TRAINING SUBJECT MATTER EXPERTS TO MAKE BETTER ESTIMATES

THESIS

Cameron R. Scott, 1st Lieutenant, USAF

AFIT-ENV-MS-21-M-268

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENV-MS-21-M-268

TRAINING SUBJECT MATTER EXPERTS TO MAKE BETTER ESTIMATES

THESIS

Presented to the Faculty

Department of Systems Engineering and Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Cost Analysis

Cameron R Scott, BS

1st Lieutenant, USAF

March 2021

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENV-MS-21-M-268

TRAINING SUBJECT MATTER EXPERTS TO MAKE BETTER ESTIMATES

Cameron R. Scott, BS

1st Lieutenant, USAF

Committee Membership:

Dr. R. David Fass
Chair

Dr. Jonathan D. Ritschel
Member

Dr. John J. Elshaw
Member

AFIT-ENV-MS-21-M-268

Abstract

Subject Matter Expert (SME) Elicitation is commonly used as a cost-estimation method when no historical data exists. The elicitation process assumes Subject Matter Experts are the resident experts in their individual fields of study. In addition, the process assumes SMEs provide reasonably trustworthy cost estimates that follow the tenets of probability theory. This last assumption is the subject of this study. A Navy sponsored e-learning course was created to help train SMEs in understanding probability distributions (O'Hagan, 2019). We conducted a study hypothesizing that taking the e-learning course would improve study participant accuracy in estimating confidence intervals for probability distributions. We found strong support for our hypothesis and significance in our results providing a 26% increase in the estimation accuracy of our SMEs. The results of the study show that training SMEs has the potential to improve SME estimation skills.

Acknowledgments

I would first like to thank my research advisor, Dr. Robert D, Fass, for his valuable guidance, support, and mentorship throughout this process. I truly enjoyed working with you and I could not have completed this milestone without you. I would also like to sincerely thank my thesis committee for their insight, constructive criticism and encouragement. Lastly, I would like to thank my classmates, friends, and family for being incredibly supportive.

Cameron R. Scott

Table of Contents

Page

List of Figures

List of Tables

TRAINING SUBJECT MATTER EXPERTS TO MAKE BETTER ESTIMATES

Taking a ride on a fire truck does not make a person a fireman. Yet, with increasing frequency, subject matter experts (SMEs) are, metaphorically, doing just that, and taking us all for a ride in the process (Lavin, Deryfus, Slepski, & Kasper, 2007, p. 1)

I. Introduction

Background

There are many different methods of cost estimation used in the Department of Defense (DoD). These methods include: Analogy/Factor, Parametric, Engineering Build-Up, and Subject Matter Elicitation. Subject Matter Elicitation is commonly used if sparse historical data exists. The quote above describes a problem that pervades expert elicitation. Some SMEs may lack the expert knowledge one would think they should possess. Miller (2020) conducted research on subject matter elicitation in cost risk analysis. He compared the efficacy of SME elicitation methods to other cost estimation methods and critically examined SME elicitation methods used within the DoD. (Miller, 2020) Nevertheless, the elicitation process assumes Subject Matter Experts (SME) are the resident experts in their individual fields of study. In addition, the process assumes SMEs provide reasonably trustworthy estimates that follow the tenets of probability theory. This last assumption is the subject of this study.

Problem Statement

Not all SMEs are created equal, of course. The level of experience, training, and other factors vary among SMEs. Nevertheless, the DoD frequently relies on SME elicitation for estimates. Although SMEs are experts in their fields, they are usually not experts in probability theory or cost estimation. This leads to the question: Can we train SMEs to make better estimates?

This topic will be explored throughout the paper. A Navy sponsored e-learning course was created to help train SMEs in understanding probability distributions (O'Hagan, 2019). A study was conducted using this course to determine if SMEs can be trained to make better estimates.

Research Questions

1. How proficient is the average SME in making estimates?
2. How will probability training affect SME estimation accuracy?

Methodology

Cost estimators rely upon Subject Matter Experts. However, SME's lack of understanding and application of probability distributions can lead to significant errors. The purpose of this research is to examine the ability of SMEs to apply probability distributions to real world examples. The data was collected through a repeated measures study. This method used only one group of participants who completed a pre-test, an e-learning course, and a posttest. Response data was collected from ten "calibration" questions and scored on a scale from 1-10 correct answers. We define this value as *estimation accuracy* for this study. Once training and post-test are complete the final data will be analyzed to see if there was improvement in the estimation skills of our SMEs.

Research Contribution

This research used repeated measures to determine if SMEs can be trained to make better estimates. If successful in the pilot study with non-SME subjects, future research may recreate the study on a larger scale with actual SMEs. Ultimately, this could lead to probability training across the Air Force, improving SME performance in cost estimation tasks.

II. Literature Review

Chapter Overview

The purpose of this chapter is to provide an in-depth review of literature related to the Subject Matter Expert (SME) elicitation process the DoD employs. Subject Matter Experts can be somewhat overconfident and sometimes even underconfident when it comes to making their estimates (O'Hagan, 2019). Subject Matter Experts (SMEs) are known to be the "resident" experts in their fields and use different techniques when providing estimates. The literature on elicitation is extensive, with important contributions made in fields as diverse as statistics, psychology, management science, economics, and environmental science. (Bellarosa & Chen, 1997; Frick, 2010; Kahneman, 2011; Kilo, 1999; O'Hagan, 2019; Tversky & Kahneman, 1974; Anderson & Elloumi, 2004). This review will cover five main topics: 1) Cost Estimating Techniques 2) Uncertainty in the DOD, 3) The Purpose of Subject Matter Experts, 4) SME Elicitation Bias, and 5) Reliability of Subject Matter Experts.

Cost Estimating Techniques

Although it is the main focus of this paper, SME elicitation is not the only cost estimating technique available. There are four primary cost estimating methods. The four cost estimating methods are: 1) Analogy/Factor method, 2) Parametric method, 3) Engineering build up, and 4) Subject Matter Elicitation. The estimating technique chosen for an estimate can significantly influence both how long it takes to develop the estimate as well as its accuracy (Air Force Cost Analysis Agency, Air Force Cost Analysis

Handbook (AFCAH), 2008, p. 27). Each cost estimating method comes with its own set of advantages and disadvantages.

The Analogy/Factor method takes data from past programs with similar characteristics and uses that as a basis for the analogy. The cost of the new system is then estimated by adjusting historical costs of the existing system to account for the differences in the new system. One advantage of the Analogy/Factor method is its ability to be used early in a program before detailed requirements are known. Only changes or differences between the new programs have to be estimated. There is not extensive data collection required. The Analogy method can also be combined with SME elicitation to make a completed factor, showing the interweaving of different cost estimating methods. The Analogy/Factor method is also difficult to refute when there is a strong similarity of traits between the historical and new program. Finally, the Analogy/Factor method also provides quick crosscheck estimates. These methods help give the analogy credibility (Air Force Cost Analysis Agency, Air Force Cost Analysis Handbook (AFCAH), 2008).

The Analogy/Factor method also has its disadvantages. The historical and new programs must have a similar scope/ effort. It also requires functional knowledge to estimate adjustments to baseline data. The data required may also be hard to find (Air Force Cost Analysis Agency, Air Force Cost Analysis Handbook (AFCAH), 2008). When the analogy process is completed incorrectly the entire project can lose its credibility.

The parametric method makes use of regressions and other statistical methods to develop an equation known as the Cost Estimating Relationship (CER). The Cost Estimating Relationship estimates a cost element by creating a relationship between one or more independent variable. The advantages of the parametric method include the ability to adjust estimates by modifying the input parameters. Statistical information is provided for estimate uncertainty and risk analysis. The statistical information also provides objective measures of estimating validity. It also allows for flexibility in the expertise of the analyst. This is because one does not need to be a technical expert to apply parametrics (Air Force Cost Analysis Agency, Air Force Cost Analysis Handbook (AFCAH), 2008).

When a cost estimator is not able to break down an estimate into the individual component costs a “black box” can be created (Air Force Cost Analysis Agency, Air Force Cost Analysis Handbook (AFCAH), 2008). The “black box” can cause an otherwise defendable estimate to lose its credibility due to the inability to break down the estimate to individual component costs. This is a clear disadvantage of the parametric method. Parametric equations must be constantly updated, validated, and calibrated for use in order to reflect current technology and programs. Even though they are easy to use, analysts still must be trained to use parametric equations and tools. The parametric method also requires functional and SME judgement to identify and adjust parametric inputs.

The Build-up method is also referred to as engineering or “grass roots” estimating. Through this method a detailed accurate basis of estimate is provided for a program. Build up estimates typically are based on detailed engineering information about the system/item being produced (Air Force Cost Analysis Agency, Air Force Cost Analysis Handbook (AFCAH), 2008). This method is typically implemented in late development and production phases. An advantage of the Build-up method is the detailed estimate that the technique tends to produce. It also accurately depicts the production process. The estimate tends to be based on recent actual cost data (Air Force Cost Analysis Agency, Air Force Cost Analysis Handbook (AFCAH), 2008).

Unfortunately, the Build-up method is prone to double counting and often omits the lower level Work Breakdown Structure (WBS) elements. The estimate provided is also data intensive and brings a heavy workload to the estimating team that requires constant maintenance and updating; which can be very time consuming. Expert knowledge of the subjects estimated is also expected of the analysts (Air Force Cost Analysis Agency, Air Force Cost Analysis Handbook (AFCAH), 2008).

Subject Matter Elicitation (also known as expert judgment) is the final cost estimating technique. Subject Matter Elicitation is the expression of expert knowledge through uncertainties and probabilities. Subject Matter Elicitation uses SME experience to develop estimates when detailed data is unavailable. It also is very useful for filling gaps in the data that are needed for other estimating methods. Subject Matter Elicitation can also be used to crosscheck other estimates.

The estimate's accuracy depends on the credibility of the SME. Documentation of recommendations and the decision process is the only defense the estimate has to stand on. The process of interviewing SMEs also tends to be time consuming. Bias is also a factor that must be considered when eliciting knowledge from a SME. Some Subject Matter Experts may not have a detailed data base of historical efforts that can help them to provide informed estimates.

Uncertainty in the DoD

In the Department of Defense (DoD), “uncertainty is an inherent, unavoidable aspect of life that has a significant impact on program or project management, and acquisition in general" (Frick, 2010, p. 1). The Department of Defense takes a logical, repeatable, and auditable approach to risk management. The DoD has access to many guidebooks that are used to help quantify risk and uncertainty for cost estimates. Two of these are the DoDI 5000.73 Cost Analysis Guidance and Procedures (2017) and the Joint Agency Cost Schedule Risk and Uncertainty Handbook (JA CSRUH) (2014).

There was a need for a set of “best” practices and techniques for estimating uncertainty, which led to the creation of the JA CRUH. The JA CRUH is used across multiple government agencies (DoD and NASA to name a few) and helps analysts apply risk and uncertainty to cost estimates. The DoDI 5000.73 and JA CRUH both have the same goal which is “to define and clearly present simple, well-defined cost risk and uncertainty analysis processes that a repeatable, defendable, and easily understood” (Naval Center for Cost Analysis, 2014, p. 1).

Both handbooks provide information about the Work Breakdown Structure. The WBS normally has three primary hierarchical levels. Sometimes a fourth and fifth level can be included if necessary (Department of Defense, 2018). The first level consists of the system or material item as a whole. The second level tends to include: integration and assembly, system test and evaluation(ST&E), system engineering/Program management(SE/PM), common support equipment(CSE), peculiar support equipment(PSE), training, data, operational/site activation, and initial spares and repair parts (Department of Defense, 2018). Level three breaks down each category from level two into more individual parts.

Purpose of Subject Matter Experts

One way in which expert opinion and judgment enters into statistical inference and decision-making is through expert knowledge elicitation. Elicitation in this context is the process of expressing expert knowledge in the form of probability distributions for uncertain quantities (O'Hagan, 2019). In the elicitation process, the expert knowledge comes from SMEs. SME estimates are often used in program offices. Program offices are used by government agencies to maintain the standards for program management throughout the organization. Subject Matter Expert estimates play a large role in the completion or the failure of a project. If utilized to complete a project, the result of using a true SME should be a noticeable reduction in the cost and development time (Pace, 2002). This is why it is important to ensure that the SMEs chosen for the elicitation process are reputable. A SME is:

> an individual who, by virtue of position, education, training, or experience, is expected to have greater-than-normal expertise or insight relative to a particular technical or operational discipline, system, or process, and who has been selected or appointed to participate in development, verification, validation, accreditation, or use of a model or simulation (Pace, 2002, p. 3).

An issue that program offices may face regarding SME elicitation is selecting the “right” SME with the proper expertise. SME expertise should encompass three domains:

structure, process, and outcome (Lavin, Deryfus, Slepski, & Kasper, 2007). Structure can be described as an expert's knowledge of her discipline. Process is the way the SMEs expand their knowledge and verify their information. Outcome is the combination of both process and structure. A SME who can combine all of these domains together would be seen as a reputable SME. There are different purposes SMEs can be utilized for when providing knowledge. "Teacher" SMEs will direct instructional meetings about the project. In these meetings SMEs will facilitate learning by interjecting comments when they know a fact that the group would not (Lavin, Deryfus, Slepski, & Kasper, 2007). These types of SMEs are ideal when information about an unknown subject is needed. They are essential in increasing the group's knowledge of a topic and help in the research of a specific aspect of the project.

SMEs can also organize activities in such a way that the other members of the group construct content in their own minds through personal contexts, rather than giving them cut and dried facts to memorize (Anderson & Elloumi, 2004). This method of elicitation encourages the members of the group to think outside of the box. Members can use the knowledge provided by the SME to come to conclusions they ordinarily would not think about. Both examples are showcases of the versatility of SMEs.

SME Elicitation Bias

As with many processes in cost estimation, bias tends to find a way to creep into SME elicitation. There have been studies conducted that took a deeper examination into these biases and their effect on the DOD. Tversky and Kahneman (1974) laid the groundwork for the heuristics and bias research program. Tversky and Kahneman found that these heuristics can lead to systematic errors of judgement, also known as biases in our more complex, everyday tasks. (O'Hagan, 2019). These biases identified in Tverky and Kahneman's research led to problems in the DoD with our SME elicitation.

People's judgments are often made on the basis of heuristics, which are quick, short-cut reasoning processes (O'Hagan, 2019). Kahneman (2011) introduces the two systems of thinking. System one is more of one's subconscious thought. It is defined as fast, unconscious, automatic and effortless (Kahneman, 2011). It can be thought of as our subconscious "scapegoat", when there is not enough time for prolonged thought. This can also be described as our "gut" feeling. For example, we tend to almost instinctively stay away from things that can cause us serious danger without even thinking about it. However, the instinctive use of heuristics can introduce systematic bias in our more complex tasks (O'Hagan, 2019; Tversky & Kahneman, 1974). The solution to this natural bias is to slow down and take the time to think. By doing this we invoke what Kahneman describes as using our system two. This is defined as slow, deliberate and conscious, effortful, controlled mental process with rational thinking. It seems like an easily performable task; however it only makes 2% of our entire thinking process (Kahneman, 2011). Some of the more common heuristics used in elicitation are Anchoring, Availability, Range-Frequency, and Overconfidence.

Anchoring is often used alongside the term "adjustment." "Anchoring and Adjustment" is when a readily available value known as the "anchor" is used as a starting point for a numerical judgement (O'Hagan, 2019). This number is then adjusted from the anchor in order to form a judgement. This judgement is inherently biased toward the anchor. If there are two successive judgements made, the first judgment typically becomes the “anchor.”

Bias is introduced when the individual fails to sufficiently adjust from her original “anchor.” It is common for the individual to make insufficient adjustments to the known value; resulting in the estimate becoming “anchored” to the known value (Kahneman, 2011). An estimating technique that is prone to anchoring is the analogy method. Experts can fail to fully adjust for the change in complexity between the historical analogy and the new effort (Frick, 2010).

Availability bias is a mental shortcut relying on immediate examples that come to an expert’s mind when providing estimates (O'Hagan, 2019). Experts who tend to fall victim to this type of estimation bias tend to base their estimates on information that is easiest to recall. For example, if the media have reported several fatal boating accidents recently, an expert may deem boating accidents to be very frequent and have a high fatality rate. In contrast, car crashes happen far more frequently but may not be broadcasted as much on the news.

O’Hagan (2019) observes that when experts are asked about an uncertain quantity, their judgements will be more influenced by evidence they can readily bring to mind. Personal experiences and recent events tend to have a heavy influence on experts when advising. Any information that is not in the forefront of these experts’ minds is more than likely to be passed over. Overconfidence is another elicitation bias that can cause issues for SMEs.

It is often said that experts are typically overconfident in their judgments.

The evidence for this is primarily from studies in which subjects were asked

> to give an interval of values for an uncertain quantity with a specified probability, and where the frequency with which the true values fell within those intervals was less than the specified probability. For instance, in studies where subjects gave 95% probability intervals, fewer than 95%, and perhaps as few as 65% of those intervals were found to contain the corresponding true values. The intervals exhibited overconfidence. (O'Hagan, 2019, p. 71).

SMEs who succumb to overconfidence bias tend to focus on everything that will go right and believe nothing will go wrong. It has been seen that overconfidence may be related to anchoring. Subjects are often asked for an interval after they have first given an estimate, which can serve as an anchor. Then, the interval being too narrow maybe due to insufficient adjustment from the anchor (O'Hagan, 2019).

Miller (2020) found that heuristics applied can be summarized in nine steps: 1) Have historical minimum, maximum, and averages on hand, 2) use multiple experts, 3) ask the expert for an upper and lower value, 4) encourage a dialog to identify various possible outcomes thus far, 5) seek the most-likely value near the end of the step for discussion, 6) select a distribution, 7) treat the SMEs input at the 70% interval, 8) Crosscheck information and challenge SMEs against historical experience, and 9) iterate the evolving conclusions with the experts as needed. (Miller, 2020)

Reliability of Subject Matter Experts

The process of expert elicitation is by no means perfect. The introduction of the different types of biases paired with SME level of expertise and other factors can hurt the reliability of expert elicitation. Biases or conflicts of interest will discredit a SME and essentially make their work suspect.

Lavin (2007) found only 55% of SMEs have actually done any sort of research within their field of expertise. The problem stems from employers. SMEs tend to be hired for their ability to state what is known about a topic, not necessarily for their true expertise in the field. Employers expect their SMEs to possess some basic knowledge in their area, but not complete expertise.

SMEs have been known to develop their expertise through either formal training or on-the-job (OJT) experience (Lavin, Deryfus, Slepski, & Kasper, 2007). There are pros and cons to both types of learning. Formal training will give SMEs a strong understanding on the theoretical aspects of their areas of study. However, they can also tend to rely too heavily on these theories, which may not be practical in every situation they encounter. This can lead to an "ivory tower" view of the world that is weak in real-world experience. (Lavin, Deryfus, Slepski, & Kasper, 2007). SMEs who gain their expertise through on-the-job experiences tend to have a strong grounding in reality and tend to know how the system truly works. They know where the shortfalls are and realistic ways to combat the problems that can lead to breakdowns. Lavin (2007) also states that SMEs who work in the field also tend to have a broader range of knowledge as opposed to their formally educated counterparts. OTJ SMEs bring a lot to the table, but they also can pose a threat to the credibility of a project. At times it can be difficult to truly separate the "firefighters" from those who are just along for the ride (Lavin, Deryfus, Slepski, & Kasper, 2007).

The disparities and issues with the SME elicitation bias have led to the creation of the 70% interval rule. The cost analyst takes the range given by the SME and treats the low and high values as a 70% interval. This interval is represented as a triangular distribution. If the distribution is symmetrical the low end is set to the 15th percentile and the high end is set to the 85th percentile. If the distribution is skewed the analyst is advised to skew the bound to match the ratio of initial values given by the SME. (Air Force Cost Analysis Agency, Air Force Cost Analysis Handbook, 2008). The addition of 30 percent enlarges the range ensuring the true cost of the estimate is captured. (Naval Center for Cost Analysis, 2014; Miller, 2020)

Development of the Hypothesis

Clearly, there is room for improvement in the elicitation process. SME estimates are not empirical findings and should not be treated as such. Kilo (1999) states that a true SME should be an opinion leader in the field whom others seek out for advice. Cost estimators additionally assume that SMEs are capable of providing reasonable cost estimates. Unfortunately, estimation expertise likely varies from one SME to another because their field expertise does not imply estimating expertise or probability theory expertise. This leads us to the question: How well trained are our SMEs in probability theory or cost estimation techniques? O'Hagan (2019) states that experts need to be trained to make necessary probabilistic judgements.

There are many ways to train SME's including workshops, in person courses, on the job experience, etc. A new online e-learning course has been identified as a potential solution to safely training experts to make better elicitations. This e-learning course is located at http://www.tonyohagan.co.uk/shelf/ecourse.html, and was sponsored by the U.S. Office of Naval Research. The course was designed to familiarize experts with making necessary probabilistic judgments and ultimately improve their elicitation process. It is composed of four modules educating the user on: probabilities, distributions, and judgements. The fourth module, titled "practice," tests the knowledge of the trainees, by providing them with three exercises to test their skills in probabilistic judgements. We identified *estimation accuracy* as the variable to be tested during this study. The goal of the study is to see an improvement in the *Estimation accuracy* of SMEs after completing the training course. This leads us to our hypothesis:

Hypothesis 1: ***Estimation accuracy will improve after probability training***

With this hypothesis we look to see if the e-learning course can prove to be a valuable asset in the training of SMEs.

Chapter Summary

In this chapter we discussed the various cost estimating methods. The different types of biases were also introduced; as well as the way they effect SMEs and the elicitation process. We identified that there is a problem with SME elicitation and there needs to be an improvement. The basis of our study, training SMEs was introduced along with the e-learning course that we hope will provide one avenue in the efforts to train SMEs to make better estimates. In the following chapters we will discuss the study that was conducted using the e-learning course.

III. Methodology

Chapter Overview

The purpose of this chapter is to give an overview of the study conducted and the processes used to obtain and analyze the data. First, the method of data collection, source, characteristics, and variables will be explained. This will be followed by the methods of statistical evaluation of the data. Finally, there will be figures accompanied throughout the chapter to give a visual representation of the data.

Data Source

Through extensive research and deliberation, we determined that the best way to test *estimation accuracy* would be to conduct a modified version of a repeated measures study using the e-learning course. In a standard repeated measures design, the members of a group are assigned a single treatment, and the results are measured over time. A repeated measures study allows us to detect a good effect size even with a small group of participants. Unfortunately, the repeated measures study also carries with it the possibility of participants dropping out of the study. This could lead to an already small sample size becoming too small to yield any significant results.

The repeated measures study was comprised of three parts: a pretest, the e-learning course, and the post test. The pre and post-test were created to measure our variable which has been identified as *estimation accuracy* (# of correct answers/10). The premise of the study was to have participants take the pre-test, complete the online e-learning course (step 2), and finally take the post test. The results from the pre and post-test would then be compared and analyzed to see if there was a positive change in the number of correct responses from pre-test to post-test. A positive change would lead one to infer that the e-learning course improved *estimation accuracy*.

Design of the Study

Once the type of study and methods were identified; it was time to design the study. Author Douglas W. Hubbard penned a book named “How to Measure Anything,” which became the basis for designing our study. Hubbard (2010) references using calibrated estimates to express uncertainty. Calibrated estimates are a good way to express uncertainty about a commonly unknown number. Calibrated estimates consist of a range of probable values that will have a particular chance of containing the correct answer. This particular “chance” of containing the correct answer is known as the Confidence Interval (CI) (Hubbard, 2010). An example of a calibrated estimate question is: How many hours per week do employees spend addressing customer complaints? Let’s say the participant replied with a 90% CI of 2-4 hours. This would mean she is 90% sure the answer lies within the range of 2-4 total hours a week addressing customer complaints. It also means that she believes there is a 5% chance that the true answer is less than 2 and a 5% chance that the true answer is greater than 4. Calibrated Estimate questions became the blueprint for the pre-test and the post-test. Ten ranged calibration estimate questions were chosen at random for both the pre and post-test. The thinking behind using only 10 questions was to keep the participant’s attention throughout the entirety of the test. If more questions were used, it was possible that participants would lose interest. We chose questions from Hubbard's calibration surveys because they have been validated in previous studies. (Hubbard, 2010)

The degree of difficulty of the questions were a main concern during the design of the study. The goal was to have a good mix of easy, medium, and hard questions. In theory, issues like one test being harder than the other could be avoided with a healthy mix of difficulty. This also helped to provide a solid foundation for the study to actually yield proper results and keep the integrity of the study. The instructions for each test were simple. Participants were asked to read each question and provide a realistic lower and a realistic upper bound with a confidence level in the 90^{th} percentile. After participants took the first test, they were instructed to complete the e-learning course. Following completion of the e-learning course, participants were asked to complete the post-test thus concluding the study. The *estimation accuracy* obtained from the pre-test was then compared to the *estimation accuracy* obtained from the post-test with the hypothesized expectation of positive improvement. The total time needed to complete the entire study ranged from 2-5 hours. Microsoft Forms was used as a medium for the pre and post-tests. In order to keep participants anonymous and still track their progress, each participant was asked to provide the last four digits of their social security number on each test. This became the identifier over the course of the study.

Collecting Participants

Referring to a study conducted by O'Hagan in his article "Expert Knowledge Elicitation: Subjective but Scientific", we determined that using non-SMEs as a proxy for SMEs would still be valid. The original proposal was to hold multiple "in person" study days in the AFIT computer labs. This would encourage all participants to complete the study in its entirety. The "in person" days would also help to boost the chances of reaching the desired goal of ≥50 participants. Due to the 2020 Covid-19 pandemic the majority of the AFIT campus were using distance learning and restricted from campus, complicating the collection of participants. To overcome this hurdle, two emails were drafted and sent to the AFIT community. Around 900+ members of the AFIT community received the emails. The International Cost Estimating and Analysis Association (ICEAA) also sent out the study to their members offering incentive credit for "continued education" to members who participated in the study from start to finish. The email sent out informed members of the study that was being conducted and included the links to the study. The study was limited to DoD personnel only. Participants were asked to complete the steps in sequential order and were given a four-week suspense to complete the study. In compliance with the Institutional Review Board (IRB) all participation was voluntary, and participants were not forced to complete the study. However, they were highly encouraged to finish.

Conducting the Study and Analysis of Data

The study was conducted over the course of four weeks. The "clock" started when the second email containing the links to the study was sent out to the masses. Of the 900+ members who received the email, 44 started the study. Sixteen participants did not complete the post-test lowering the sample size (N) from 44 to 28. Microsoft Forms provided live updates on the completion of the pre-test and post-test. Once the post-test deadline had passed, the data was downloaded as an excel file from Microsoft Forms. Refer to Table 1 and 2 for examples of the pre-test and post-test.

Table 1 Pre Test Example

Pre-Test Example		
	Lower Bound	Upper Bound
The first probe to land on Mars, Viking 1 landed there in what year?		
How old was the youngest person to fly into space?		
When was Elvis Presley born?		

Table 2 Post Test Example

Post-Test Example		
	Lower Bound	Upper Bound
How many yards wide is a football field?		
How many inches long is a 20 dollar bill?		
What percentage of the atmosphere is oxygen by weight?		

Once the data was downloaded from Microsoft Forms the data was organized in ascending order by last the last four numbers of the participant's socials. Next the 16 incomplete participants were removed from the data set. To calculate *estimation accuracy* each range was examined one by one and compared to the answer key. If the answer was correctly inside the provided range it was counted as a "correct" answer. At the end of the test, the number of correct answers were tallied up to determine the *estimation accuracy* for each participant. This process was completed for the pre and post-test. Next the mean accuracy for each test was calculated by taking the average of the total correct answers.

Next, the change in *estimation accuracy* for individual participants was compared. To calculate this, the number of correct answers in the pre-test were subtracted from the number of correct answers in the post test. A negative result would indicate that the participant's *estimation accuracy* declined, a positive result would be indicative of the *estimation accuracy* increased, and a "0" would mean there was no change in *estimation accuracy.*

Once the *estimation accuracy* from each test was compared, they were analyzed in JMP. The change in *estimation accuracy* mean was then analyzed using the Shapiro-Wilkes test for normality and the Wilcoxon signed rank test. The Shapiro-Wilkes test is used to determine if the data is normally distributed. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used to determine whether two dependent samples have the same distribution. The results of these tests were then analyzed with the hypothesized expectation of improvement in *estimation accuracy.*

Summary

As previously mentioned the study was sent out to 900+ members on the AFIT network and members of ICEA. Of those members only 44 members started the study. Out of the 44 members only 28 completed the full study. One limitation of the design was that there was no sure way to ensure participants completed the full e-learning course. Also due to the anonymity of the participants, there was no way to reach out specific participants to encourage them to complete the study. Questions were already prepared and validated in previous studies, ensuring good mix in the degree of difficulty between both the pre and post-test. Collection and analysis of the data was a simple process. Overall, with exception of the participation, the study was very simple to perform and can easily be replicated or expanded in future research.

IV. Analysis and Results

Chapter Overview

The purpose of this chapter is to provide the analysis and results of the methodology outlined in Chapter III. In this chapter, the results of the study will be discussed. The Wilcoxon-Signed ranked test performed will also be discussed.

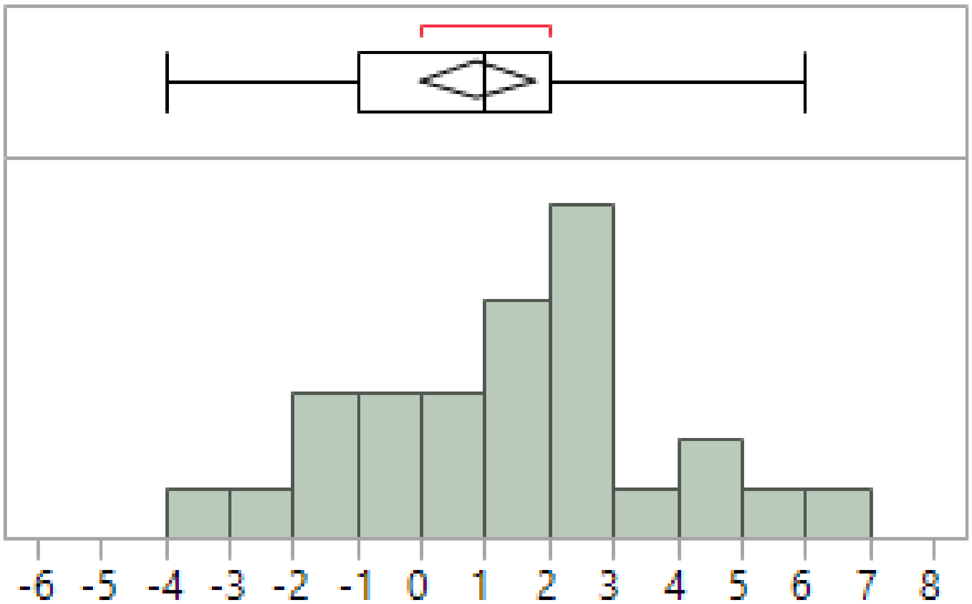
Pre/Post Test Analysis

Once the data was reviewed it was analyzed in search of any change in *estimation accuracy*. At first glance, there was a positive change in some of the participants' accuracy. There were also a few who had negative trend. Overall there were 17 participants whose scores improved, eight who showed a negative trend, and three showed no change.

Descriptive Statistics

The mean of the pre-test was 3.42 and the mean of the post-test was 4.32. This shows an improvement of .89 in the means from pre to post-test. The median of correct answers from the pre-test was 3, with the post-test median being 4. The mode for pre and post-test were 2 and 3 respectively.

Figure 1. Estimation accuracy



Note: This figure is a visual representation of the improvement in *estimation accuracy*

Table 3 Descriptive Statistics

Group			
	Mean	Median	Mode
Pre-Test	3.428571	3	2
Post-Test	4.321429	4	3

The improvement mean as previously mentioned was 0.89. There was a standard deviation of 2.37. This high standard deviation shows that the scores had a spread respectively above and below the mean. The median was 1 with an interquartile range of 3. The interquartile range is derived by subtracting the median of quartile 1 from the median of quartile 3.

Table 4 Estimation Accuracy Descriptive Statistics

	Mean	Std Dev	Median	N	Interquartile Range
Estimation accuracy	0.892857	2.377974	1	28	3

Difference in Means Test

Overall, there was an improvement in the scores from Pre to Post-Test. A difference in means t-test was run to assess the significance of the change in the post to pre-test. This column was created by subtracting the post-test scores from the pre-test scores. The test returned a test statistic of 1.98 and a Prob< t of .0286. Significance was found at the .05 level. Significant results at the .05 level show support for the hypothesis: *Estimation accuracy* will improve after probability training. Refer to table 5 for the difference in means test.

Table 5 Difference in Means t-test

Difference in Means t-test				
	Test Statistic	Prob > |t|	Prob > t	Prob < t
	1.9868	0.0572	0.9714	0.0286*

The difference in means test assumes normality exists in the data. The significance in the difference of means test supports this and in turn supports the hypothesis. To be certain about normality a Shapiro-Wilkes test was conducted using JMP.

Shapiro-Wilk Test

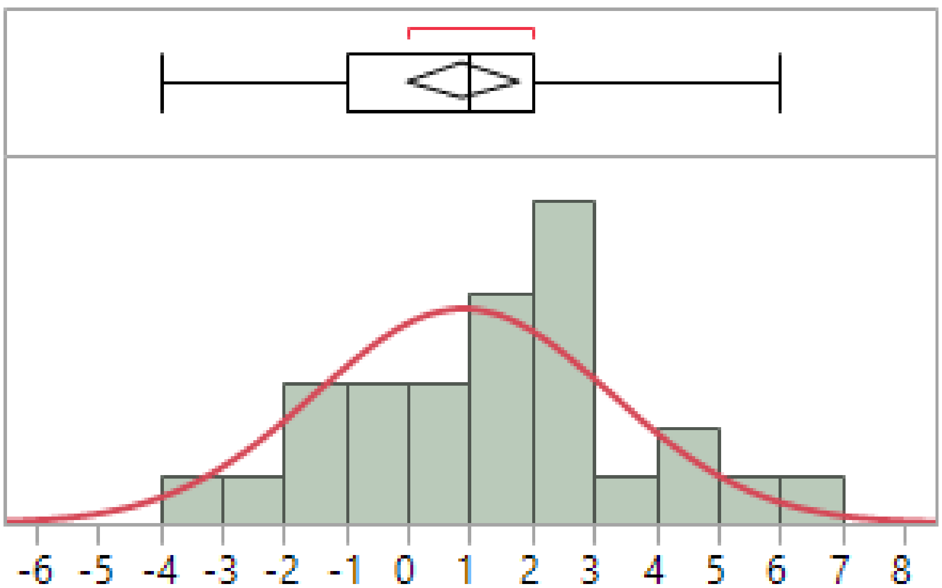
A Shapiro-Wilk test is used to test the null hypothesis that a sample came from a normally distributed population. In this test we were looking to get a p value of larger than our alpha of .05. This would fail to reject the null hypothesis thus proving our data to be normally distributed. After conducting the Shapiro-Wilk our results provided a p value of 0.7340. This p value is greater than 0,05 which fails to reject the null hypothesis. Therefore, our data set is normally distributed and passes the normality test. (Table 6).

Table 6 Shapiro Wilks Test

Shapiro Wilk W Test

	W	Prob < W
	0.975	0.734

Figure 2 Estimation Accuracy w/ Normalization (Shapiro-Wilk)



A Wilcoxon-Signed test was also performed on the data. This was one final stop check for the data even though significant results supporting the original hypothesis were already found. The small N of the data raised small concerns for the need of the test.

Wilcoxon-Signed Rank Test

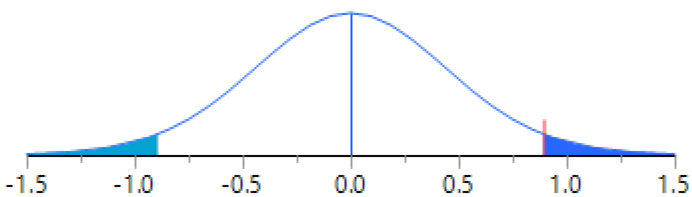
The Wilcoxon-Signed ranked test is a non-parametric statistical hypothesis test. It is used to compare to related, matched, or repeated measurement samples. It also tests to see if the population mean ranks differ. There are three assumptions related to the test. 1.) Data are paired and come from the same population. 2.) Each pair is chosen randomly and independently. 3.) The data are measured on at least an interval scale when, as is usual, within-pair differences are calculated to perform the test. This test was performed on the results of the pre and post-test. Specifically, the "change" in the pre and post-test column. The hypothesized mean value tested was 0. The results of the test yielded a significant result of .0315 at the .05 level supporting the hypothesis. Figure 4 and Table 7 provide a visual representation of the data.

Table 7 Wilcoxon-Ranked Signed Test

Wilcoxon-Signed Ranked Test

	Test Statistic	Prob > |t|	Prob > t	Prob < t
Signed-Rank	80.5	0.063	0.9685	0.0315*

Figure 3 Wilcoxon-Signed Ranked Test



Note: This figure is a visual representation of the Wilcoxon-Signed Ranked Test.

Hypothesis and Research Questions

Hypothesis

The hypothesis stated earlier in the paper was: ***Estimation accuracy will improve after probability training.*** The results of the three tests performed on the data support the hypothesis. We found a 26% increase in *estimation accuracy* in our participants. In layman's terms this means that the training that was conducted in the study had a positive effect on the *estimation accuracy* of our participants.

Research Question 1

How proficient is the average SME in making estimates?

We used non-experts estimating confidence intervals for probability distributions as a proxy for SMEs making cost estimates. Based on the research findings in (O'Hagan, 2019), we determined that the results could reasonably be expected to be similar for experts and non-experts. When applied to this study this question is targeted at the *estimation accuracy* from the pre-test. According to the results our average SME answered 3 questions correct out of 10. A perfectly calibrated SME would answer about 9 out of 10 correctly when generating a 90% confidence interval. From these results, we can be reasonably assured that the average SME is overconfident and not proficient when estimating probability distributions.

Research Question 2

How will probability training affect SME estimation accuracy?

The e-learning course was the main factor of this study. The results of the study provide strong evidence that the e-learning course has a positive effect on training some of the participants in our study. There was an improvement of .89 in the means from pre test to post test. The Shapiro-Wilk test for normality proved that the data was normally distributed. The difference in means test and the Wilcoxon-sign test also provided significant results that showed improvement to SME *elicitation accuracy*.

V. Conclusions and Recommendations

Chapter Overview

This chapter concludes from results the value and usefulness of self-assessment by discussing the implications, limitations of the research, recommendation for future research, and final thoughts on self-assessment.

Research Implications

The results of this study can have many implications in the field and academics. Practitioners can take this data and use it to assess their own SMEs. SMEs can use this study to educate themselves and asses where they need to improve their own elicitation processes. This can lead to a better cost estimation process as a whole.. As previously mentioned there is significant room for improvement. In Academics this study can be used as an example of the importance of training our SMEs. The avenue explored ion this study is just one of many trainings available to our SMEs. Training our SMEs will help the problems with overconfidence, under confidence, and biases that plague the SME elicitation process.

Limitations

While the findings of this study found significance at the .05 level there were plenty of limitations going into this study. The Covid-19 pandemic also posed a big issue because the study had to be performed completely online. The original plan was to perform the study in person. This would have made it easier to ensure that each participant completed each step especially the e-learning course. There was also no way to tell if participants completed each test truthfully without any outside sources. Another limitation was the small number of participants that actually completed the study. We were originally planning to have an N ≥50. It seemed like we were going to reach that however we only had 44 participants start the study and only 28 actually complete the study. Another limitation is the risk and reward structure connected to SMEs in the field creating actual estimates. In the field SMEs have more at stake than just getting an estimate right or wrong. Their credibility along with their very own livelihood may be on the line if they provide a bad estimate. SMEs also have pressure from those above them to do their due diligence and provide legitimate estimates. In our study we could not recreate this risk and reward structure. If we were able to recreate the risk and reward structure test results would potentially look vastly different than the ones we received.

Recommendations for Future Research

Further research could tie into recreating this study on a larger scale with preferably using real SMEs. While our participants served as a useful proxy for SMEs; It would improve the study to conduct the experiment with SMEs, using questions pertinent to the SMEs' respective fields. This would give a proper perspective on the ability of the e-learning course to improve *estimation accuracy* among our SMEs. We also recommend finding a way to ensure that the e-learning course is completed in its entirety including the "practice" module.

Policy Recommendations

Subject Matter Experts are a very important component to the cost estimating community. Unfortunately, there is no way to know how proficient our SMEs are with their estimates. Too often do we end up tailoring the estimates given by our SMEs when we should not need to. An example of this is the 70% AF CRUH rule discussed earlier in this thesis. There needs to be an improvement in the way we train our SMEs to make better estimates. Preparing our SMEs to be better equipped to answer our questions will lead to better inputs and overall a better cost estimating process. We believe that this e-learning course is one of many trainings that can be used to train our SMEs to make better estimates. The results of the study show that training SMEs has the potential to improve SME estimation skills. An overhaul of the training of SMEs can change the landscape of estimating overall and lead to a revision of the 70% AF CRUH rule. A new certification process can be implemented. This would verify that a SME has been sufficiently trained to provide more reliable estimates. Therefore the 70% interval rule could be changed to 80% or even 90% interval.

Final Thoughts

This thesis explored the possibility of training SMEs to make better estimates. An e-learning course identified in O'Hagan's article "Expert Knowledge Elicitation: Subjective but Scientific" became the basis for the entire study. Though performed on a small scale and not with actual SMEs, significant results were found. Future researchers could build upon the findings from this study in field conditions, using SMEs in a cost estimation context. The e learning course is not the final solution to our SME estimation issues, but it has its place. There is much more research that can be done in the field of training our SMEs.

Bibliography

Air Force Cost Analysis Agency. (2008). *Air Force Cost Analysis Handbook.* Washington DC.

Air Force Cost Analysis Agency. (2008). *Air Force Cost Analysis Handbook (AFCAH).* Washington D.C.

Anderson, T., & Elloumi, F. (2004). *Teaching in an online learning context.* Athabasca University.

Bellarosa, C., & Chen, P. Y. (1997). The effectiveness and practicality of occupational stress management interventions: A survey of subject matter expert opinions. *Journal of occupational health psychology*, 247-262.

Department of Defense. (2018). *Work Breakdown Structures for Defense MAteriel Items.* Washington DC.

Frick, D. E. (2010). Embracing CUncertainty n DoD Acquisition. *Defense Ar Journal*, 355-374.

Hubbard, D. W. (2010). *How To Measure Anything.* Hoboken: John Wiley & Sons, Inc.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.

Kilo, C. M. (1999). Improving Care Through Collaboration. *PEDIACTRICS*, 384-393.

Lavin, R. P., Deryfus, M., Slepski, L., & Kasper, C. E. (2007). Said Another Way Subject Matter Experts: Facts or Fiction. *Nursing Forum*, 189-195.

Miller, C. (2020). An investigation into Subject Matter Expert Elicitation in Cost Risk Analysis. *Theses and Dissertations*.

Naval Center for Cost Analysis. (2014). *Joint Agency Cost and Schedule Risk and Uncertainty Handbook.* Washington DC.

O'Hagan, A. (2019). Expert Knowledge Elicitation:Subjective but Scientific. *The American Statistician*, 69-81.

Pace, D. K. (2002). *Subject Matter Expert (SME)/Peer Use in M&S V&V.*

Tversky, A., & Kahneman, D. (1974). Judgement under Uncertainty: Heuristics and BIases. *Science*, 1124-1131.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 074-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From – To)
25-03-2021	Master's Thesis	August 2020 – March 2021

TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Training Subject Matter Experts to Make Better Estimates	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Scott, Cameron R., 1LT, USAF	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)	8. PERFORMING ORGANIZATION REPORT NUMBER
Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/ENV) 2950 Hobson Way, Building 640 WPAFB OH 45433-8865	AFIT-ENV-MS-21-M-268

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
AIR FORCE LIFECYCLE MANAGEMENT CENTER 1865 Fourth St, Bldg 14, WPAFB, OH 45344 937-986-5473 and zachariah.sayre@us.af.mil ATTN: Zachariah Sayre	AFLCMC/FZC
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Subject Matter Expert (SME) Elicitation is commonly used as a cost-estimation method when no historical data exists. The elicitation process assumes Subject Matter Experts are the resident experts in their individual fields of study. In addition, the process assumes SMEs provide reasonably trustworthy cost estimates that follow the tenets of probability theory. This last assumption is the subject of this study. A Navy sponsored e-learning course was created to help train SMEs in understanding probability distributions (O'Hagan, 2019). We conducted a study hypothesizing that taking the e-learning course would improve study participant accuracy in estimating confidence intervals for probability distributions. We found strong support for our hypothesis and significance in our results providing a 26% increase in the estimation accuracy of our SMEs. The results of the study show that training SMEs has the potential to improve SME estimation skills

15. SUBJECT TERMS
Cost Estimates,

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Robert D. Fass, AFIT/ENV
U	U	U	UU	38	19b. TELEPHONE NUMBER (Include area code)