quickly analyzed and the results displayed to the expert group. Experts are encouraged to discuss these results, challenging each other's assumptions and problem solving processes while defending their own positions. Meanwhile, analysts tailor the next round of questions based on the results from the previous round, and the process iterates. consensus provides a simple, theoretically sound way of aggregating response distributions into a single probability distribution in a Delphi elicitation setting DeGroot.[67]

[63]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[64]Idem, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.; Rebecca A O'Leary et al. "Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby Petrogale penicillata". In: *Environmetrics* 20.4 (2009), pp. 379–398; O'Hagan, "Elicitation", op. cit.

[65]Parnell, Driscoll, and Henderson, *Decision Making in Systems Engineering and Management*, op. cit.; Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

[66]Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[67]DeGroot, "Reaching a consensus", op. cit.

The second means of aggregating expert responses is through data analysis and statistical methods applied to the response data after the elicitation. These methods are discussed in section C.

Once an analyst has an idea of what type of information he or she needs, along with the project's CLA, the analyst can begin considering potential elicitation methods. In choosing a method, the analyst should consider Meyer and Booker:[68]

- The degree of interaction among the participants.
- The amount of structure imposed by the moderator or interviewer.
- The number of meetings.
- The time allotted for structuring the problem and eliciting expert judgment.
- Who structures the problem and elicits feedback, the analysts or the experts.
- The response mode.
- Whether the expert's reasoning is requested or not.
- The level of detail elicited.
- Whether the expert judgment undergoes some translation in a model and is returned to the experts for the next step.
- The elicitation media; whether some or all of the elicitation is conducted in person or by mail, e-mail, online systems, or telephone.

Elicitation Preparation

Preparing for the elicitation includes rehearsals, pilot-testing, and training. Rehearsals should include introducing the experts to the elicitation process, elicitation and documentation procedures, mathematical aggregation of answers (perhaps using notional data), and analysis using the models selected for the research Meyer and Booker.[69] Pilot-testing serves several purposes. It verifies the experts' understanding of the problem or question, their use of the response mode, their understanding of the elicitation procedures, and their ability to accommodate any required documentation formats. It can also be used to validate the survey (see discussion of Cronbach's alpha, section C). If experts are not available, the survey or elicitation instrument can still be pilot-tested within TRAC in order to garner feedback on clarity Rubemeyer et al.[70] Training is required when the people that execute the elicitation are not the same people that planned it, when

[68]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.
[69]Ibid.
[70]Rubemeyer et al., *Capability Gap Assessment; Blending Warfighter Experience with Science*, op. cit.

elicitors feel uncomfortable with their roles, or when multiple people will perform the same technical tasks Meyer and Booker.[71]

Other preparations include obtaining a list of the participating experts and arranging for proper protocol. The study team should communicate in advance with the experts, providing appropriate background materials, guidance and relevant information concerning the elicitation process, and a note thanking them for their participation Meyer and Booker and Ayyub.[72] The study team also must secure the facilities, supplies, and technical support required to carry out the elicitation. If the elicitation is conducted in a distributed fashion, the team must still prepare to provide timely technical support and answers to questions as they arise Brenda M. Wenzel and Kathy L. Nau.[73]

In some cases, the SME elicitation might qualify as human subject research and require documentation and approval as such. Department of Defense (DoD) regulation pertaining to human subject research and Institutional Review Board (IRB) requirements is provided in ??32 CFR 219 *Title 32–National Defense; Code of Federal Regulations 219, Protection of Human Subjects (32 CFR 219).*[74] This reference specifies what circumstances require an IRB, and provides exceptions that often apply to TRAC SME elicitation events. TRAC has implemented a human research protection plan that provides additional guidance on research involving human subjects Jebo, Krondak, and McGrath.[75] Local organizations and installations might also generate their own human subject research approval requirements. For example, the Naval Postgraduate School (NPS) reviews all human subject research according to the following guidance:

> The NPS IRB has jurisdiction over all research involving human subjects. A human subject is a living individual about whom an investigator conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information. No human subject research in any form can take place without proper review and approval by the NPS IRB and NPS President. The NPS IRB is authorized to review, recommend approval to the NPS President, require modifications in, or withhold approval or suspend approval of research involving human participants. For a list of specific human research protection instructions and regulations reference NAVPGSCOL Instruction 3900.4A (authentication required) *Institutional Review Board for the Protection of Human Subjects.*[76]

[71] Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[72] Idem, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[73] Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

[74] *Title 32–National Defense; Code of Federal Regulations 219, Protection of Human Subjects (32 CFR 219).* Government Printing Office. 2013. URL: `http://www.ecfr.gov/cgi-bin/text-idx?c=ecfr&tpl=/ecfrbrowse/Title32/32cfr219_main_02.tpl`.

[75] Dr. Jennifer Jebo, Ms. Sara Krondak, and Ms. Amy McGrath. *TRADOC Analysis Center Human Research Protection Program Plan.* Tech. rep. TRAC-L-TM-13-028. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2013.

[76] *Institutional Review Board for the Protection of Human Subjects.* Naval Postgraduate School website.

Whether or not an IRB is required, analysts conducting any SME elicitation must ensure that they respect respondents and consider any risk they incur in their participation. For example, personal questions about certain experiences can cause discomfort in some people. Prior to participation, potential respondents must be briefed on the details of the elicitation and the use of the resulting data so that they can make an informed decision on whether to participate Brenda M. Wenzel and Kathy L. Nau.[77] Respondents must be aware of whether any response data will be attributional, and analysts must consider potential negative impacts of such attribution. They must also understand that there will be no adverse effects if they choose not to participate. At a minimum, potential respondents should be briefed on the following, as they apply Brenda M. Wenzel and Kathy L. Nau:[78]

- An explanation of the purpose(s) of the study.
- The approximate amount of time the survey will take.
- A description of what the respondents will be asked to do.
- A description of any foreseeable risks or discomforts.
- A description of any benefits to the respondents or others.
- A statement of the voluntary nature of the participation.
- A statement describing how confidentiality of responses will be attained.

Naturally, analysts must always adhere to organization and institutional policies and regulations, which help to ensure that all human subject research meets ethical standards and protects the privacy and rights of the people involved.

Executing the Elicitation

Executing the elicitation should go, to the extent possible, as rehearsed. The event should begin with introductions and provide the experts with a clear purpose, format, and expectations. Each member of the elicitation team should be introduced and his or her role defined. The elicitation team must give the experts opportunities to ask questions throughout the process. The elicitation team should generally provide any materials and support needed by the experts in order to provide their response data or otherwise participate in the elicitation Meyer and Booker.[79] The team should carefully record the elicitation in accordance with the plan and the rehearsals.

During execution, the elicitation team should also be aware of indicators of bias, which we discuss in the next section.

2013. URL: `http://www.nps.edu/research/IRB.htm`.

[77]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

[78]Ibid.

[79]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

Elicitation Bias

Elicitation Bias is a difference between the response data collected during an elicitation event and reality. Much of the research on SME elicitation has been concerned with minimizing this bias, which can come from a variety of sources Bedford, Quigley, and Walls.[80] In general, we encounter two types of elicitation bias: *motivational* bias and *cognitive* bias Meyer and Booker.[81] Motivational bias occurs when there is a difference between the expert's judgment and the expert's response. This difference can result from peer influences or from a desire to impress the interviewer. Cognitive bias is when there is a difference between the real world and the expert's judgment. Inaccurate, incomplete, or incorrect mental models of the situation can produce a cognitive bias, which is often difficult to detect because often in an elicitation there is not objective, external data for comparison against expert responses. One indicator of cognitive bias is a lack of consistency; if an expert gives responses that contradict each other, it might be an indication of a cognitive bias.

Specific causes of motivational bias

practicalguide provide some causes of motivational bias Meyer and Booker:[82]

- Group think. This occurs when one or more individuals dominate the discussion, with the remainder generally agreeing.
- Wishful thinking. Wishful thinking occurs when the experts perceive they have something to gain by responding a certain way. The "something" could be intangible, such as credibility, or tangible, such as money.

Specific causes of cognitive bias

Some specific causes of cognitive bias stem from methods of estimation that are well documented in the literature. They include:

- Availability bias. In availability bias, experts base their input (typically probability or frequency of an event's occurrence) on how easily they can recall relevant events Garthwaite, Kadane, and O'Hagan; Kadane and Wolfson; Kynn; and Ayyub.[83] For

[80] Tim Bedford, John Quigley, and Lesley Walls. "Expert elicitation for reliable system design". In: *Statistical Science* (2006), pp. 428–450.

[81] Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[82] Ibid.

[83] Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.; Joseph Kadane and Lara J Wolfson. "Experiences in elicitation". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 3–19; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

example, a TRAC analyst attempting to elicit information on risk might ask an expert about the likelihood of a certain catastrophic event occurring under certain conditions. Suppose in reality the likelihood of occurrence is very low. If the expert recently witnessed or was otherwise involved in a catastrophic event similar to the one described in the question, he or she is likely to inflate that probability, even if the conditions or the event itself differ somewhat from the one described in the question. The event comes to the expert's mind easily because it happened recently and because it had catastrophic consequences.

- Representative bias. This bias is similar to availability bias, resulting from the tendency of people to relate the probabilities of two events based on some similarity between them, however irrelevant to probability Kynn and Ayyub.[84] For example, an electronic jamming device might disrupt a cell phone signal at certain distance 80% of the time. An expert, knowing this quantity, might report the same probability for disrupting a hand-held radio at the same distance.

- Judgment by anchoring and adjusting. People widely use this method as a heuristic for estimating unknown quantities, by starting with a familiar, similar scenario (the *anchor*) and then updating the quantity of interest to adjust for the unfamiliar circumstances. Most of the time, these adjustments are too small, and the resulting quantity is biased toward the anchor value Garthwaite, Kadane, and O'Hagan; Kadane and Wolfson; Kynn; and Ayyub.[85]

- Overconfidence. This bias often occurs by underestimating the weight of the tails in a probability distribution. In other words, the expert fails to correctly assess his or her own degree of uncertainty. For example, elicitation events often involve asking experts for a confidence interval, or range or plausible values for an unknown quantity, e.g., the average number of hours a proposed vehicle can operate before needing servicing. Suppose an elicitation asked experts for a 95% confidence interval for this quantity, i.e., a range for which they would wager up to \$0.95 for a \$1.00 reward for containing the true average. Multiple studies have shown that the interval experts provide tends to contain the true value with a much lower probability O'Hagan; Garthwaite, Kadane, and O'Hagan; Kadane and Wolfson; and Kynn.[86]

- Conjunction fallacy. Conjunction fallacy stems from a misunderstanding of the additive nature of probability applying only to disjoint events. Experts will provide probabilities that do not follow the axioms of probability for a variety of reasons, including the

[84]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[85]Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.; Kadane and Wolfson, "Experiences in elicitation", op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[86]O'Hagan, "Elicitation", op. cit.; Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.; Kadane and Wolfson, "Experiences in elicitation", op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

failure to account for probability assigned to the intersection of two events Kadane and Wolfson and Kynn.[87]

- Hindsight bias. This bias occurs when people are asked to give an a priori probability of an event after it has occurred. For example, a vehicle commander might have experience a track failure on his vehicle during a mission. During a subsequent investigation, the vehicle commander is asked what he thinks the probability of having the track failure during the mission would have been prior to conducting the mission Kadane and Wolfson and Garthwaite, Kadane, and O'Hagan.[88]

- Conservativism. This bias is the Bayesian analogue to anchoring and adjusting. Humans, when presented with new data and information, tend to remained bias toward a priori beliefs in spite of the new evidence Meyer and Booker.[89] heuristics gives two potential reasons for this bias: (1) misaggregation, in which the subjects (or experts) understand the usefulness of each new data point, but fail to discern the aggregate effect of the new data according to Bayes' theorem, and (2) misperception, in which the subjects fail to understand the applicability of the new data points Kynn.[90]

- Conditioning. Conditioning occurs when the expert applies conditions to the scenario or question that were not intended in the elicitation Meyer and Booker.[91]

practicalguide detail a program for handling bias Meyer and Booker.[92] This approach includes anticipating bias and taking action in the elicitation planning phase to make the event less prone to bias; familiarizing experts with potential sources of bias and, as necessary, with basic probability constructs and interpretations that apply to the elicitation; monitoring the elicitation for signs of bias and taking action in real time to counter it; and analyzing the data following the elicitation for indicators of bias.

Preventing bias in planning

Careful structuring of the questions, as mentioned in section C, is one important way to prevent bias in the data coming from SME elicitation Ayyub.[93] The TRAC *Code of Best Practices for Survey Efforts* provides many useful tips on designing and structuring surveys Brenda M. Wenzel and Kathy L. Nau.[94] The background information, operational scenarios, context, framing, and question prompts should fall in the experts' area of

[87]Kadane and Wolfson, "Experiences in elicitation", op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[88]Kadane and Wolfson, "Experiences in elicitation", op. cit.; Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.

[89]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[90]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[91]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[92]Ibid.

[93]Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[94]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.

expertise, be specifically worded, and be measured in units familiar to the experts Kynn.[95] Questions should not be leading or provide anchors to the experts (see section C) Meyer and Booker, Kynn, and Ayyub.[96]

elicitation provides a detailed interview technique aimed at training the expert to understand and provide the desired quantities. The author also makes suggestions to keep the precise language of probability out of the questions, e.g., ask for a "range of plausible values" instead of a "95% confidence interval" O'Hagan.[97] The analysts can also design the questions so that they allow for checks for internal consistency, either by asking several questions designed to elicit the same expert judgment data point Garthwaite, Kadane, and O'Hagan.[98] If inconsistencies are identified during the elicitation event, experts can be given a chance to explain and correct them Garthwaite, Kadane, and O'Hagan.[99]

Other references distinguish between *direct* and *indirect* methods of elicitation Chesley and Ayyub.[100] Direct methods assume that an expert can and will provide the truth when asked for a specific probability, range, quartile, histogram, or distribution function. This method has the advantage of simplicity. However, research has shown that accurate subjective probabilities are not obtainable by simply asking a person to provide a probability Bedford, Quigley, and Walls; Jenkinson; Kynn; and Ayyub.[101] One direct method that appears often in the literature is to ask experts for an odds ratio, instead of a probability Chesley,[102] assuming that odds ratios are more intuitive than probabilities. Direct questioning might be preferred, however, when eliciting information that is less abstract, such as ranking preferences Anderson and Johnson.[103]

Indirect methods, on the other hand, require more work on the part of the study team. They involve establishing betting schemes, relative frequency analogies, tradeoff scenarios, or other practical applications of probability in order to avoid the biases stemming from probability misunderstandings Chesley and Abdellaoui.[104] These methods enable the

[95]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[96]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[97]O'Hagan, "Elicitation", op. cit.

[98]Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.

[99]Ibid.

[100]Chesley, "Elicitation of subjective probabilities: a review", op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[101]Bedford, Quigley, and Walls, "Expert elicitation for reliable system design", op. cit.; Jenkinson, "The elicitation of probabilities-a review of the statistical literature", op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.; Ayyub, "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities", op. cit.

[102]Chesley, "Elicitation of subjective probabilities: a review", op. cit.

[103]Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

[104]Chesley, "Elicitation of subjective probabilities: a review", op. cit.; Mohammed Abdellaoui. "Parameter-free elicitation of utility and probability weighting functions". In: *Management Science* 46.11 (2000), pp. 1497–1512.

analysts to discern subjective probabilities based on the experts' decisions in scenarios provided. literaturereview provides many methods for constructing indirect and direct-indirect hybrid questions aimed at obtaining uncertain quantities Jenkinson.[105]

comparison address methods of minimizing bias when eliciting weights for application in a multi-attribute decision problem. The authors conclude that direct rating (rating each attribute independently on a common scale) is preferable to point allocation (distributing a fixed number of points among the attributes), and concludes that the "Max100" weighting method, in which the best attribute is assigned 100 points and the remaining attributes rated against it on a 0-99 scale, performs the best with respect to subject's internal consistency. This performance is attributed to the way that Max100, unlike direct rating, forces experts to fix the value of their most important attribute from the beginning. The subjects in the study also indicated they preferred Max100 and direct rating to another method, Min10, which performed poorly Bottomley and Doyle.[106]

Preventing bias in preparation

Rehearsals and pilot testing of survey instruments can provide useful insights into opportunities for bias to enter into an elicitation event. During these activities, analysts should watch for indicators of bias including misunderstandings among participants or ambiguity in the response mode. The results of the pilot tests and rehearsals can cause a study team to go back and refine questions and instructions on the elicitation procedures.

Training the elicitation personnel, including the experts, can also serve to prevent bias. The study team can prepare the experts for the elicitation event by conducting several exercises to help calibrate their probability scales against overconfidence, conjunction fallacies, conservativism, and anchoring and adjusting bias. Several methods aimed at preparing experts and improving expert calibration include Kynn:[107]

- Providing calibration training questions in advance of the elicitation. These must be directly related to the actual elicitation questions if they are to be of any value.

- Providing scoring rules (i.e., indirect methods of questioning) that help calibrate or explain probabilities. These must be transparent to the expert.

- Providing experts with a brief review of probability Jenkinson.[108]

[105]Jenkinson, "The elicitation of probabilities-a review of the statistical literature", op. cit.
[106]Paul A Bottomley and John R Doyle. "A comparison of three weight elicitation methods: good, better, and best". In: *Omega* 29.6 (2001), pp. 553–560.
[107]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.
[108]Jenkinson, "The elicitation of probabilities-a review of the statistical literature", op. cit.

Preventing bias during execution

There are several indicators of bias that the elicitation team might observe during an elicitation event. Very little discussion or dissenting opinions is an indication of group think. In this case, the facilitator could encourage more discussion by presenting opposing points of view to the group, effectively introducing an anchoring counter-bias Meyer and Booker and Kynn.[109] Alternatively, the analysts can adjust the elicitation to use anonymous response and, according to the Delphi method, present an analysis of the results to the experts to counterbalance group think Parnell, Driscoll, and Henderson.[110]

If experts are providing feedback quickly without due consideration, it might be an indicator of wishful thinking. Having experts explain their reasoning in detail might help in identifying and mitigating wishful thinking Meyer and Booker.[111] elicitation recommends asking experts up front about any financial or personal interests in the outcome of the elicitation in order to remind them of the need for unbiased data and to document any conflict of interest O'Hagan.[112] These potential conflicts of interest can infiltrate TRAC SME elicitations, because experts are often from the user community. The users might have the impression that their responses might lead to results that ultimately impact them in some way, either positive or negative. TRAC analysts conducting SME-elicitations should be mindful of this potential source of bias and plan to control for it to the extent possible, even if simply by making the experts aware of it or having them suggest methods for mitigating it during the elicitation.

Inconsistency in the experts' responses, sometimes called a lack of coherence, refers to situations in which experts contradict themselves, and is a potential indicator of a conditioning bias. If a team encounters a lack of self-consistency during an elicitation event, the team can address the inconsistency with the expert in order to find out why the expert's answer changed. Another alternative is to have experts monitor their own consistency as one method for maintaining self-consistency. The lack of coherence could also occur because the expert changed his or her mind during the elicitation, which is interesting information worth further exploration Kynn.[113] However, it is more likely that the expert is conditioning and will only require some clarification, or that the expert is fatigued. If it is fatigue, the team should review the demands the elicitation event is placing on the experts. In either case, expert inconsistency is an indicator of inadequate elicitation planning, as obtaining coherent responses is largely dependent on context, framing, and the specific details of an elicitation Kynn.[114]

Availability bias might become apparent during an elicitation if an expert bases his or her judgment on one or two specific things pertaining to a more general question. This

[109]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.
[110]Parnell, Driscoll, and Henderson, *Decision Making in Systems Engineering and Management*, op. cit.
[111]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.
[112]O'Hagan, "Elicitation", op. cit.
[113]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.
[114]Ibid.

behavior would indicate the expert has a memory that is quickly coming to mind and influencing his or her decision-making. The facilitator, interviewer, or discussion with other experts can help counter availability bias by investigating the source of the expert's memory association, or by brainstorming facets of the problem that are different than those emphasized by the expert Meyer and Booker and Kynn.[115] These same techniques can also help mitigate representation bias.

Anchoring or conservativism biases are apparent when an expert does not update his or her response based on additional data Kynn.[116] Techniques to counter this bias are similar to those used to counter the availability bias Meyer and Booker.[117]

Methods to detect overconfidence involve asking experts a large number of questions for which objective data is readily available, and then comparing the results for consistency Kynn.[118] While this might be impractical, it might be worthwhile to take action to counter overconfidence even in the absence of evidence. One method is to have experts decompose uncertainty into different sources. The aggregation of these different sources of uncertainty can help experts quantify the total uncertainty in their estimates Meyer and Booker.[119] Exercises to improve calibration can also be useful in mitigating overconfidence, but analysts should use caution as documentation of expert calibration has produced mixed results Kynn, Meyer and Booker, and Kadane and Wolfson.[120]

Calibration exercises can serve several purposes within TRAC elicitations. First, it can serve to verify that expert responses are consistent with measurable, external data. Second, it can help establish a baseline for variance withing the group of experts. Third, it can serve as expert screening criteria. For example, an analysis team might assemble a group of field artillery officers in order to help evaluate a new projectile design. The team is attempting to collect quantitative data on the accuracy of the prospective projectile. In order to calibrate the officers, the analysts provide them with the characteristics of a round that is already in production, with known, well-documented trajectory data. Analysts elicit accuracy data from the officers and compare their estimates to the known quantities. If the resulting data has a large variance, the analysts should expect a large variance when eliciting the unknown quantities as well. If one or more experts provide very bad estimates of the known quantity, the analysts might want to consider excluding their estimates for the unknown quantity or quantities based on this benchmark.

Finally, the conjunction fallacy bias becomes apparent when expert responses violate the

[115]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[116]Idem, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[117]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[118]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[119]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[120]Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.; Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.; Kadane and Wolfson, "Experiences in elicitation", op. cit.

axioms of probability Kadane and Wolfson.[121] If analysts observe such an obvious indicator of incoherence during an elicitation, they should begin breaking down the questions into more sub-parts or asking the experts to explain their incoherent responses in order to reveal flaws in the experts' logic and assist them in correcting their responses.

Section Summary

In this section we reviewed the available literature on gathering data from SMEs in order to support analytical efforts. However, gathering data is only half of the elicitation problem. Analysts must apply, interpret, or analyze the response data provided by the SMEs in order to answer the issues for analysis and the EEA. Incorrect analysis of SME response data leads to bias and, ultimately, flawed findings. In the next section of this technical report we present a review of the literature on applying SME elicited data in analyses.

Review of Literature on Use of SME Elicited Data in Analyses

There are many different ways to analyze data collected during SME elicitation. practicalguide recommend using data driven methods, i.e., selecting the analysis models to use based on the data collected Meyer and Booker.[122] Another approach is to determine, from the issues for analysis, what statistical models will provide most insight in answering them and then scope the elicitation to collect exactly the input data needed for the models. All statistical models come with some embedded assumptions; it is important for the analysts to know these assumptions and make sure they are *valid*, *necessary*, and *accepted* in accordance with TRAC standards TRADOC Analysis Center.[123] When possible, analysts should avoid assumptions on population distribution and independence Meyer and Booker.[124] Using non-parametric analysis models or simulations offer creative ways to avoid having to make assumptions of these types.

The literature is full of statistical methods that can be useful in analyzing SME elicited data. The following sections discuss a few general types of analyses and provide a few examples.

First Analysis

Before committing to specific models and analytical tools, the analysts should review the data, encoding and collapsing categories as necessary, and perform descriptive statistics in order to gain initial insights into the characteristics and distributions of the data. The

[121] Idem, "Experiences in elicitation", op. cit.
[122] Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.
[123] TRADOC Analysis Center. *Constraints, Limitations, and Assumptions, Code of Best Practice*. Technical memorandum TRAC-H-TM-12-033. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345, 2012.
[124] Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

TRAC *Code of Best Practices for Survey Efforts* provides some techniques for scrubbing, analyzing, encoding, and summarizing survey data Brenda M. Wenzel and Kathy L. Nau.[125]

Encoding

Encoding data refers to mapping the expert responses to real numbers in a way that makes the data useful while retaining its original meaning. Even if the response mode included quantification of expert input, it is a good idea to review the encoding to ensure the resulting distribution accurately depicts the subjective data collected. In addition to making the data quantifiable and measurable, encoding is also used to reduce data to a common scale in preparation for aggregation and other statistical analyses. surveycobp presents a process for encoding survey responses Brenda M. Wenzel and Kathy L. Nau.[126]

One important distinction to make when quantifying expert data is between *ordinal* data and *cardinal* dataMeyer and Booker.[127] Ordinal quantities only provide relative measures; we might know that a "2" rates better than a "1," but we cannot quantify from these numbers how *much* better. Expert rankings among a set of alternatives and Likert-scale responses provide an example of ordinal data. This data can be used in non-parametric statistical tests, aggregations, and visual displays that don't account for distances between values, such as the and other percentiles. Ordinal data is essentially qualitative in nature Anderson and Johnson.[128] Ordinal response data can also be directly mapped to a value function for multi-attribute decision making, but analysts justify the value function they choose Anderson and Johnson.[129]

Cardinal data, on the other hand provides absolute measures according to some scale, imputing meaning on the differences between two values. For example, if we have the quantities "30 seconds" and "40 seconds," we can not only say that 30 seconds is faster than 40 seconds, we can also quantify that difference: it is *10 seconds faster*. madm further decompose cardinal data into a *ratio* scale, which is relative to an absolute zero, and an *interval* scale, which is relative to an arbitrary zero Anderson and Johnson.[130] Most statistical methods and measures are appropriate when using cardinal data, but the analysts must justify the underlying assumptions for any method employed. madm provide the linear scale transformation, the vector normalization transformation, and the non-proportional transformation as three methods of creating value functions from cardinal data for use in multi-attribute decision making Anderson and Johnson.[131]

[125]Brenda M. Wenzel and Kathy L. Nau, *Code of Best Practices for Survey Efforts*, op. cit.
[126]Ibid.
[127]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.
[128]Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.
[129]Ibid.
[130]Ibid.
[131]Ibid.

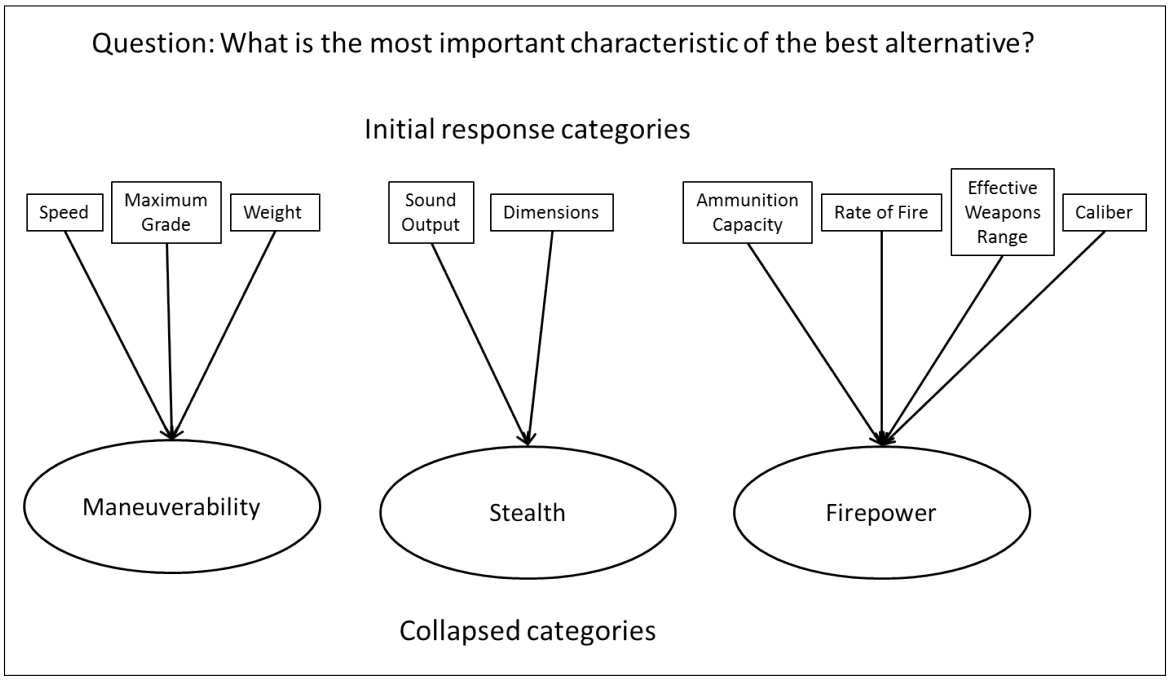


Figure C–1. Example of collapsing categorical data.

Categorical data

The elicitation data might also be *categorical*, with no intuitive mapping to the real number line. For example, an elicitation might include the question: "what is the most important factor in determining the best alternative?" Answers might fall into a few categories, such as "cost," "speed," and "capacity." Because there is no clear numerical ordering of these three alternatives analysts should *not* attempt to assign values to these nominal, qualitative categories Anderson and Johnson.[132] A useful technique in dealing with subjective data of this type is to first assign categories to the responses, and then to collapse categories into broader designations until all response data is grouped into a few, broad categories Meyer and Booker.[133] Figure C gives a notional, simple example of collapsing categories for data elicited concerning the most important factor in determining the best alternative from a selection of combat vehicles.

Qualitative Data

Free response data from the experts is also useful but can be more difficult to quantify. One way to quantify free response data is to categorize it and then collapse categories as described in section C above. The analysts can read through all responses, group those that are similar in nature, and use the groupings to comment on trends and commonalities among the responses Rubemeyer et al.[134] Once the team has established the trends in the

[132]Ibid.
[133]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.
[134]Rubemeyer et al., *Capability Gap Assessment; Blending Warfighter Experience with Science*, op. cit.

comments, it can choose comments that represent each of the categories to help define each category to the study sponsors. These representative statements should be those that best suggest a feature, capability, or quality that might distinguish among alternatives in the analysis or among categories in the response data Rubemeyer et al.[135]

Descriptive statistics

In conjunction with response quantification and collapsed category considerations, the analysts should produce simple summary statistics and create visual depictions of the data, such as histograms or box and whisker plots, appropriate to address the research questions with the data provided as part of the initial data analysis Deavens et al. and Rubemeyer et al.[136] Outlying data points should be investigated for conditioning (see section C) or other sources of bias. The team should also look for and investigate correlations between variables or experts, multi-modality, and clustering Meyer and Booker.[137] The initial insights from this first analysis can assist the study team in verifying initial assumptions or study hypotheses, determining which aspects of the study require further investigation, and ultimately in answering the study issues. The first analysis also is used to verify any modeling assumptions required for the statistical methods that the analysts will employ in follow-on analyses.

Investigation for Bias

While all elicitation events might incur bias, it is not always necessary for the analyst to expend effort in finding and accounting for it in the elicitation data. If an analyst chooses not to focus on bias in the analysis, he or she should state the assumptions that provide justification for ignoring potential bias. Generally, an analyst should avoid focusing on bias unless the scope of the study problem specifically includes becoming aware of bias, preventing or inhibiting the occurrence of bias, avoiding criticism on the quality of the knowledge base, or analyzing the data for the presence of bias. If the study does not explicitly require an investigation into bias, and evidence or risk of bias significant enough to affect the study results does not exist, the analysts can make a statement to this effect in the paper and omit analysis of bias Meyer and Booker.[138]

Detecting bias after the elicitation event is a difficult endeavor. In many cases the methods employed must be tailored to detect a specific bias suspected by the study team. Statistical investigations of correlations within expert answers, between expert demographic data and response data, and between responses from different experts can provide indicators of motivational bias, but these correlations can also have other causes that are relevant to the

[135]Ibid.

[136]Deavens et al., *Support for the Expeditionary Military Intelligence Brigade Commanders' Assessment Workshop*, op. cit.; Rubemeyer et al., *Capability Gap Assessment; Blending Warfighter Experience with Science*, op. cit.

[137]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[138]Ibid.

issues for analysis. If the removal of a single expert or relatively small group of experts from the sample results in significant changes in correlations, it might be an indication of bias or other conditioning effect Meyer and Booker.[139]

Analysis of correlations between similar items, designed tot measure the same attribute, for each expert provide measures of an individual's internal consistency, also called *within-person consistency* or *coherence*. Following the elicitation, it might be difficult to ascertain the reasons for incoherence, which is why they would be better addressed during the elicitation (see section C). The best insurance for obtaining coherent data during an elicitation is good planning: providing the appropriate context, framing, and the specific details to obtain the data of good quality Kynn.[140]

Cronbach's alpha, described in the next section, provides a measure for consistency among all of the experts based on variances among similar elicitation items.

Cronbach's Alpha

Cronbach's alpha provides a measure of internal consistency by comparing correlations among elicitation items aimed at measuring the same attribute. Mathematically, it is defined as

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right),$$

where k is the number of elicitation items measuring the attribute, s_i^2 is the variance of the ith item, and s_T^2 is the variance of the total formed by summing the scores from all of the k items Bland and Altman.[141] If the items are independent, then probability theory tells us that the sum of the variances of the individual items will equal the variance of the total, making $\alpha \to 0$ (it is worth noting that pathological dependencies could cause $\alpha < 0$). On the other hand, if there is correlation among the times, this sum of the item variances will be less than the variance of the total, with $\alpha \to 1$ as the correlation increases.

Applications of Cronbach's alpha typically involve evaluation of a survey for internal consistency, also called survey *reliability* or *inter-rater* reliability, as opposed to evaluation of the coherence of the individual survey takers' (or SMEs') responses Anderson and Johnson.[142] Use of Cronbach's alpha to evaluate internal consistency for an individual over a set of items is possible, but not well documented in the literature. According to alpha, α values from 0.7 to 0.8 are generally considered to be satisfactory for concluding items are measuring the same response in most survey applications, with values higher than 0.8

[139] Ibid.

[140] Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[141] J Martin Bland and Douglas G Altman. "Statistics notes: Cronbach's alpha". In: *BMJ* 314.7080 (Feb. 1997), p. 572. DOI: 10.1136/bmj.314.7080.572.

[142] Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

being even more desirable Bland and Altman.[143] Cronbach's alpha can provide a useful measure of internal consistency for TRAC surveys, either after an elicitation event or for refining surveys following pilot testing (see section C). Cronbach's alpha has the desirable characteristic of only needing one set of test data, as opposed to other methods of survey validation such as test-retest Tavakol and Dennick.[144]

While generally accepted as a measure of survey reliability in the elicitation field, Cronbach's alpha has several limitations. One potential problem is that more items generally lead to a higher value for α, which could result in artificial sense of internal consistency Tavakol and Dennick.[145] Conversely, a short test could result in artificial deflation of α as a measure or reliability.

Another limitation is that α does not provide a measure of unidimensionality, instead unidimensionality is an underlying assumption in using α Tavakol and Dennick.[146] A multi-dimensional set of items, i.e., a set of items measuring multiple latent factors, can still produce a high value for α, but this value will typically underestimate reliability and, more importantly, will not reveal the multiple factors (factor analysis can be helpful in this situation, see section C) Schmitt.[147]

Finally, while many practitioners use a value of 0.7 as an acceptable value for concluding internal consistency among items, there are other sources of variability that contribute to α that provide argument against any universal standard. For this reason, when using α as a measure of reliability, analysts should also provide the correlation matrices or other supporting statistical analyses for the items pertaining to each latent attribute as additional evidence of survey reliability Schmitt.[148]

Point Estimators

Typically, point estimation is applied to estimate a measure of location, variation, or covariation Meyer and Booker.[149] Measures of location include the mean, median, mode, and percentiles. These measures serve as useful ways to aggregate the expert data following the elicitation. For example, suppose an expert elicitation involves asking experts to approximate the maintenance costs associated with a piece of equipment in development. The mean response can be used as an estimator for the mean maintenance costs, or as a way of describing or aggregating the set of expert data collected.

Measures of variation and covariation can be useful too. Standard deviation is a measure of

[143]Bland and Altman, "Statistics notes: Cronbach's alpha", op. cit.
[144]Mohsen Tavakol and Reg Dennick. "Making sense of Cronbach's alpha". In: *International Journal of Medical Education* 2 (2011), pp. 53–55.
[145]Ibid.
[146]Ibid.
[147]Neal Schmitt. "Uses and abuses of coefficient alpha". In: *Psychological assessment* 8.4 (1996), pp. 350–353.
[148]Ibid.
[149]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

variation in a data set and estimates the standard deviation of the population. For example, experts might be asked how much time they typically spend doing a certain task. The standard deviation of their responses provide an estimate of the standard deviation of the population. Covariance measures how much the variation in one variable is related to that of another. Perhaps the analysts collect two times from each expert: the amount of time the typically spend doing "task A" and the amount of time the typically spend doing "task B." The analysts could compute the covariance or the correlation coefficient as an estimate of how related these times are for an individual. For more information on the computation and distributions (given certain assumptions) of these estimates, see Rice[150] or other statistics reference.

Point estimates provide useful insights, but all point estimates have limitations. Measures of location, in particular, are used widely, often without due consideration of their limitations. Any method of aggregating response data into a single measure masks information about the underlying distribution of the data. Furthermore, if the goal of the analysts is to infer from the sample information about the population, as it often is, the analysts must consider the underlying distribution of the statistic they are employing. It might be the case, based on the underlying assumptions in the study and the size of the sample, that the statistical measure employed (e.g., the mean) is not a very reliable estimate of the associated population parameter. In summary, analysts must exercise caution when employing any point estimate or data aggregation method, even when using a simple average in a study. Without consideration of the underlying assumptions in their methods and the distributions of the point estimates employed, the resulting analysis might become flawed.

Significance Tests

Significance testing, also called hypothesis testing, is a method of judging the validity of a specific hypothesis about the population from a given sample. It involves the following steps Ferber:[151]

1. State the goal of the statistical test.

2. Select the *null* and *alternate* hypotheses. These must be contradictory, with the alternate hypothesis aligned with the goal of the statistical test.

3. Choose a value for α, the probability of concluding the alternative hypothesis is true when, in fact, the null hypothesis is true.

4. Identify the test statistic and its distribution based on the null hypothesis. Consider also what values the test statistic is likely to take under the alternative hypothesis, compared to the likely values under the null hypothesis.

[150]Rice, *Mathematical Statistics And Data Analysis*, op. cit.
[151]Robert Ferber, ed. *Handbook of Marketing Research*. New York: McGraw-Hill Book Company, 1974.

5. Compute the value of the test statistic (let's call it x) and its associated p-value. The p value is the probability of obtaining, from a population conforming to the null hypothesis, a sample with a test statistic at least as favorable to the alternative hypothesis as x.

6. Compare the p-value to α and state the conclusion of the test. If $p \leq \alpha$, the analyst concludes that, based on the low probability of obtaining a random sample from a population conforming to the null hypothesis, there is statistically significant evidence in support of the alternate hypothesis. If $p > \alpha$, the analyst concludes that there is not significant evidence in support of the alternate hypothesis.

Recent TRAC studies, such as those presented later in this paper, often investigate how proposed force structure changes will effect certain capabilities within the Army. Significance testing can be useful in these cases. The study team can design a test in which the null hypothesis is that the proposed changes will have no effect. The alternate hypothesis is that the changes will result in an increase in a certain capability, a decrease in cost, or some other effect that can be approximated using SME input. Results from these significance tests can be informative in estimating the effects of the proposed changes.

All significance tests involve certain assumptions which must be justified by the analysts Ferber.[152] Typically, the analysts must assume that the data constitutes a random sample from the population, i.e., that each data point has been drawn independently from a hypothetical, infinite population. These tests also involve some distributional assumptions on the population in forming both the null and alternative hypotheses. Tests on the population median require only hypothesized values for the median as assumptions. Similar, non-parametric tests can be constructed for the mean but are not as powerful as tests involving more specific distributional assumptions on the population. The reason parametric tests on the population mean are more powerful stems from the fact that mean is very sensitive to outlying values with low probability mass. Assumptions on the shape of a distribution serve put a limit on the locations and probabilities associated with outlying values, limiting their potential impact.

In the field of hypothesis testing, different methods exist for interpreting and reporting the results of hypothesis or significance tests. The notable statistician, Sir Ronald Fisher, advocated for reporting the p-value as the level of significance in the results of the hypothesis test, de-emphasizing the need to either reject the null hypothesis or not based on some arbitrary value α Rice.[153] Furthermore, other methods exist for conducting hypothesis tests, such as Bayesian methods that employ likelihood ratios. These methods do not require the analyst to assume one hypothesis is true in order to gather evidence in support of the other, instead using prior distributions to indicate the perceived likelihood that each hypothesis is true. The conclusions of these hypothesis tests are based on posterior probability ratios, which can support either hypothesis.

[152]Ibid.
[153]Rice, *Mathematical Statistics And Data Analysis*, op. cit.

Distributions

In many cases, the analysts seek to construct a probability distribution from the experts' response data. For example, a study might involve a Bayesian statistical analysis. Expert elicited data can provide these distributions and is essential in cases in which limitations in data availability cause the prior distribution to significantly influence the shape of the posterior distribution O'Hagan.[154] Often, a parametric distribution or statistical model is assumed on a certain phenomenon and the parameters are treated as Bayesian random variables, termed *hyperparameters*. The goal of the elicitation becomes to discern a priori distributions for the hyperparameters, which will then be updated using available data. statisticalmethods and literaturereview each provide several methods for eliciting prior distributions Garthwaite, Kadane, and O'Hagan and Jenkinson.[155] There are also multiple methods documented in the literature for eliciting a priori distributions for parameters in a Bayesian linear regression model Garthwaite and Dickey; Kadane et al.; and Garthwaite, Kadane, and O'Hagan.[156]

In simple cases, the response data might consist of a range of plausible values (see section refbias), from a to b where $a < b$ O'Hagan.[157] If the analyst wants to represent this range with a probability distribution, he or she can use the uniform distribution, which has the probability density function

$$f(x) = \frac{1}{b-a}, \ a \leq x \leq b,$$

and is otherwise defined as 0 Garthwaite, Kadane, and O'Hagan.[158] This distribution assumes probability is distributed uniformly along the range from a to b.

Of course, many times the uniform distribution is inappropriate because probability is clearly not uniformly distributed along the entire range of plausible values for some unknown quantity. Often there is a "most likely" value somewhere inside the range of plausible values. For example, an elicitation might seek information about how many inches of steel a new munition should penetrate. The expert might provide the following information:

- Most of the time, the round will penetrate about three inches.
- The maximum penetration will be about four inches.

[154] Anthony O'Hagan. "Eliciting expert beliefs in substantial practical applications". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 21–35.

[155] Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.; Jenkinson, "The elicitation of probabilities-a review of the statistical literature", op. cit.

[156] Paul H Garthwaite and James M Dickey. "Quantifying expert opinion in linear regression problems". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1988), pp. 462–474; Joseph B Kadane et al. "Interactive elicitation of opinion for a normal linear model". In: *Journal of the American Statistical Association* 75.372 (1980), pp. 845–854; Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.

[157] O'Hagan, "Elicitation", op. cit.

[158] Garthwaite, Kadane, and O'Hagan, "Statistical methods for eliciting probability distributions", op. cit.

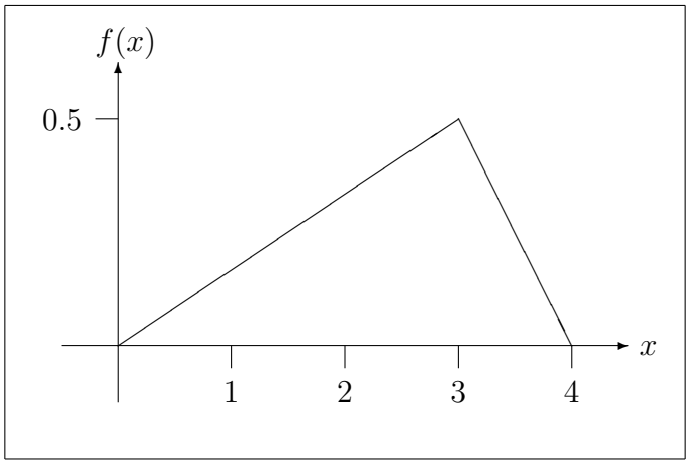


Figure C–2. Plot of the triangular probability density function with $a = 0$, $b = 4$, and $c = 3$.

- In the worst case, the round will not penetrate the steel.

Given this information (a minimum, a; a maximum, b; and a *mode*, c), the analyst can construct a triangle distribution. This distribution has the following probability density function Garthwaite, Kadane, and O'Hagan:[159]

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)} & c < x \leq b \\ 0 & \text{Otherwise.} \end{cases}$$

The triangle distribution has a piecewise linear probability density with the highest probability concentrated in the region near the mode, c. Going back to our example using round penetration, we know from the data provided that $a = 0$, $b = 4$, and $c = 3$. Figure C–2 depicts a plot of the probability density function used to model this uncertain quantity. wolferisk provides a good example of a relevant application of eliciting data to build triangle distributions in support of TRAC studies involving risk analysis Wolfe.[160]

Eliciting additional probabilities or values, such as upper and lower quartiles, from the expert can assist in fitting a more precise distribution to the expert response data. elicitation gives an example of using a gamma distribution to provide a very good fit, provided five expert input values: plausible range (minimum and maximum), median, a specific upper tail probability, and a specific lower tail probability O'Hagan.[161] While the gamma distribution is a "nice" result and well-known parametric distribution, it would be difficult to show how this method would perform better than a similar-shaped triangle distribution based on the minimum, maximum, and mode. This is especially true when one considers the many research findings concerning the difficulty in obtaining accurate

[159]Ibid.
[160]Wolfe, *Operational Risk Analysis, A New Approach*, op. cit.
[161]O'Hagan, "Elicitation", op. cit.

subjective probabilities from humans (see sections C and C) Bedford, Quigley, and Walls; Chesley; and Kynn.[162]

On the other hand, if the analysts elicit data points from many experts, they can fit continuous probability distributions to the empirical distribution coming from these data points. This can be done manually in Microsoft® Excel or using software with distribution-fitting capability. gapassessment use this technique in attempt to quantify risk severity for performance capability measured on a continuous scale, and then use the results to determine transition points between different severity levels Rubemeyer et al.[163]

Monte Carlo Simulation

Monte Carlo methods are useful for approximate modeling of problems involving stochastic events with known or assumed probabilities Ferber.[164] It is an appropriate tool to use when expert data provides information about the probability distributions of unknown quantities or stochastic events (see section C). Often the analyst wants insight on the distribution of an outcome resulting from interactions between multiple stochastic events, i.e., the distribution resulting from a function of multiple unknown quantities for which experts have provided distributions. Analytically deriving the distributional results of complicated functions of random variables is tedious and often intractable, while producing large data sets from a known distribution is usually relatively easy using widely available software such as R or Microsoft® Excel Leemis.[165] The distribution of a function of multiple random variables is thus approximated by evaluating the function many times using the simulated data sets as inputs.

Monte Carlo simulation is very useful in approximating distributions resulting from aggregating or performing other operations involving several unknown quantities with known (or assumed) distributions. However, the analyst must be careful to ensure that the simulation accounts for any underlying correlations between the input random variables. Unless the analyst includes a function forcing correlation between the randomly generated data sets, the input quantities in each computation will be independent. Analysts should document their assumptions on independence or correlation models and how they implemented them into their simulation engine when using Monte Carlo methods.

wolferisk uses SME input to construct triangular distributions (see section C) modeling both severity and likelihood of failure in specific scenarios, and then uses Monte Carlo simulation to approximate the distribution of the overall risk, which is a function of both attributes Wolfe and Headquarters, Department of the Army.[166]

[162]Bedford, Quigley, and Walls, "Expert elicitation for reliable system design", op. cit.; Chesley, "Elicitation of subjective probabilities: a review", op. cit.; Kynn, "The 'heuristics and biases' bias in expert elicitation", op. cit.

[163]Rubemeyer et al., *Capability Gap Assessment; Blending Warfighter Experience with Science*, op. cit.

[164]Ferber, *Handbook of Marketing Research*, op. cit.

[165]Lawrence M. Leemis. *Probability*. Lightning Source Incorporated, 2011. ISBN: 978-0982-91740-4.

[166]Wolfe, *Operational Risk Analysis, A New Approach*, op. cit.; Headquarters, Department of the Army,

Bootstrapping

Bootstrap methods are appropriate when attempting to determine the distribution of a statistic or other function of the response data without making arbitrary assumptions on the distribution on the population Rice.[167] Many traditional statistical methods rely on such assumptions. One of the most common assumptions is that the population follows a normal distribution. The bootstrap method simply replaces these distributional assumptions with another assumption: that the empirical distribution of the data is a good approximation of the population's distribution. This assumption can be difficult in practice to verify or justify (see section C). Nevertheless, the bootstrap method is a powerful tool and is known to perform well under a variety of conditions. Just as in the case of other distributional assumptions, once the analyst has assumed a distribution on the population, he or she can derive the distribution of the statistic or function of interest.

In most cases, direct derivation of the statistic's distribution proves difficult Efron and Tibshirani,[168] although improvements in technology are increasing in capability in this area. The accepted alternative to direct derivation is Monte Carlo simulation, which is almost always applied in bootstrapping. This simulation consists of generating a large number of samples from the empirical distribution, computing the value of the statistic for each, and then using the resulting distribution as an approximation for the distribution of the statistic.

In SME elicitation, bootstrapping can be useful in investigating correlations and bias in the data, characterizing and analyzing uncertainties in the expert's estimates, and exploring the distribution of aggregations and other point estimators resulting from the response data Meyer and Booker.[169] Using the bootstrap to examine differences in distributions for a certain statistic, e.g., the median, based on different bootstrap sampling strata can assist an analyst in identifying sources of conditioning and bias. practicalguide give more details on this method Meyer and Booker.[170]

Bootstrapping is a very useful statistical method, but, as with any model, has its disadvantages. Analysts using this method should consider and state their assumption that the empirical distribution of the response data provides a reasonable approximation of the response population (the population must be clearly defined). If the original sample is not a good representation of the distribution of the population, the bootstrap method will not provide accurate results. Another potential problem with the bootstrap method pertains to the sensitivity of the statistic of interest: if the statistic is very sensitive to minor fluctuations of specific percentiles (especially outer percentiles) of the underlying population's distribution, the bootstrap might not provide an accurate representation of

FM 5-19: Composite Risk Management, op. cit.

[167]Rice, *Mathematical Statistics And Data Analysis*, op. cit.

[168]Bradley Efron and Robert Tibshirani. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy". In: *Statistical science* (1986), pp. 54–75.

[169]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[170]Ibid., p. 280.

the distribution of the statistic Rice.[171] Before using the bootstrap method, analysts should justify their assumption that the empirical distribution of the response data provides a good approximation for the underlying distribution of the response population, and that the target statistic is not overly sensitive to percentiles that will be approximated in the empirical distribution as a result of the sample size.

Multivariate Techniques

This section discusses several common multi-variate statistical methods that might prove useful in analyzing expert judgment.

Regression

Regression analysis involves an assumption that a phenomenon (the *response*) of interest occurs as a function of one or more variables (the *predictors*), which can be measured, combined with some uncertainty, or noise. The uncertainty is assumed to have an average value of zero, so that the regression function can be used to provide the *expected* value of the response based on specified input values for the predictor variables. Regression can also be thought of as a method of fitting data points to a curve defined by a parametric function; the function provides an assumed relationship in the data set, while the goal of the analyst is to estimate the parameters. The method of linear least squares is the most common, but not only, method for determining parameters in regression problems. For linear models, it provides unbiased estimates that exhibit other desirable statistical qualities Rice.[172]

Linear regression is probably the most well-known and widely used regression model. It assumes that the expected value of the response, conditioned on the predictors, is a linear function of the predictor variables Hastie, Tibshirani, and Friedman.[173] In other words, the assumed relationship is

$$Y = \beta^T \mathbf{x} + \epsilon,$$

where Y is the response variable, β is the vector of coefficients (the parameters), $\mathbf{x}$ is the vector of predictor variables, and ϵ is the independent, normally distributed noise with mean equal to zero. The standard deviation of ϵ is constant and provides another parameter for this model, so that if there are n predictor variables, there are $n + 1$ parameters that must be estimated by the analyst. The resulting model is useful for quantifying the relationships between the variables in the model, ascertaining the significance of the uncertainty, and for making predictions.

Regression can be useful in TRAC SME elicitations when trying to determine a

[171] Rice, *Mathematical Statistics And Data Analysis*, op. cit.
[172] Ibid.
[173] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. New York: Springer Science, 2001. ISBN: 978-0387-95284-0.

relationship between two quantities. For example, analysts might be interested in whether the number of years of experience an expert has is related to his or her response to a certain question (note that this might be an investigation into conditioning bias; see section C). Regression offers a method of investigating this relationship, in which number of years experience is explored as a potential predictor of an expert's response to a specific question.

When using regression, the analyst must justify the assumptions embedded in the relationship between the quantities of interest embedded in the selected model. Justification can include investigations of relationships using scatter plots as well as goodness of fit and model utility tests, many of which are standard in any regression software. Even if the results of the regression indicate a strong relationship between two variables, the regression itself does not establish a causal relationship between the predictor and response variables. Additional investigation is required in order to determine the cause of the correlation between the two quantities.

There are many regression models beyond simple linear regression. Logistic regression is another widely-used linear statistical model that is used to classify an element based on its attributes. While we will not go into detail in this paper, logistic regression has become a popular tool in statistical analysis and has been used in conjunction with SME elicitation O'Leary et al.[174] Bayesian linear models provide another means for statistical analysis that is often coupled with SME elicitation. Techniques for eliciting a priori distributions for the parameters of a Bayesian linear model are discussed in section C.

Cluster Analysis

Most of the statistical methods and models discussed in this paper relate to *supervised* learning. These methods involve assumed relationships between the variables. Cluster analysis is an *unsupervised* learning method that attempts to group the data into subsets that are more closely related to one another Hastie, Tibshirani, and Friedman.[175] The goal is to discover how items tend to group, and which attribute or combination of attributes best distinguish between the different clusters. Cluster analyses can be useful in analyzing the demographic information on the experts, investigating reasons behind multi-modality in the data, investigating evidence of conditioning, and providing initial insights into relationships among the different variables. Most cluster analyses require very few assumptions, but the analyst must decide in advance how to measure "closeness" between two or more data points.

[174]O'Leary et al., "Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby Petrogale penicillata", op. cit.

[175]Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, op. cit.

Factor Analysis

Factor analysis provides a means of variable reduction and can be used to identify and analyze "latent" variables within a data set Hastie, Tibshirani, and Friedman.[176] It consists of two variants: *exploratory* factor analysis and *confirmatory* factor analysis.

Exploratory factor analysis is the unsupervised learning version of factor analysis. It provides a method for discovering meaningful commonality among subsets of the survey questions. The method assumes that an expert's response to each question is a linear combination of unobserved, uncorrelated latent factors plus some random noise, i.e.,

$$X = \mathbf{A}S + \psi E,$$

where X is the vector of n responses for a single participant, $\mathbf{A}$ is the $n \times m$ coefficient matrix (also called *factor loadings*), S is the vector of the m latent variables ($m < n$), ψ is the vector of loadings on the random noise, and E is a vector providing the random noise associated with each response, each distributed according to the standard normal distribution.

Typically, the analyst assumes that these latent factors follow a normal distribution and then uses the method of maximum likelihood to fit the model to the data Hastie, Tibshirani, and Friedman.[177] The analyst must decide on the number of latent factors, m, in advance. Because there are many degrees of freedom, there are multiple best-fit models, which are essentially geometric rotations of each other. One method of choosing the rotation in fitting a factor analysis model is to choose the one that maximizes the highest factor loadings and minimizes the lowest Fricker Jr., Kulzy, and Appleget.[178] By keeping the factor loadings relatively high or low, with few moderate values within the range, it becomes easier to separate the survey questions into groups that are explained by each of the factors. fricker give a very good explanation and methodology for applying exploratory factor analysis to survey data Fricker Jr., Kulzy, and Appleget.[179] The TRAC *Africa Knowledge, Data Source, and Analytic Effort (KDAE) Exploration* provides an example of factor analysis to support TRAC research Deveans et al.[180]

The exploratory factor analysis model involves many assumptions, such as assumptions on the distributions of the factors and the simple assumption that latent factors exist. It also involves several, seemingly arbitrary decisions on the part of the analyst, such as the number of factors to use to build the model and which rotation of the maximum likelihood best fit solution set to use. This degree of subjectivity has resulted in skepticism and reluctance of use among the statistics community Fricker Jr., Kulzy, and Appleget and

[176] Ibid.

[177] Ibid.

[178] Ronald D. Fricker Jr., Walter W. Kulzy, and Jeffrey A. Appleget. "From Data to Information: Using Factor Analysis with Survey Data". In: *Phalanx* 45 (4 2012), pp. 30–34.

[179] Ibid.

[180] MAJ Thomas Deveans et al. *Africa Knowledge, Data Source, and Analytic Effort (KDAE) Exploration*. Tech. rep. TRAC-M-TR-12-037. 700 Dyer Road, Monterey, CA 93943-0692: TRADOC Analysis Center, Monterey, 2012.

Hastie, Tibshirani, and Friedman.[181] Ultimately, factor analysis can provide useful insights into expert response data, especially when the responses are in a survey format. However, the analyst must justify the assumptions in the model as well as the subjective decisions made.

Confirmatory factor analysis can be used by analysts to test whether a hypothesized survey structure adequately fits the observed data Fricker Jr., Kulzy, and Appleget.[182] It is very similar to exploratory factor analysis except that the number of factors and the rotation of the solution are dictated by a hypothesis on the response data. The resulting loadings are used as evidence to support or discredit the hypothesized survey structure.

Discriminant Analysis

Discriminant analysis is a method for producing a set of functions, or boundaries, designed to classify data points into different categories Hastie, Tibshirani, and Friedman.[183] The categories form the response. This analysis could be useful in attempting to figure out what characteristics had the most impact in expert's responses. For example, we might know an expert's years experience, basic branch, number of deployed months, number of vehicle operation hours. Suppose we ask these experts what the most important feature is for the new ground combat vehicle. We might collapse the responses into three categories as shown in figure C. We could then find best fit discriminant functions to categorize, or predict, and expert's response based on his or her years experience, branch, deployment history, and vehicle operation history.

Discriminant analysis assumes a multivariate normal distribution for all variables, and does not account for noise. Neither of these assumptions is often realistic. practicalguide recommend using discriminant analysis as an exploratory tool only, backing up any initial insights with more rigorous follow-on analyses Meyer and Booker.[184]

Analysis of Variance

Analysis of variance (ANOVA) is primarily concerned with comparisons of means among different groups Rice.[185] This analysis begins by assuming that the means are the same for each group, i.e., group distinctions have no effects on the distribution of the population. For example, a TRAC study might involve asking vehicle operators (group A), vehicle commanders (group B), platoon leaders (group C), and company commanders (group D) a

[181]Fricker Jr., Kulzy, and Appleget, "From Data to Information: Using Factor Analysis with Survey Data", op. cit.; Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, op. cit.

[182]Fricker Jr., Kulzy, and Appleget, "From Data to Information: Using Factor Analysis with Survey Data", op. cit.

[183]Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, op. cit.

[184]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.

[185]Rice, *Mathematical Statistics And Data Analysis*, op. cit.

question about vehicle maneuverability. The analyst might want to know if the responses are different based on the distinctions of these four groups. Using ANOVA, the analyst can determine whether evidence exists to reject the null hypothesis that:

$$\mu_A = \mu_B = \mu_C = \mu_D.$$

ANOVA computes the sample means of the four groups and then compares the variations between these four means to variations within each group. Under certain assumptions, including the null hypothesis, the ratio between standardized measures of these two sources of variation follows and F-distribution and can be used to find a p-value for the hypothesis test. As with any hypothesis test, a low p-value gives evidence against the null hypothesis.

There are many different ways to conduct ANOVA. The previous paragraphs describes the simple case of a one-way layout, i.e., one dimension of groups in which the analyst is interested. A two-way layout would support analysis of a second set of groupings. Continuing with the previous example, suppose that two units fielded this particular vehicle. One unit used it in a major combat operations scenario, the other in a stability operations scenario. Now the analyst wants to test the responses for effects based on the position of the respondent *and* the unit; the analyst might use a two-way ANOVA.

ANOVA relies on several assumptions. First, the variance within each of the groups is assumed to be equal. A common assumption is that the data in each group follows a normal distribution; in this case the one-way ANOVA model is

$$X_{i,j} = \mu_X + \alpha_i + \epsilon_{i,j},$$

where $X_{i,j}$ is the jth measurement in the ith group, μ_X is the underlying population mean, α_i is the group effects for group i (essentially the group i effect on the population mean), and $\epsilon_{i,j}$ is the normally distributed noise, with mean equal to zero and constant standard deviation. The null hypothesis then translates to $\alpha_i = 0 \, \forall i$ Rice.[186]

In addition to the assumptions, one of the disadvantages of using ANOVA is the output. The results might provide significant evidence that the different groupings have an effect on the mean. However, the test does not give much additional insight into what that difference is, forcing the analyst to conduct follow-on statistical investigations Rice.[187] Several methods exist for further quantifying the differences between the groups.

Finally, non-parametric methods for conducting ANOVA exist, allowing the analyst to compare groups without the assumption of normality. The Kruskal-Wallis test provides a method for ANOVA that uses rank data instead of the measurements. This test can be of particular use in TRAC SME elicitations because often the response data consists of ordinal values, such as rank data.

[186]Ibid.
[187]Ibid.

Relative Importance or Weights

Some TRAC studies involve applications of multi-attribute decision theory to evaluate and compare courses of action or alternatives Anderson and Johnson.[188] We present two methods for determining relative importance among a set of attributes from SME elicited responses. The rank ordered centroid assessment relies only on attribute rankings provided by the SMEs. Saaty's technique for determining weights requires the SMEs to make comparisons between each pair of attributes.

parameter-free provides a non-parametric method for deriving weighting functions that also relies on pairwise comparisons, which we do not cover in detail in this paper Abdellaoui.[189] comparison compare three methods for eliciting weights: Max100, Min10, and direct rating Bottomley and Doyle.[190] This reference also mentions the point allocation method. Because these methods deal more with how to conduct the elicitation than the analyses using the data, they are discussed in section C.

Rank Ordered Centroid Assessment

Rank ordered centroid assessment provides an estimation technique that converts ordinal data to relative ratio data, seeking the minimum loss of fidelity. This method can be used to determine attribute weights from SME rank data in a multi-attribute decision making problem Anderson and Johnson.[191] Given n attributes for which a SME provides rankings $1 \ldots n$, the rank order centroid produces the following weight, w_i, for the ith ranked attribute:

$$w_i = \frac{1}{n} \sum_{j=i}^{n} \left(\frac{1}{j} \right)$$

The advantage to using this method is that it requires only rank data to be elicited from the SMEs. The disadvantage is that it quantifies, somewhat arbitrarily, the differences between the ranked attributes using a heuristic approach.

Saaty's technique for pairwise data analysis

The idea behind Saaty's technique is that experts have an easier time comparing options in pairs, rather than comparing multiple options against each other at the same time Meyer

[188] Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

[189] Abdellaoui, "Parameter-free elicitation of utility and probability weighting functions", op. cit.

[190] Bottomley and Doyle, "A comparison of three weight elicitation methods: good, better, and best", op. cit.

[191] Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.

and Booker.[192] This method can be of use to TRAC because it's goal is to develop a relative value or weight from a list of alternatives or attributes, a common requirement in TRAC studies and a common goal of a TRAC SME elicitation. Suppose there are n alternatives for comparison. An expert fills out the upper triangle of an $n \times n$ matrix in which each entry indicates a pairwise comparison. The matrix has ones along the diagonal, representing equality when comparing each alternative with itself. For the entry in row i, column j, for $i < j$, the expert must:

- Determine which alternative is better, i or j.
- Determine how *much* better, on a numerical scale from one to nine. One indicates the two alternatives are the same, while nine indicates the chosen option is absolutely superior in all aspects. Let x be the number selected by the expert.
- If i is better, enter x in row i, column j. If j is better, enter $1/x$ in row i, column j.

Once the expert finishes determining the values for the upper triangle, the analyst can fill in the lower triangle with the reciprocal values, then find the matrix's eigenvalues. These eigenvalues provide the relative weights or values for the importances Meyer and Booker.[193] madm provide alternative methods for solving for the weights from the comparison matrix Anderson and Johnson.[194] Saaty's technique also provides a method of calculating a consistency ratio, which provides a measure of within-person consistency. Correlation matrices and ANOVA methods are also useful for measuring inter-rater reliability when using Saaty's method with multiple expertsAnderson and Johnson.[195]

Saaty's method has the advantage of being more mathematically rigorous than the rank ordered centroid assessment, because the elicited data is already cardinal, as opposed to ordinal, in nature. The disadvantage of Saaty's method is that it involves more work on the part of the experts, and is therefore open to more forms of bias which could counteract any gains made in using a more rigorous method. A study team considering using Saaty's method of pairwise comparison should consider that, in order to determine weights for n attributes, experts will have to make $\frac{n(n-1)}{2}$ comparisons. Remaining consistent for a large number of comparisons is naturally more difficult than simply ranking the attributes.

Bayesian Methods

Bayesian methods differ from traditional statistical methods by assuming that the population characteristics of interest are random variables, rather than fixed quantities. In

[192]Meyer and Booker, *Eliciting and Analyzing Expert Judgement, A Practical Guide*, op. cit.
[193]Ibid.
[194]Anderson and Johnson, *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation)*, op. cit.
[195]Ibid.

order to use a Bayesian method, the analyst *must* assume a prior distribution on the quantity or quantities of interest. This prior distribution can be "uninformed." For example, an analyst might be interested the percentage of experts that think that a low malfunction rate is more important than a high cyclic rate of fire. The uninformed prior would be that this percentage, θ, comes from a standard uniform distribution. The expert would then collect the data (i.e., elicit expert judgment) and then use the resulting data to update the prior distribution:

$$f(\theta|k) = f(\theta) \cdot \frac{p(k|\theta)}{\int_{-\infty}^{\infty} p(k|\theta) f(\theta)\, d\theta} = 1 \cdot \frac{nk\theta^k(1-\theta)^{n-k}}{\int_0^1 nk\theta^k(1-\theta)^{n-k} \cdot 1\, d\theta},$$

where k is the number of experts that indicated the low malfunction rate was more important, n is the number of experts surveyed, $f(\theta|x)$ is the posterior probability density function of θ conditioned on the data, $f(\theta)$ is the prior (standard uniform) probability density function of θ, and $p(k|\theta)$ is the probability mass function of k conditioned on θ (note that this follows a binomial distribution).

This posterior distribution reduces to the well-known beta distribution with $a = k + 1$ and $b = n - k + 1$ Rice.[196] The analyst can use the mode of this distribution as a maximum likelihood estimate for θ, or form a 95% *credible* interval (the Bayesian analogue to a confidence interval) for θ.

For many traditional methods of statistical inference, analogous Bayesian methods exists aimed at answering the same statistical questions. Bayesian methods are most appropriate when the the analyst can quantify prior beliefs or knowledge about a quantity of interest using a probability distribution, and that quantification somehow serves to benefit or inform the analysis. Because quantification of beliefs is often a goal in SME elicitation, Bayesian methods can be useful in some TRAC analyses. In particular, if an analysis is to gather empirical data for Bayesian statistical analysis, SME elicitation can be used to produce informed prior distributions (see section C) Gelman et al.; O'Leary et al.; O'Hagan; Garthwaite and Dickey; and Kadane et al.[197] If an analyst chooses to use Bayesian statistical methods, he or she should include sensitivity analyses on any a priori distribution assumptions made to support the analysis.

[196] Rice, *Mathematical Statistics And Data Analysis*, op. cit.

[197] A. Gelman et al. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2003. ISBN: 9781420057294; O'Leary et al., "Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby Petrogale penicillata", op. cit.; O'Hagan, "Eliciting expert beliefs in substantial practical applications", op. cit.; Garthwaite and Dickey, "Quantifying expert opinion in linear regression problems", op. cit.; Kadane et al., "Interactive elicitation of opinion for a normal linear model", op. cit.

This page intentionally left blank.

Appendix D
Interviewing Techniques

Guidelines for TRAC Analysts for Conducting Interviews and Focus Groups White Paper

Patricia Kinney, PhD

TRADOC Analysis Center

Introduction

Interviews and focus groups are methods for eliciting information from people, obtaining their viewpoints, opinions, expert judgments, etc. for a specific purpose. Well known TV and radio journalists (e.g., Barbara Walters, Terry Gross,) may make it look easy; but it is more complicated, complex, and involved than it appears.

In the context of this white paper, the interview is one of several possible methods for collecting required information for a research study or analysis. Specific procedures must be followed to obtain the needed information while at the same time limiting bias and not influencing responses from the people being interviewed (Soldiers, Subject Matter Experts (SME), or participants). Inexperienced or untrained interviewers and poor question wording can have a dramatic effect on the outcome of the interview, which in turn can affect the validity and reliability of the data. With respect to interviews, validity is the accuracy to which the questions obtain the information they are intended to acquire. Synonyms for reliability include consistency, dependability, and trustworthiness. With interviews, reliability is the precision or consistency of measurements of the data collected across interviews (i.e., from interview to interview). With reliable techniques and question wording, the interviewers can trust that they are obtaining the same general information from all participants with little variation in data elicited from person to person. To maximize validity and reliability, interviewers must know how to properly conduct interviews and the questions must be worded to reduce differences in interpretation of those being interviewed.

Typically, an interview is conducted with one individual and in some circumstances, two individuals. Interviewing more than two people at the same time is considered a focus group. Essential interviewing techniques apply to focus groups as well; however, additional considerations and guidelines are needed for conducting focus groups. These are presented in a separate section below.

Purpose

The purpose of this white paper is to provide guidelines for conducting interviews and focus groups that are based on the accumulated education and years of experience of TRAC analysts in combination with the views of other professionals in the field (see bibliography). This white

paper provides more of an overview than a treatise. For the interested reader who desires a more thorough discussion, there are numerous books and articles that provide more depth and detail.

Scope

The focus of this white paper is an overview on the procedures and processes for conducting interviews and focus groups for the purposes of obtaining data from Soldiers and civilians, including SME, for TRAC studies and analyses. This paper does not cover arranging access to the SMEs or persons to be interviewed, procedures for adherence to human research protection protocols (HRPP), obtaining permission or waivers for conducting human research, complying with directives for collecting personally identifiable information (PII), adhering to informed consent protocols, detailed instructions on constructing questions, and data analysis. The analyst must be knowledgeable of these procedures to collect information with both surveys and interviews. In addition, the interviewing procedures discussed here apply after the following have been accomplished. These

- The measurement space has been developed.
- Data requirements have been identified.
- A determination has been made that interviews and/or focus groups are the appropriate data collection method and are worth the time and costs.
- The Soldiers or people who have the knowledge to provide the needed data have been identified.

Presentation of Information

The information is presented in the following sections.

- Types of Interviews
- Preparing for the Interview
    - Constructing Questions and Question Sequence
    - Practicing the Interview
    - Training Interviewers
    - Coordinating Interview Administration
- Conducting the Interview
    - Opening the Interview
    - Asking the Questions and Interviewer Effects
    - Recording
    - Ending the Interview
    - Special Situations and Other Considerations
- Focus Groups

Types of Interviews

Interviews generally fall into one of three types; structured, unstructured, and semi-structured.

Structured

These types of interviews are structured in the sense that there is a structured plan for collecting the information. It is an art and a science at the same time. This type of interview is used when the researcher has a good understanding of the topic of interest, knows what data need to be obtained, and has a clearly specified set of questions to be asked. All of the questions that will be asked are within the framework of the problem and have been written in advance. The interviewer typically asks the same questions of each participant in the same order. Nonetheless, experience has shown that rigid adherence to asking the questions in the same order in all situations can interrupt the flow of information and the person's thought process. (More discussion on that is presented below in the Asking the Questions section.) Some questions may require the participant to select among a set of pre-determined answers and each person is presented with the same set of possible responses. The questions are asked and answered within the same context which reduces the possibility of interviewer effects and increases reliability. Open-ended questions can be included; but, their focus is on the specific issues of the research topic or they allow the participant to elaborate further on specific issues. For example, if you first ask "how satisfied are you with your unit training?" and the Soldier says "not satisfied", possible open-ended questions would be: "please explain your answer" or "what are the reasons you are dissatisfied?" Also if you are guiding SMEs through a set of questions to obtain effectiveness assessments of a system, you may want to ask them to explain their assessments.

Unstructured

This type of interview is used to explore lines of inquiry, probe a topic in more depth, conduct in-depth interviews on a particular topic, tap into the experiences of the participant, obtain a range of perspectives on an issue from several people, or obtain a better understanding of an issue that will be investigated further. This is when you can "pick someone's brain". Unstructured interviews can help narrow the scope for future research efforts and define the data requirements for subsequent data collection or structured surveys. The interview contains mostly open-ended questions, a brief set of prompts to deal with a range of topics, few or no questions with pre-determined answers, and there is no requirement that the questions be asked in any order. Additional questions may be added as the interviewer deems necessary. The interview is flexible and can allow the participant to determine what is relevant and take the interview in the direction of what is significant to him/her.

Semi-structured

A semi-structured interview is more flexible than a structured interview and not as organized. It is more orderly than an unstructured interview; but, is not as probing and exploratory. It is typically used when the researcher has a list of fairly specific topics to be covered that are used as a guide for the interview. You use this type of interview when you have a good idea of the

data you need and the direction you need to take, but you also need to explore some issues further. The order of questioning should follow a logical path, but the questions do not have to follow exactly as organized and additional questions may be added as the interviewer deems necessary. Furthermore, you allow the participant to explain or expound on what he/she sees as important.

Preparing for the Interview

Preparing for the interview is a critical step in the interview process to maximize the validity and reliability of the data and reduce error. Preparation involves constructing the interview questions, organizing the question format, coordinating administration of the interviews (which includes arranging for the correct people to be interviewed, scheduling the dates and times of interviews, and arranging for a favorable environment), pilot testing, practicing, and training the interviewers.

Constructing the Questions and Question Sequencing

This section provides an overview of the key issues regarding constructing questions and question sequencing. Many of the guidelines on constructing survey questions also apply to constructing interview questions. Complete guidance can be found in two published TRAC documents: Survey Code of Best Practices (April 2014) and Enhancing Subject Matter Expert Elicitation Techniques (June 2013) and in The Art of Asking Questions (Payne 1951).

Constructing the Questions. It is important that all of the interviewers are asking the same questions in the same context during each interview. This applies even if there is only one interviewer. Proper question wording and sequencing go a long way in achieving data consistency. The wording of questions largely determines the answers that will be given. There can be different meanings and interpretations of words and phrases due to a participant's background, experience, education, etc. Use simple language and language that is meaningful to the target population. The interview should have a conversational tone; but the questions should also be grammatically correct. Avoid jargon, abbreviations, and acronyms that might not be known to everyone. Using common Army acronyms and acronyms familiar to the Soldier or person can help establish rapport and trust. Be sure to define any acronyms the first time you use them.

The intent is to have questions that really get the person to address what the interviewer wants. Equally important is that the questions are totally neutral, in that the wording of each question does not affect what the person will say. What you ask and the way you ask affects the answer, because all people have feelings. We want the person to reflect on his/her thoughts and feelings unhampered by reactions to the interviewer or how the interviewer states things.

Some interviewing experts suggest avoiding asking questions that can be answered with a "yes" or "no" because, usually, the interviewer is really interested in the explanation behind the answer

and because inexperienced interviewers may record only the "yes" or "no" response, without asking for the reasons. However, using a "yes" or "no" question can be useful because it helps to frame the issue and direct the person's thought process. For example, if you first ask the question "did you have any problems (difficulties) firing the new weapon?" If they say yes, then ask them to explain ("what difficulties did you have?"). In any case, ensure that a follow-up question for each "yes/no" question is included to obtain the reasons for their responses. Avoid asking questions that begin with "why" because some people may feel they are being evaluated. Instead, you could ask "in what ways" or "what do you think". If presenting the person with a list of choices to select from, it is better to use an unfolding technique than to expect the participant to remember everything you say (Frey and Oishi,1995). Instead of presenting a long list of possibilities, give them a short list and then ask clarifying questions. For example, first ask "was the system easy or difficult to employ?" If they say difficult, then ask was it "very difficult, moderately difficult or somewhat difficult?"

In unstructured interviews, the following examples can be used: "take me through the process", "describe your experience", "start from the beginning and tell me about___".

Question Sequencing. Consideration must be given to the order and flow of questions. The way you order the questions will depend in some part on the type of interview (structured, unstructured, or semi-structured). Unstructured interviews are more free-flowing and do not always need to be asked in a specific order. A factor to keep in mind is that a participant's exposure to one question might influence how he/she answers subsequent questions.

Start the interview with general and/or factual questions and ask sensitive questions later. This helps to put the participant at ease. In most interviews, questions should be grouped in a meaningful way such as by topic and organized in a way that does not require the interviewer to page back and forth. Also, provide a meaningful transition from one question to the other. Use transition statements that introduce and separate sections or topics. Examples are: "now I am going to ask you about system effectiveness", "I would like to learn more about your most recent deployment", "the next set of questions is about___". For most interviews, use a format that transitions from more general to specific questions within each grouping; each succeeding question is related but has a narrower focus. Also, some interviewers ask the most complex questions early-on in the interview process to assure they are obtaining the critical information before fatigue becomes an issue. Frey and Oishi (1955) suggest that you will better establish a rapport with the participant if you postpone asking the demographic questions later. However, my experience with interviewing military personnel has shown that I need to know their frame of reference (rank, military occupational specialty (MOS), current duty assignment, previous duty assignments, etc.) to better understand their responses.

If you want to see some examples of question wording and sequencing, the TRAC Technical Reports Office (TRO) has copies of most of TRAC-WSMR's Training Effectiveness Analysis

(TEA) reports. Many of the TEAs used interviews as a method for collecting some of the data and in most cases, the reports include the interview questions in the appendices.

Coordinating Interview Administration

Administration of the interview is a factor in quality control, and affects data quality. Confirm your visit and interview administration requirements with your coordinator or interview participants. Prior to your data collection visit, determine the location and accommodations of where you will be interviewing and, if necessary, request the accommodations you will need. Strive to conduct the interview in a quiet, secluded area. The participants will be more at ease if they can talk with you privately. Each interview should be no longer than 60 minutes. Any longer than that, people become fatigued or fidgety which can affect data quality and completeness. Schedule the interviews to allow for breaks in between every interview. Too many interviews right after the other can affect your mental process and you will become fatigued. Breaks also allow you time to go over your notes and fill in details where necessary while your memory is still fresh and to jot down other observations.

Pilot Test the Interview Questions. Pilot testing the interview allows you to find flaws or parts that need changing in instructions, question wording, question flow, etc. so you can correct them before you interview the target population(e.g., Soldiers, SMEs). Have other analysts and military personnel review the final draft of the interview document. Follow the same procedures for pilot testing a survey, which can be found in TRAC's Survey Code of Best Practices (April 2014).

Practice. Before conducting the actual interview, it is essential that you and the interviewers fully understand the subject matter and the information that must be obtained. Practicing the interview will help with that and will help you become comfortable when conducting the interview. It is a good idea to practice asking the interview questions aloud. The goal is to ask the questions without stumbling over the words and eliminate any indication that you are unprepared or unfamiliar with the subject. Your comfort in asking the questions will also help the participant feel more at ease which will facilitate the interview process.

Training the Interviewers. If more than one person will be conducting the interviews, an interviewer training course is essential for consistency among interviewers and for reducing potential errors. The interviewer needs to know that he/she can unintentionally bias the results. Participants will react to subtle cues that the interviewers may not know they are projecting. Conducting an interview is not the appropriate circumstance for learning on the job or through trial and error. When possible, team members unfamiliar with interviewing techniques can gain initial experience as an observer and/or a note taker. Start the training session with the purpose of the study/analysis and the purpose of the interview; i.e., the data the interview is designed to collect. Provide administrative details such as dates, time to administer, scheduling, etc. Cover the guidelines for conducting the interview, asking questions, demeanor, appearance, interviewer

effects, solutions to problems, recording procedures, and making notes after each interview on important issues, such as conditions, noise, interruptions, and impressions. Go over the questions step by step so that interviewers understand the needed data and how it will be used and possible problems with the question or responses. A useful procedure is to have the interviewers observe the project lead or lead interviewer conduct the interview with someone else. After that, the interviewers should practice giving the interview several times with different people and also serve as the person to be interviewed. See Frey and Oishi (1995, pp117-143) for more detailed instructions on conducting interview training.

Conducting the Interview

This section covers opening the interview, asking the questions, interviewer effects, recording the interview, ending the interview, and special situations, and considerations.

Opening

Make sure the participants you are interviewing are comfortable and establish eye contact with them. Do not allow observers or listeners. Introduce yourself, briefly describe your job, your organization, and the purpose of the study/analysis effort. Emphasize the importance of their participation to the outcome of the study/analysis effort. Assure confidentiality of answers. Tell them how long the interview will take. Ask if there are any questions before starting. Answer questions in an honest, straight forward manner unless your answers could bias the outcome of the interview. Provide as brief of an answer as possible that will suffice and does not introduce bias. If you think you cannot answer their questions, simply state that you cannot answer the question. Some small talk may help in establishing rapport and break the ice, but should be short and not reveal your opinion on the subject. This is not the time to exchange war stories or tell the participant about your background and experiences. If you do, the participants may want to please you with their responses, provide only those responses they think you want to hear, and so on. Before you start asking the questions, make sure you are familiar with the participant to include background, current duty position, situation, etc. so you can understand what the person is saying in his/her own terms.

Asking the Questions

Correct interviewing is not having a conversation or chat with someone. The role of the interviewer is to listen, obtain the needed information in an unbiased way, control the situation, steer the course of the interview, explore, direct, redirect, encourage relevant answers and discourage irrelevant responses without disrupting the person's thought processes, and ensure that the same general areas of information are collected from each participant. Interviewing is a delicate balance because the interviewer must encourage the participant to answer the questions, but not influence the responses.

The interviewer should speak clearly and present a positive, confident, self-image with a need to know. Be friendly, respectful and courteous, but remain professional and maintain the tone of a natural conversation. Do not tolerate rudeness or obnoxious behavior from the participant. The interviewer's demeanor/deportment is critical. The interviewer needs to be tolerant, sensitive, patient, neutral, non-judgmental, and exercise good listening skills. A neutral manner is one that does not imply or convey criticism, surprise, approval, or disapproval. It is important to be encouraging, yet, not react to the participant's response. Do not be condescending or arrogant and do not judge or criticize the response. It is also important that the interviewers know how their behaviors, attributes, and characteristics could influence the responses.

Good listening skills are the foundation of good interviewing (Guba and Lincoln, 1982). "Listening is the most important skill in interviewing. The hardest work for most interviewers is to keep quiet and to listen actively" (Seidman, 1998). Listen carefully and record as precisely as you can what the participant is saying. Be flexible in listening because people have different ways of expressing themselves and explaining things, and may not answer the question directly, instead answering in a more round-about way. Concentrate on the substance of the responses to fully understand and to assess whether what the Soldier/person is saying is as complete as you need (Seidman, 1998).

The interviewer should say nothing that is not absolutely necessary to obtaining the required information. Interviewers should not respond or discuss the questions or answers with the person and not submit their opinions or experiences. That biases the responses and contaminates the results. The interviewers may restate the question, rephrase the question slightly, request more detail or the reasons, and paraphrase the person's response to make sure the interviewer heard it right. Sometimes the participant may not understand the question and the interviewer may have to re-phrase the question. This must be done in a way that does not introduce bias or change the meaning of the question. Changing the meaning of the question could affect the validity of the response.

There are many interviewer behaviors, attributes, and traits that could influence the responses. The interviewer is part of the interview process because it is an interaction between people. Interviewer effects are inevitable, so it is the interviewer's job to keep those effects to a minimum. The interviewers' sheer physical presence combined with visible characteristics, such as age, dress, rank, job status, race, etc. may influence the Soldier's/person's responses. Dress and appearance can affect answers because most people assume certain characteristics of the interviewer based on those observations. Rank, status, and job position might influence the participant to be cautious in answering and/or to provide answers in an attempt to please the interviewer. If military personnel will be the interviewers, it might be helpful for them to wear civilian clothing when conducting the interview to limit any potential influence they may have on the participants. Interviewers could also influence the responses with their tone of voice or facial expressions or by chuckling at an answer, all of which may indicate agreement, disagreement, approval, or disapproval.

Do not interrupt when the participant is talking, even when the person pauses. Instead, jot down key words and go back to them when the person has finished. And do not rush them. Give them time to think after asking the question.

When the responses do not provide the needed data, probing may be required. Interviewer silence and gestures such as nodding or saying "uh-huh" may induce the person to elaborate more. Or you can ask: "In what ways?", "anything else?", "please continue". Probes must be neutral and every respondent should be given the same probe to the specific question (Babbie, 1973).

Be attentive to the person's non-verbal communication such as facial expressions, body movements, and eye contact during the interview that may help you understand the participant's true meaning and viewpoints. When the non-verbal cues indicate incongruence with what is being said, ask for clarification. (Frey and Oishi, 1995).

If you do not understand their answers or they use acronyms you do not know, ask for clarification and definitions.

The structured interview should not be so structured that it interrupts the flow of information and participant's thought processes. In unstructured interviews, the focus on the big picture issues and you should let the participants go where they want, based on their views. If the participant does not answer the questions you need, then ask the specific ones. The atmosphere should be one where the interviewer is there to learn from the experienced and knowledgeable participant or SME. In all types of interviews, the interviewer cannot allow the session to turn into a gripe session. It is the interviewer's job to keep the participant on track (see the section "Let Them Talk" below for techniques).

Be aware of how much has been covered and how much there is to cover in the allotted time. If you are running out of time, you may have to alter your strategy and identify the highest priority questions.

Recording

Most of us cannot write as fast as people talk, which means we need an effective method for recording the person's responses. As previously stated, you need to maintain a conversational momentum during the interview, yet you need to document the answers as soon as possible. This creates a dilemma and is easier said than done. If you pause frequently to record the answers in your notes, you lose the momentum and the participant may forget what he/she wanted to say next, and it conveys the message that you are not interested in everything he/she has to say. On the other hand, if you write while the person is talking, you can miss some of what the person is saying and you lose eye contact. Two options for dealing with this are not ideal: using two interviewers and tape recording. There is no one right method for every circumstance. The interviewer will have to decide what method to use given the circumstances.

Two Interviewers. Many professional interviewing companies use two interviewers. One interviewer asks the questions and the other records the answers. This is the most frequently recommended method; but, some respondents may be uncomfortable and feel intimidated with two interviewers. To minimize that possibility, the interviewer who is recording should be as unobtrusive as possible and not ask the questions. To help with the process of writing down the answers, you could use shorthand if you know it, or develop you own method that is similar to shorthand. For example use @ for "at", w/ for "with", bw for " between", and develop your own shortcuts.

Tape recording. Professional interviewers are divided on whether or not to use a tape recorder. The advantage to tape recording is that it captures everything you asked and what the participant said so you can listen later to fill in the places you may have missed with note taking or your notes are confusing. It allows you to focus on the questions and what the person is saying and not break eye contact. The main disadvantage is the reluctance of some people to be recorded. They are constantly aware of the recording device, which tends to inhibit the free flow of information. Some are completely distrustful of having their opinions recorded and they believe their answers will be shared with their leadership. Other considerations are the fidelity of the recording device and ambient noises which could affect the sound quality. It might be tempting to use a concealed recording device, but that is unethical. If you choose to record, you must first ensure it is allowed in the location where the interviews will occur and you must also obtain the participant's permission. Have a back-up plan in case the person you are interviewing refuses or is reluctant to be tape recorded.

Ending the Interview

At the end of the interview, take the time to go over your list of questions to ensure you asked everything you wanted to ask. Close the interview with a short summary of what the participant said. Do not end the interview abruptly because the person might feel that his/her contribution was not important. Finalize by expressing your appreciation and offering the person an opportunity to elaborate further by asking if he/she has additional comments. It is a good idea to send a letter of thank you or appreciation to the person you interviewed and his/her chain of command.

Special Situations and Other Considerations

This section presents some special situations you might encounter during an interview and other aspects to consider regarding the interviewing process.

Complex Issues or Questions. There may be situations when the participant is required to answer questions regarding a complex issue or the participant has to read something or have something explained to them before answering questions. This may be necessary because some participants may not be completely familiar with the topic/issue or may have inaccurate information. In those situations, the interviewer should summarize the issue using key words

before asking the specific questions. This will help to reduce error because all the participants being interviewed have been presented with the same keyword summary.

Recall Techniques. There are a few recall techniques that help the participant to not omit or forget essential information. After the participant responds to an open-ended question, you can present a list of other possible details, facts, or events in case he/she overlooked them. You can also provide them with a reference point when recalling an event or situation. For example, their first duty station, the time they were in Advanced Individual Training, their last promotion, etc. You could also use specific well-known events or timelines, such as major disasters, the holidays, prior to or after 9/11/2001. Just be sure to use the same reference points for everyone.

Interviewing Two Participants at One Time. Despite your diligent efforts to coordinate and confirm a schedule, sometimes things do not go as planned. You may be faced with having to interview more than one person at a time. This may happen because of scheduling errors despite your thoroughness in establishing a schedule. For example one person shows up at the wrong time but cannot wait or return later, the instructions and schedule you coordinated were misinterpreted or changed, several Soldiers show up at the same time, etc. The lead interviewer will have to be creative and determine if multiple people can be interviewed at the same time without that being detrimental to obtaining the needed information. Experience shows that interviewing two people at the same time is doable, but the interviewer's role will be to ensure that the views of both people are obtained and limit the influence one person might have over the other. Generally, interviewing more than two people becomes a focus group and there are different guidelines for conducting those (briefly discussed below). Whether you are interviewing two people at one time or conducting a focus group, the interview will take more time because you need to obtain the views of everyone. Check with both participants to see if a joint interview is acceptable and advise them it will take longer because you will need to get both of their responses. Furthermore, the interviewer needs to be comfortable interviewing multiple people at once.

Let Them Talk. You may encounter some participants who, after the first few questions, continue to talk and address all of the topics/issues you were going to cover without being prompted. In those situations, let them talk and proceed without interruption as long as the responses answer the questions you need answering. This is why it is important that you fully understand the subject matter and know the information you need to obtain. By interrupting, you are disrupting the person's train of thought and he/she might not remember what he/she wanted to tell you later in the interview. Also, you may gain useful information you did not anticipate or intend on collecting. If the person strays off topic, then you should find an appropriate time to interrupt to get him/her back on track. Let them finish a sentence and say something such as "I would like you to also tell me about___", or if you pick up on a keyword the participant used, you could mention that keyword and direct the interview back to the question related to that keyword. When this situation occurs, take the time during or after the interview to code the

responses to correspond to your interview questions in your notes. For example, you can use question numbers or brief topic headings.

Ending the Interview Early. You may need to end the interview early. Some people do not have the necessary verbal skills for you to understand what they are trying to tell you or they do not have the knowledge and experience required to answer the questions. It is best to not probe excessively or belabor an issue. If you feel you are not getting anywhere, gradually terminate the interview in a courteous manner (Frey and Oishi, 1995 and Siedman, 1998). You may also encounter participants that are dealing with an emotionally charged issue and need to vent[1]. Allow them to talk awhile and politely terminate the interview. To close the interview, say thank you, that is all the information I need from you.

Coding Interviewer Observations. If the interviewer makes observations, they should be coded or labeled in a way to distinguish them from the responses of the participant being interviewed. Otherwise, you run the risk of not knowing who said what and introducing error into the data because the observations may be inadvertently recorded as the person's responses.

Wait to Bin Responses. Avoid trying to categorize responses during the interview, unless you are using pre-determined responses. It is better to record the answers exactly as given as much as possible. You can re-phrase and code later. You won't know the best way to categorize the responses until you see all of them.

Large Sample Sizes Not Always Necessary. Striving for large sample sizes is not always the most prudent thing to do for some analyses. In some instances, it is better to "pick the brains" of a few SMEs who have the information and insights needed to inform the issues. If you strive to obtain a large sample size to make the statistics look good, you may end up obtaining the views of people whose knowledge and experience are limited and you run the risk of getting their best guesses. You could filter the responses out later assuming you collected the correct demographics, but that would take time and resources.

Focus Groups

This section will briefly cover only key issues of conducting focus groups because most of the guidelines on conducting interviews and asking questions also apply to focus groups and, there are several good books and articles that provide thorough guidelines (e.g., Krueger and Casey (2000). The purpose of a focus group is to obtain people's opinions and perspective on a specific topic. For example: the opinions of squad leaders on using a new piece of equipment in a large exercise/experiment. The ideal group size is between 6 and 10 participants and the focus group

[1] For example, during a focus group with Soldiers who had just returned from deployment, the interviewer determined (based on the Soldiers' responses) that it was more appropriate to allow the Soldiers to vent about their deployment than it was to try and force them to respond to the interview questions.

should last no longer than 90 minutes. There should be two moderators, one to ask questions and the other to record, whether with hand notes, a tape recorder, or with a computer.

It is better to have several homogenous groups than trying to get information from groups of people with vastly different backgrounds, roles, experiences, etc. Advantages with homogenous groups are that you will be able to organize the responses by the groupings you deemed important, you will understand the issue better, and it will take less time than with a heterogeneous group. Examples of homogenous groups are battalion commanders, company commanders, squad leaders, and S3s. In the Army, it is important to not mix higher ranked with lower ranked Soldiers because the lower ranked Soldiers will feel intimidated and be reluctant to voice their opinions.

Krueger and Casey (2000, p24) provide situations when focus groups should and should not be used.

Focus groups should be used to:

- Obtain a range of ideas
- Uncover or discern factors that influence opinions, behaviors, provide insight into complicated topics
- Obtain differing perspectives. Understand differences in perspectives.
- Obtain ideas and information you do not already have.
- Pilot test ideas, plans, etc.
- Obtain information to design a quantitative study.
- Shed light on quantitative data already collected.
- Capture the way people think about a topic and the language they use.

Focus groups should not be used to:

- Obtain a consensus.
- Educate people.
- Obtain sensitive information.
- Make statistical projections. (The sampling is not valid and there are not enough participants.)

Furthermore, focus groups are not appropriate when:

- The environment or topic is emotionally charged.
- The confidentiality of sensitive information cannot be ensured.
- Other methods can obtain the same information more economically or can produce better quality information.

The interviewer is a moderator who must be independent of the issue and not in a power position. The moderator follows a set of carefully predetermined questions and uses many of the

techniques used in interviewing presented above to elicit responses from the Soldiers or participants. The moderator sets the environment and ground rules to facilitate a free flowing discussion. The ground rules are:

1. Everyone should participate.
2. Respect others when they are talking.
3. Do not interrupt.
4. Everyone's experiences and opinions are important.
5. There are no wrong or right answers.
6. Do not attack others you disagree with.
7. Speak from your own experience.
8. Do not be afraid to present a differing opinion.
9. The goal of the focus group is to gain a deeper understanding of an issue and a wide range of perspectives and not to come to an agreement.
10. You may be called on if you haven't spoken in a while.
11. What is said here stays here.
12. Your responses are confidential. We do not identify anyone by name or any other identifiable trait in our report.

Offer light refreshments if you can, such as bottles of water, cookies, or candy. This will help make the participants feel comfortable. When collecting demographic information, use a survey sheet that each person fills out and hands in. It is a quicker way to collect that information and participants should not have to present that information in the group.

Start with introduction of yourself, purpose of focus group, how information will be used, informed consent, and confidentiality. Start with an easy, general question to help break the ice and get participants to feel comfortable. It could be helpful to have each participant briefly describe his/her job or role. Use discretion in asking them to provide their names (albeit, Soldiers' names are on their uniforms). Order questions from general to more specific. Most of the questions should be open-ended to obtain opinions perspectives. Use words and phrases such as "what", "how", "in what ways", "how do you feel", "describe your experiences". To obtain further elaboration, ask questions such as "could you tell me more about that", "please give me an example", "help me understand". As with interviewing one person, after the session ends, clarify your notes, fill in blanks, and label your notes and any tapes with dates, times, name of group.

As a moderator, you will have to manage the group dynamics and deal with the self-appointed expert, the dominator, the rambler, and those that are reluctant to speak. If some people do not participate, ask them individually for their opinion. The self-appointed expert and dominator tend to dominate the discussion and are typically the first person to respond to each question. Some give lengthy responses and interrupt others. To help deal with that, the moderator can direct a specific question to different person or say such things as "let's hear from someone else this time", "are there other views on that?", "remember, do not interrupt". For the rambler, the moderator can use the techniques discussed above for getting the person back on track.

If permitted and every participant agrees, using a tape recorder to record is the best method. Otherwise, when taking notes, assign letters or numbers to the participants to use when recording their answers. This will allow you to get a sense of the importance or frequency of a response. You want to know everyone's opinion and not get a consensus. If there is consensus, record that everyone was in agreement. You can also ask "does everyone agree with that?" "Are there any other views?"

Conclusions

The intent of this paper was to provide adequate, but, not exhaustive, guidelines on conducting interviews and focus groups for the purposes of collecting information for Army studies and analyses. The role of the interviewer is to collect the needed information in a way that assures the validity and reliability of the data. To assist the interviewer in fulfilling that role, this paper discussed the procedures, major aspects to consider, and some techniques to use

Author's Biography:

Dr. Patricia A. Kinney received her education in New Mexico obtaining her BA in Psychology from the University of New Mexico, followed by her MA in Experimental Psychology and her PhD in Educational Research from New Mexico State University. She began her government service in 1985 at the state level working as the Director of Statistics and Research for the NM Human Services Department. Her federal government employment began when she came to the TRADOC Analysis Center (TRAC) in May of 1987 as a GS-11. During her 27 years of service, Ms. Kinney has risen through the ranks to the grade of GS-13 and has provided invaluable support as the Project Lead or as the Senior Operations Research & Systems Analyst (ORSA) for many studies to include: NCO 2020 Survey Support for Institute for NCO Professional Development, TRADOC HQ; A cost and training requirements analysis on the Training Aids, Devices, Simulators, and Simulations (TADSS) required for the FCS spin out systems; A TEA on TADSS used to support FCS spin out system training; New Army Learning Model TEA- Battle Staff NCO Course portion; Assignment Oriented Training TEA; Army Reserve Expeditionary Force Training Requirements Analysis; MOS 92A TEA; Standard Army Retail Supply System (SARSS) TEA; Combat Service support Automation Office CSSAMO Training Analysis which identified the staffing and training needs to stand up "help desks" for CSS automation systems; CGSC pilot mentoring program analysis; AMSC /SSC training requirements study; and the Video Tele-training Reserve Component TEA. During her career, Dr. Kinney provided direct support to United States (U.S.) Army Soldiers through her extensive work in Training Effectiveness Analysis (TEA) and other Soldier centric analysis efforts, allowing her to gain valuable experience in interacting with this particular group of Subject Matter Experts (SME).

Appendix E
The Sheffield Elicitation Framework (SHELF) Overview

The Sheffield Elicitation Framework (SHELF, v2.0)

An "Off the Shelf" package for eliciting probability distributions

The SHELF package comprises a number of components:

1. This overview document, which should be read carefully before using SHELF.
2. Some pre-session briefing notes.
3. Blank templates and the same templates with added notes, for the following:
    a. Pre-session pro forma to be sent out with the briefing notes to experts;
    b. Elicitation record Part 1, to record the context and purpose of the elicitation;
    c. Elicitation record Part 2 (in several different forms), to record the elicitation of each probability distribution.
4. Software for fitting distributions using the R package, and instructions for its use.

What's new in version 2?

SHELF version 2.0 is a significant development from version 1, featuring:

- Completely revised and more powerful software for fitting distributions
- Templates for more than twice as many different elicitation methods
- Updated guidance in all documents

Elicitation and SHELF

Elicitation is the process of capturing expert knowledge about one or more uncertain quantities in the form of a probability distribution. It can be done informally, but when the expert judgements are sufficiently important it is necessary to employ a formal procedure in the interests of quality and defensibility. SHELF is such a formal procedure for elicitation.

But SHELF is more than this. Good elicitation generally requires a *facilitator* who has expertise in the process of elicitation. The facilitator guides the expert(s), manages the process and at the end delivers the elicited probability distribution. SHELF provides not only the tools for a facilitator but also copious advice on their use.

The developers of SHELF are co-authors of one of the leading textbooks in the field:

> "Uncertain Judgements: Eliciting Experts' Probabilities", by A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley and T. Rakow. Published in 2006 by Wiley. ISBN: 978-0-470-02999-2.

SHELF draws extensively on the research and conclusions for best practice in this book, but is also updated to reflect more recent experience.

By using SHELF, the facilitator can say that the elicitation has been conducted in an open and well-structured way that accords with best practice in the field. The SHELF name is a mark of quality.

The elicitation process

The facilitator needs to plan carefully in order to ensure a successful elicitation. The process begins with identifying the experts. SHELF is a framework for eliciting beliefs of one or more experts, such that if there are several experts they are brought together as a group. There are other frameworks for elicitation in which the experts are kept separate, but the SHELF view is that it is preferable to elicit their beliefs together as a group. A group should ideally not comprise more than about five experts: having too many experts will lead to unnecessarily long discussion, and it should be possible to cover all relevant expertise with just a few experts.

Having identified the experts and fixed a date when they can all attend, some preliminary *briefing material* should be sent to the experts well before the meeting. They should be told that they are expected to read this carefully in advance (and, where the experts are being paid for taking part, the payment should allow for the time to absorb this material). Briefing material may be short or may include substantial training documents (for instance concerning probability and its use to represent expert knowledge). Experts should be invited to contact the facilitator about any matters raised in the briefing.

SHELF is not prescriptive about the initial briefing material, but we emphasise that this is an important part of the process. SHELF includes a possible set of pre-session briefing material, in the form of a short set of briefing notes, a pro forma template and the same template with explanatory notes. The pre-session form is to be completed partly by the facilitator and partly by the expert, and is to be returned by the expert prior to the meeting. The briefing notes are designed to provide at least some minimal orientation for the expert. The form is intended to gather some information that may allow the facilitator to complete a part of the elicitation record in advance.

The elicitation session begins with housekeeping business for which a template "SHELF 1 (Context).doc" is provided in the SHELF package. Much of the ground covered in this first part of the elicitation session is extremely important. In particular, this document will note what orientation and training has been given. Training is essential, but depends on the experts. Some will need very little explanation of what probability is, for instance. Training that is additional to the orientation material is often given in this first part of the session, and handouts or slides used should be attached to the elicitation record. Unless the elicitation extends over several sessions involving the same experts, it is always advisable in this first part of the session to conduct at least one practice elicitation.

Quantities used for practice elicitation should be such that the facilitator knows the true value but the experts do not (so that the true value can be revealed at the end as part of the debriefing and discussion of that exercise). Although it is common to use general knowledge quantities such as the area of France, the population of Australia or the birth date of Tolstoy, there is much to be said for choosing practice quantities in the experts' area of expertise.

The first part of the session ends with identifying precisely the quantities whose probability distributions are to be elicited. This can be a complex exercise involving the technique of structuring, to express the quantities of primary interest in terms of others whose distribution(s) may be easier to elicit. The elicitation then moves into the second part, in which these distributions are elicited in turn. Each requires the completion of a "SHELF 2" document. SHELF accommodates several different protocols for eliciting a distribution, identified by the letters P (Probability), Q (Quartile), R (Roulette) and T (Tertile).

- In the P (Probability) method, the facilitator asks the experts for some specified probabilities.
- In the Q (Quartile) method the facilitator asks the expert(s) for their median and upper and lower quartiles.
- In the R (Roulette) method the facilitator asks the expert(s) to indicate their probabilities for ten ranges of values, known as bins, by placing chips in the bins.
- In the T (Tertile) method the facilitator asks the expert(s) for their median and upper and lower tertiles.

Also, the SHELF framework involves first eliciting individual distributions from each expert and then a group elicitation. Different methods may be used for the individual and group stages, so there are several forms of the "SHELF 2" document:

- In "SHELF 2 (Distribution) P.doc" the facilitator uses the P method for both individual and group elicitation.
- In "SHELF 2 (Distribution) Q.doc" the facilitator uses the Q method for both individual and group elicitation.
- In "SHELF 2 (Distribution) QP.doc" the facilitator uses the Q method for the individual elicitation but the P method for group elicitation.
- In "SHELF 2 (Distribution) R.doc" the facilitator uses the R method for both individual and group elicitation.
- In "SHELF 2 (Distribution) RP.doc" the facilitator uses the R method for the individual elicitation but the P method for group elicitation.
- In "SHELF 2 (Distribution) T.doc" the facilitator uses the T method for both individual and group elicitation.
- In "SHELF 2 (Distribution) TP.doc" the facilitator uses the T method for the individual elicitation but the P method for group elicitation.

There are blank templates for all these elicitation record documents, and there are also the same templates with added explanatory notes. In addition to describing what should be entered in each part of the template, these notes include (a) advice [in square brackets] to the facilitator on carrying out the relevant task, and (b) notes *in italics* on the rationale for this task, and how it contributes to good elicitation.

The facilitator is free to choose which SHELF 2 protocol to use. We have tried in version 2.0 to offer all the versions that we think can lead to good elicitation. However, it is worth saying that our current favourites are QP, R, RP and TP. There are some comments in the annotated templates comparing the protocols.

Time in the elicitation session itself is at a premium, so the facilitator should be well prepared. In particular, he/she should have blank templates (SHELF 1 and a SHELF 2 for each distribution to be elicited) with basic data filled in already where possible. He/she will also find it useful to have hard copy printouts of the annotated versions of the forms, to refer to during the session.

The forms should be filled in live during the session, and preferably projected so that all participants can see the record as it is built up. The facilitator will find it helpful to have an assistant for this purpose, and/or to compute fitted distributions.

Techniques

The SHELF protocols require distributions to be fitted, distributions to be plotted and feedback data to be computed in real-time. It is essential that the facilitator ensures that he/she has access to suitable software, and is fluent in its use. SHELF includes some basic software in the form of procedures in the R language, together with notes on their use, but the facilitator or his/her assistant may use whatever they find best suited to the task and/or their skills.

Ideally, graphs of fitted distributions should form part of the elicitation record. The facilitator should also check that he/she knows how to produce graphs and to paste them into the SHELF 2 documents (or can produce graphs as attachments to those documents).

It goes without saying, though, that the facilitator's most important technique is simply the ability to manage the process and the interaction between the experts. The notes for the various SHELF documents try to highlight things that the facilitator needs to watch out for, but there is no substitute for practical elicitation experience.

Using SHELF with one expert

Although a formal and careful elicitation process such as SHELF is generally used for elicitation from several experts, it is straightforward to adapt the SHELF materials for use with a single expert. The elicitation should be recorded on a "SHELF 1" form and either P, Q, R or T "SHELF 2" forms. All of the fields in the "SHELF 1" form are relevant. In the "SHELF 2", leave blank the "Group elicitation" field, but the subsequent feedback (and possible iteration) remains important.

The R software routines include functions to assist with single expert elicitation, including a probability-based method that is not used in multi-expert elicitation.

About SHELF

This is only the second version of the Sheffield Elicitation Framework. It will continue to evolve in response to the experiences of people using it and according to the wish of its developers to extend its capabilities.

Comments are welcomed by Tony O'Hagan (shelf@tonyohagan.co.uk) and Jeremy Oakley (j.oakley@sheffield.ac.uk).

We would particularly welcome offers of additional materials or suggested amendments to components of SHELF.

The SHELF package is available from the website http://tonyohagan.co.uk/shelf/.

Copyright

The appropriate form of citation in published work is:

Oakley J. E. and O'Hagan, A. (2010). SHELF: the Sheffield Elicitation Framework (version 2.0), School of Mathematics and Statistics, University of Sheffield, UK. (http://tonyohagan.co.uk/shelf)

Appendix F
References

[1] MAJ Christopher Marks, Ms. Kristen Smead, and LTC Jonathon Alt; *Enhancing Subject Matter Expert Elicitation Techniques*. Tech. rep. TRAC-M-TR-13-048. 700 Dyer Road Monterey, California 93943: TRADOC Analysis Center - Monterey, 2013.

[2] Ralph L. Keeney. *Value-Focused Thinking: A path to creative decisionmaking*. Harvard University Press, 2009.

[3] Craig W Kirkwood. *Strategic Decision Making*. Duxbury Press Belmont, CA, 1997.

[4] Lawrence D. Phillips and Carlos A. Bana e Costa. "Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing". English. In: *Annals of Operations Research* 154.1 (2007), pp. 51–68. ISSN: 0254-5330.

[5] Paul L Ewing Jr, William Tarantino, and Gregory S Parnell. "Use of decision analysis in the army Base Realignment And Closure (BRAC) 2005 military value analysis". In: *Decision Analysis* 3.1 (2006), pp. 33–49.

[6] Claude Elwood Shannon. "A mathematical theory of communication". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001), pp. 3–55.

[7] Jianhua Lin. "Divergence measures based on the Shannon Entropy". In: *Information Theory, IEEE Transactions on* 37.1 (1991), pp. 145–151.

[8] Lou Jost. "Entropy and diversity". In: *Oikos* 113.2 (2006), pp. 363–375.

[9] DB De Araujo et al. "Shannon Entropy applied to the analysis of event-related fMRI time series". In: *NeuroImage* 20.1 (2003), pp. 311–317.

[10] Lotfi A Zadeh. "Fuzzy sets". In: *Information and control* 8.3 (1965), pp. 338–353.

[11] Glad Deschrijver and Etienne E Kerre. "On the relationship between some extensions of fuzzy set theory". In: *Fuzzy sets and systems* 133.2 (2003), pp. 227–235.

[12] HJ Zimmermann. *Fuzzy Set Theory and Its Applications Second, Revised Edition*. Springer, 1992.

[13] William J Tastle and Mark J Wierman. "Consensus and dissention: A measure of ordinal dispersion". In: *International Journal of Approximate Reasoning* 45.3 (2007), pp. 531–545.

[14] William J Tastle and Mark J Wierman. "Using consensus to measure weighted targeted agreement". In: *Annual Meeting of the North American Fuzzy Information Processing Society, 2007*. IEEE. 2007, pp. 31–35.

[15] William J Tastle and Mark J Wierman. "Consensus and dissention: a new measure of agreement". In: *Annual Meeting of the North American Fuzzy Information Processing Society, 2005.* IEEE. 2005, pp. 385–388.

[16] William J Tastle, Jack Russell, and Mark J Wierman. "A new measure to analyze student performance using the Likert scale". In: *Information Systems Education Conference: Proceedings of ISECON.* 26 (2005), p. 2007.

[17] W. N. Venables, D. M. Smith, and the R Core Team. *An Introduction to R.* Version 3.0.3. CRAN. 2014.

[18] Dr. Michael R. Anderson and Mr. Eric E. Johnson. *Multi-Attribute Decision Making (MADM) and Associated Assessment Techniques in Support of Army Studies and Analyses, Code of Best Practices (COBP) (Powerpoint presentation).* TRADOC Analysis Center Methods and Research Office. Powerpoint presentation. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center Methods and Research Office, 2009.

[19] Mary A. Meyer and Jane M. Booker. *Eliciting and Analyzing Expert Judgement, A Practical Guide.* 3600 University City Science Center, Philadelphia, PA 19104-2688: Society for Industrial and Applied Mathematics, 2001. ISBN: 0-89871-474-5.

[20] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. "Statistical methods for eliciting probability distributions". In: *Journal of the American Statistical Association* 100.470 (2005), pp. 680–701.

[21] Tony O'Hagan. "Elicitation". In: *Significance* 2.2 (2005), pp. 84–86.

[22] Bilal M Ayyub. "A practical guide on conducting expert-opinion elicitation of probabilities and consequences for Corps facilities". In: *Institute for Water Resources, Alexandria, VA, USA* (2001).

[23] A. O'Hagan et al. *Uncertain Judgements: Eliciting Experts' Probabilities.* Statistics in Practice. Wiley, 2006. ISBN: 9780470033302.

[24] Gregery S. Parnell, Patrick J. Driscoll, and Dale L. Henderson, eds. *Decision Making in Systems Engineering and Management.* Second. Hoboken, NJ: John Wiley & Sons, Inc., 2011.

[25] Ms. Lynn Leath. *Study Directors' Guide; A Practical Handbook for Planning, Preparing, and Executing a Study.* Tech. rep. TRAC-F-TM-09-023. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2013.

[26] Michael F. Bauman et al. *Measurement Space Code of Best Practice (CoBP).* Tech. rep. TRAC-H-TM-12-034. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2012.

[27] Ph.D. Brenda M. Wenzel and Ph.D. Kathy L. Nau. *Code of Best Practices for Survey Efforts.* Tech. rep. TRAC-W-TM-12-001. Martin Luther King Drive, White Sands Missile Range, NM 88002-5502: TRADOC Analysis Center, White Sands Missile Range, 2011.

[28] David Jenkinson. "The elicitation of probabilities-a review of the statistical literature". In: *Bayesian Elicitation of Experts' Probabilities (BEEP) working paper* (2005).

[29] Mary Kynn. "The 'heuristics and biases' bias in expert elicitation". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171.1 (2007), pp. 239–264.

[30] Glenn G Shephard and Craig W Kirkwood. "Managing the judgmental probability elicitation process: a case study of analyst/manager interaction". In: *Engineering Management, IEEE Transactions on* 41.4 (1994), pp. 414–425.

[31] GR Chesley. "Elicitation of subjective probabilities: a review". In: *The Accounting Review* 50.2 (1975), pp. 325–337.

[32] Leonard J Savage. "Elicitation of personal probabilities and expectations". In: *Journal of the American Statistical Association* 66.336 (1971), pp. 783–801.

[33] Morris H DeGroot. "Reaching a consensus". In: *Journal of the American Statistical Association* 69.345 (1974), pp. 118–121.

[34] Mr. Adam Rubemeyer et al. *Capability Gap Assessment; Blending Warfighter Experience with Science*. Tech. rep. TRAC-F-TR-13-022. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2013.

[35] J.A. Rice. *Mathematical Statistics And Data Analysis*. Third. Advanced series. Brooks/Cole CENGAGE Learning, 2007. ISBN: 9780534399429.

[36] MAJ Matther Dabkowski et al. *Force Design/Force Mix: Building the Best Army Possible with Reduced End-Strength*. Tech. rep. TRAC-F-TR-11-020. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2011.

[37] MAJ Tom Deavens et al. *Support for the Expeditionary Military Intelligence Brigade Commanders' Assessment Workshop*. Tech. rep. TRAC-M-TR-13-029. 700 Dyer Road, Monterey, CA 93943: TRADOC Analysis Center, Monterey, 2013.

[38] Ms. Michele Wolfe. *Operational Risk Analysis, A New Approach*. Tech. rep. TRAC-F-TR-13-026. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2013.

[39] Headquarters, Department of the Army. *FM 5-19: Composite Risk Management*. Government Printing Office.

[40] Rebecca A O'Leary et al. "Comparison of three expert elicitation methods for logistic regression on predicting the presence of the threatened brush-tailed rock-wallaby Petrogale penicillata". In: *Environmetrics* 20.4 (2009), pp. 379–398.

[41] *Title 32–National Defense; Code of Federal Regulations 219, Protection of Human Subjects (32 CFR 219)*. Government Printing Office. 2013. URL: `http://www.ecfr.gov/cgi-bin/text-idx?c=ecfr&tpl=/ecfrbrowse/Title32/32cfr219_main_02.tpl`.

[42] Dr. Jennifer Jebo, Ms. Sara Krondak, and Ms. Amy McGrath. *TRADOC Analysis Center Human Research Protection Program Plan*. Tech. rep. TRAC-L-TM-13-028. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345: TRADOC Analysis Center, Fort Leavenworth, 2013.

[43] *Institutional Review Board for the Protection of Human Subjects*. Naval Postgraduate School website. 2013. URL: http://www.nps.edu/research/IRB.htm.

[44] Tim Bedford, John Quigley, and Lesley Walls. "Expert elicitation for reliable system design". In: *Statistical Science* (2006), pp. 428–450.

[45] Joseph Kadane and Lara J Wolfson. "Experiences in elicitation". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 3–19.

[46] Mohammed Abdellaoui. "Parameter-free elicitation of utility and probability weighting functions". In: *Management Science* 46.11 (2000), pp. 1497–1512.

[47] Paul A Bottomley and John R Doyle. "A comparison of three weight elicitation methods: good, better, and best". In: *Omega* 29.6 (2001), pp. 553–560.

[48] TRADOC Analysis Center. *Constraints, Limitations, and Assumptions, Code of Best Practice*. Technical memorandum TRAC-H-TM-12-033. 255 Sedgwick Avenue, Fort Leavenworth, KS 66027-2345, 2012.

[49] J Martin Bland and Douglas G Altman. "Statistics notes: Cronbach's alpha". In: *BMJ* 314.7080 (Feb. 1997), p. 572. DOI: 10.1136/bmj.314.7080.572.

[50] Mohsen Tavakol and Reg Dennick. "Making sense of Cronbach's alpha". In: *International Journal of Medical Education* 2 (2011), pp. 53–55.

[51] Neal Schmitt. "Uses and abuses of coefficient alpha". In: *Psychological assessment* 8.4 (1996), pp. 350–353.

[52] Robert Ferber, ed. *Handbook of Marketing Research*. New York: McGraw-Hill Book Company, 1974.

[53] Anthony O'Hagan. "Eliciting expert beliefs in substantial practical applications". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 21–35.

[54] Paul H Garthwaite and James M Dickey. "Quantifying expert opinion in linear regression problems". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1988), pp. 462–474.

[55] Joseph B Kadane et al. "Interactive elicitation of opinion for a normal linear model". In: *Journal of the American Statistical Association* 75.372 (1980), pp. 845–854.

[56] Lawrence M. Leemis. *Probability*. Lightning Source Incorporated, 2011. ISBN: 978-0982-91740-4.

[57] Bradley Efron and Robert Tibshirani. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy". In: *Statistical science* (1986), pp. 54–75.

[58] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. New York: Springer Science, 2001. ISBN: 978-0387-95284-0.

[59] Ronald D. Fricker Jr., Walter W. Kulzy, and Jeffrey A. Appleget. "From Data to Information: Using Factor Analysis with Survey Data". In: *Phalanx* 45 (4 2012), pp. 30–34.

[60] MAJ Thomas Deveans et al. *Africa Knowledge, Data Source, and Analytic Effort (KDAE) Exploration*. Tech. rep. TRAC-M-TR-12-037. 700 Dyer Road, Monterey, CA 93943-0692: TRADOC Analysis Center, Monterey, 2012.

[61] A. Gelman et al. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2003. ISBN: 9781420057294.

This page intentionally left blank.

Appendix G
Glossary

ANOVA	Analysis of Variance
ARCIC	U.S. Army Capabilities and Integration Center
CLA	Constraints, Limitations, and Assumptions
CNA	Capability Needs Analysis
COE	Center of Excellence
COA	Course of Action
CON	Measure of Consensus
CPR	Capability Portfolio Review
CSA	Chief of Staff of the Army
CV	Coefficient of Variance
DoD	Department of Defense
Dom	Dominance
EEA	Essential Elements of Analysis
Eff	Efficiency
Exp	Expeditionary
FoA	Freedom of Action
MCDA	Multi-Criteria Decision Analysis
MU	Arithmetic Mean
NLT	No Later Than
PMJ	Professional Military Judgment
RC	operational risk
SD	Standard Deviation
SHELF	The Sheffield Elicitation Framework
SME	Subject Matter Expert
S & T	Sceince and Technology
SW	Shapiro-Wilkes Test
TOE	Table of Organization and Equipment
TRAC	Training and Doctrine Command Analysis Center
TRAC---LEE	U.S. Army Training and Doctrine Command Analysis Center---Fort Lee
TRAC---FLVN	U.S. Army Training and Doctrine Command Analysis Center---Fort Leavenworth
TRAC---MTRY	U.S. Army Training and Doctrine Command Analysis Center---Monterey
TRAC---WSMR	U.S. Army Training and Doctrine Command Analysis Center---White Sands Missile Range
TRADOC	U.S. Army Training and Doctrine Command
TWV	Tactical Wheeled Vehicle
VFT	Value Focused Thinking
WfF	War-fighting function