RISK-3824

Calibration Assessments: Validation of Subjective Probabilities and Impact Ranges in Risk Analysis

Francisco Cruz Moreno, PE

Abstract–Risk analysts and project teams must rely on expert judgment to collect subjective probabilities and ranges of potential cost and schedule impacts, especially when they are performing quantitative risk analyses. The main reason for using subject matter experts is the lack of reliable historical data associated with risk impacts. Research shows that subjective probabilities and risk impact ranges consistently yield overconfident and underconfident results, which, in turn, generate inaccurate cost values at selected confidence levels and confident intervals.

This paper explores the limitations of current elicitation approaches to collect and use subjective probabilities and impact ranges to assess uncertainty and risks. It provides several examples of different calibration assessment results and their adequate use to improve the strength of risk input data. It also presents a case for risk analysts to use sound scientific rigor in respect to inputs when performing qualitative risk assessments and quantitative risk analyses in support of decision-making. The author suggests the use of calibration assessment in any modeling approaches using subjective inputs whether they be decision trees, parametric models, Monte Carlo simulation, reference class forecasting, or system dynamics.

Table of Contents

Introduction

Risk analysts consult several team members and experts from different disciplines to validate cost estimates, schedule forecasts, propose means and methods, and assess internal and external risks that could impact a particular project or program. Risk analysts are often faced with limited availability of sound historical data to create probability density functions (PDF); additionally, analysts survey peer risk analysts and colleagues informally to ensure that the impact ranges are reasonable and that the PDF they chose are adequate.

While there is an abundance of project data, risk analysts rely almost exclusively on subject matter experts (SME) to collect probabilities of occurrence and expected ranges (e.g., three-point estimates), which are then captured as PDF to represent potential cost and schedule impacts. Input from experts includes their best guesses that are usually agreed upon by consensus and then combined to create a final PDF for a particular risk. This exercise is repeated for every critical risk that will be further assessed during a quantitative risk analysis (QRA).

Project teams and risk analysts have been blending historical data, computer-based forecasting, and subjective expert judgments for decades to determine cost and schedule contingency amounts. In the last decade, the risk management community has collected a vast amount of cost overrun data and has performed research spanning many industries. However, this research does not show that QRA practices on average have changed since Monte-Carlo simulation became practical 30 years ago. The result has been that there is no improvement on average as to how well risk analysis are predicting cost growth [1].

When the QRA method relies upon subjective judgments, the most common omission in contemporary risk management practice is neglecting to perform expert calibration assessments to validate risk inputs.. This paper will provide insight regarding the tools and methods currently used by risk analysts and consider how bias introduced by experts is compromising risk analysis results.

The Case for Calibration Assessments

A body of experimental and long-term work in social psychology, economics, and statistics has focused on decision-making and on combining expert judgment estimates (forecast or fortune-telling). Since experts provide input for probabilities and potential impacts in risk analysis, it is important to understand the meaning of the term *expert*. Below are two definitions in the domains of cost engineering and forecasting.

- A person or persons recognized, either formally or informally, as having specialized knowledge or training in a specific area [RP 10S-90].
- An expert is someone who makes especially accurate forecasts [2].

Notwithstanding the reliance on expert judgment, 40 years of research show that SMEs are poorly calibrated due to limited information and insufficient memory capacity [3]. Additionally, in the last decade, there has been a major focus by risk analysts on creating a statistical model that can better forecast project outcomes, putting emphasis on data cleaning, normalization, and model calibration but discounting bias introduced by subjective judgment. For instance, Recommended Practice RP 42R-08: Risk Analysis and Contingency Determination Using Parametric Estimating [4] properly addresses the calibration of a statistical model based on historical data; it mentions the word "calibration" more than 20 times. However, it does not address the calibration of SMEs. In other instances, such as in Recommended Practice RP40R-08: Contingency Estimating – General Principles [5], the word "calibration" is not mentioned even once.

Bias and Noise

Behavioral scientists, particularly psychologists, sociologists and economists, have studied biases in thinking and how bias affect the way people remember, evaluate, understand, judge, and use information. In the 1970s, psychologists Kahneman and Tversky worked on the concept of *cognitive bias* after assessing how people struggled with judging objectively. They found that many errors in judgement occurred even though subjects from several experiments were encouraged to be accurate and received rewards for correct answers [6].

Most recently in 2021, Kahneman, Sibony, and Sunstein discussed the concept of *noise*, a type of judgment error that produces unwanted variability in judgments that should be identical. They postulated that both bias and noise need to be comprehended to understand error in judgment [7]. Flyvbjerg goes further by calling for debiasing estimates and decisions given that cost overruns and schedule growth are not a product of error but of bias, and he states that organizations will not see any improvements until decision makers address and understand behavioral bias [8]. As of these writings, there are approximately 200 cognitive biases that have been identified; almost all of them have been captured in the *cognitive bias codex* [9], which depicts cognitive errors and allocates them into four quadrants: information overload, memory to remember facts, need for speed to act fast, and not enough meaning.

As it pertains to risk management, there are prevalent biases that can help to explain inaccurate judgments leading to cost overruns and schedule growth results. These include anchoring, attribution asymmetry, availability bias, base rate fallacy, confirmation bias, group think, inertia, optimism bias, overconfidence bias, planning fallacy, repetition bias, selective perception, and underconfidence bias. This paper does not define these biases since other sources exist that provide a description, a guidance to find evidence of their occurrence and how to respond to them [10].

There are tools currently available for risk analysts to identify and address certain biases such as overconfidence, underconfidence, group think, confirmation, and strategic misrepresentation. However, there is ongoing research aiming at defining and clarifying the biases that tend to

overlap and make it difficult to pinpoint the key bias drivers [11]. One such tool is the use of calibration assessments, a form of structured elicitation.

What Is a Calibration Assessment?

A calibration assessment is a method used to elicit and quantify each expert's uncertainty to measure the statistical accuracy of their judgments. The goal of a calibration assessment is to identify reliability and accuracy of the judgments provided by experts, and to minimize the inherent biases of expert judgments to improve the accuracy of the forecast.

Calibration assessments are performed by a facilitator, who is most likely a risk analyst, with the approach of distributing between 10 to 15 predetermined seed questions to the SMEs before a risk workshop to assess overconfidence, underconfidence, planning fallacy, and availability biases. The SMEs return their responses to the facilitator who, in turn, compares the responses with the known *true value* of each seed question to determine if the answers were accurate.

There are several approaches to perform a calibration assessment that include mathematical and behavioral procedures with the goal of combining, or aggregating, expert judgments; these include the *classical model* [12] and the *Investigate, Discuss, Estimate, and Aggregate* (*IDEA) protocol*. The steps to perform a calibration assessment include identifying and selecting the subject matter experts, training experts in probability elicitation and probability assessment [13], and combining their responses.

Decision and Quantitative Risk Analysis Methods Impacted by SMEs Judgments

As projects go through their stage-gate process and are further developed, risk analysts have a myriad of risk assessment methods and tools at their disposal to perform qualitative risk assessment and quantitative risk analysis. The decision to use one tool or method over another, or even combining two or more methods, is usually based on the organization's quantitative risk maturity level [14], the available historical data and SMEs, the size and complexity of the project, and the time available to complete the assessment. The following decision, benchmarking and quantitative risk analysis methods, which are perceived as either subjective or empirically based, contain various degrees of bias that risk analysts need to consider, identify, and reduce to improve their forecasts.

Decision Trees

A decision tree is a decision analysis methodology that is used to solve problems involving multiple decision paths that carry uncertainty. Organizations apply this quantitative technique to many uncertain situations, such as when determining what type of equipment they should buy; choosing between building a new structure or renovating an existing area; choosing between several proposals; or even evaluating whether or not to submit a claim. Decision trees have enjoyed a revival in recent years [15], and they are commonly used to solve classification problems in a machine-learning domain.

When it comes to developing decision trees, risk analysts need to identify the major decisions and their alternatives along with their probability of impact and potential costs. Ideally, the organization will have historical data that could be used to calculate the probability of occurrence of each decision action along with their associated costs. The tree structure will include all possible outcomes with the main objective of finding the most favorable path that yields the least cost or the highest return.

Figure 1 shows a basic decision tree structure identifying the decisions related to building a new facility or renovating the actual facility. It includes the available alternatives to address the question at hand, cost estimates for each alternative, and uncertainties associated with each alternative. The numbers highlighted in green represent cost estimates while the percentages highlighted in yellow represent the probabilities of each potential outcome. When looking at the cost estimate and the probabilities, *bias* may manifest in the form of overconfidence, availability, optimism, anchoring, planning fallacy and premature termination. Risk analysts should evaluate and validate the assumptions made by the decision makers to determine the probabilities of each event, the sources used to determine cost estimates, and confirm whether other possible alternatives exist.

This simplified decision tree structure does not show a completed decision tree analysis nor a solved tree since the focus is in evaluating the sources of bias, which resides within the inputs and the assumptions initially considered to answer the main question (i.e., build a new facility versus renovating the existing facility).

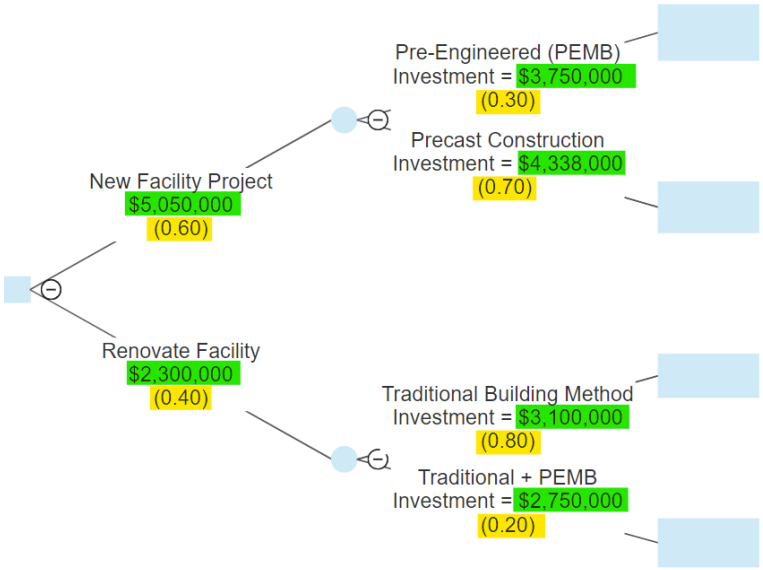

Figure 1–Build New or Renovate – Highlighted Areas with Potential Bias

Reference Class Forecasting

Reference class forecasting (RCF) is a method that was introduced by Bent Flyvbjerg and was created from the theories developed by Daniel Kahneman and Amos Tversky [1]. The goal of the

RCF method is to address the planning fallacy, optimism bias and strategic representation to minimize their impact on a project's cost, schedule, and benefits targets. While most of the research and literature about RCF is dedicated to megaprojects, RCF could also be used on medium to large projects. Planners, estimators and risk analysts use historical cost and schedule data to build probability distributions and to create the RCF model. The author has used the RCF method for Class 10 and Class 5 estimates in transit projects, when there is limited availability of project details and scope definition, and when the level of uncertainty is very high. Once the RCF model is completed, decision makers determine the overall funding amounts including reserves, usually the P80 confidence level, after confirming the risk appetite of their organization.

The following points summarize the steps used to create an RCF model:

1. Assess the historical data; use same reference classes
2. Clean the data and remove inconsistent values; do not remove outliers
3. Normalize the data and strip the cost or schedule of any contingency
4. Deflate the cost values back to the date of the estimate depending on the available economic data

Risk analysts need to avoid collecting data that is not representative of their projects and which is statistically significantly different from each other; this may lead to sampling bias. Optimism and uniqueness biases might affect the reference class forecast if the decision makers believe that their projects are not as risky as past projects, or if they are optimistic about treating risks early in the project. These biases may render the RCF exercise useless.

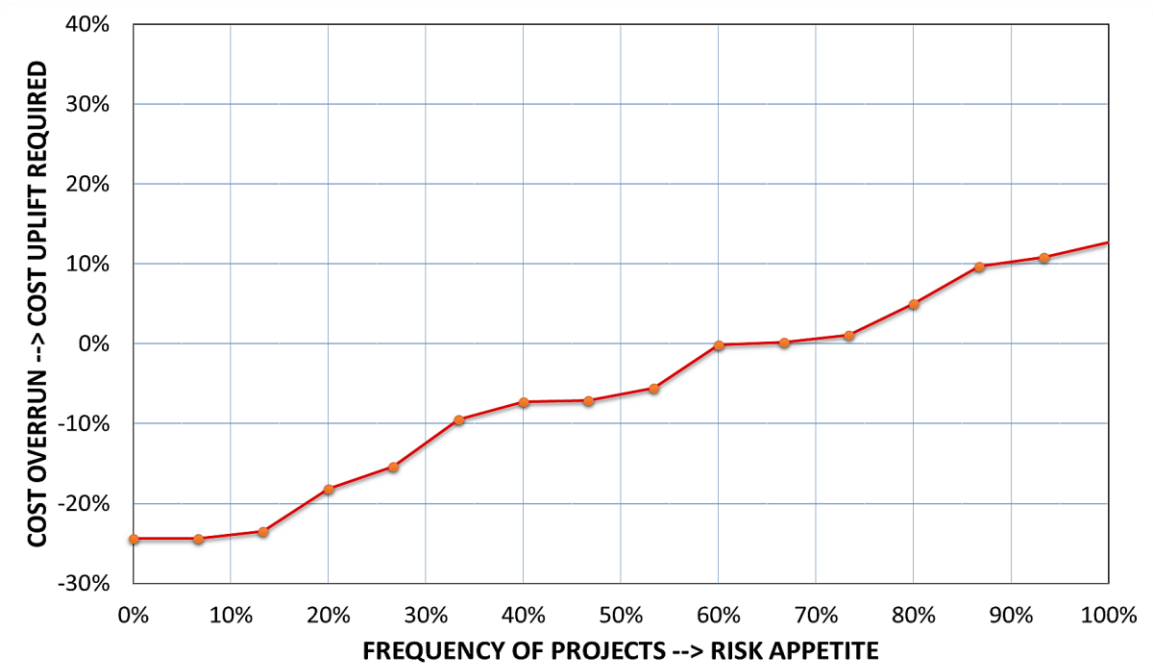

Figure 2–RFC Model for Bus Rapid Transit Projects in the US

Figure 2 shows an RCF model created by the author to determine the cost uplift required for bus rapid transit (BRT) projects in the United States. The RCF model was created with data from projects where the revenue service dates fall between 2007 and 2021; it shows that BRT projects have not suffered from cost estimates inaccuracies that have plagued transit projects. For instance, using the 80% confidence level would only require a cost uplift of ~5%. This is lower than what international benchmarks show [16], where the average cost uplift for BRT projects is ~ 41%. The main hypothesis drawn by the author is that transit organizations have overallocated funds for BRT projects, indicating that there may be a bias towards cost overruns.

Linear Regression Models and the Parametric Method

Parametric models are created using historical data that are fit into either a simple linear regression model that contains only one independent variable, or a multiple linear regression model that contains two or more independent variables. In the case of cost engineering, the goal is to identify cost estimating relationships between dependent variables (e.g., total project cost, manufacturing cost) and independent variables (e.g., length, weight, design cost, size).

While a level of empiricism when developing parametric models does exist, great potential for the expert developing the model to introduce bias and noise also exists, which could ultimately invalidate the model. This paper will not delve into how to create, test, and calibrate a parametric model, nor it will provide the steps to interpret it. It will point out, however, the issues related to the assumptions that need to be satisfied to trust the model, as well as potential biases that could impact the model's accuracy.

Risk analysts may introduce errors from the initial step of formulating a hypothesis for the variable of interest that will be tested (e.g., cost growth, schedule growth). Additionally, risk analysts must perform cost and schedule data normalization that includes inflating or deflating costs; adjust cost and schedule baselines to account for bias; and adjust for complexity, materials and equipment, and contract types (e.g., cost plus).

After the selection of the dependent and independent variables, risk analysts will test the relationships using diagnostic plots, assess the regression's functional form, and evaluate the curve for the best-fit pattern. Note that the term curve fitting is used in cost estimating to describe the evaluation of the best-fit pattern, although the curve may be a straight line.

A general linear regression model is given by equation 1 for *n* unknown parameters:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \qquad \textbf{Equation 1}$$

Where,
Y = Dependent variable
β_0 = Constant or Intercept
β_1 = Coefficient 1 or Unknown parameter 1
X_1 = Independent variable 1

β_2 = Coefficient 2 or Unknown parameter 2
X_2 = Independent variable 2
ε = unobservable error term

Risk analysts need to validate and satisfy the following main assumptions to avoid violating them, thereby rendering the parametric model as unreliable.

1. *Linearity*, so that the regression is linear in the parameters and presents a correct functional form. The main issue with getting a linearity assumption wrong is that the coefficients and standard errors of the results have a high likelihood of being inaccurate. Likelihood ratio tests or residuals plots are needed to validate this assumption. In some instances, when linearity is not confirmed by the risk analyst, the functional form may need to be adjusted to represent the correct relationship between variables. As the risk analyst starts creating the regression model, there would be various attempts of trial and error. For instance, a risk analyst may start with one independent variable but may need to add additional variables if the linearity assumption is violated. In other instances, risk analysts may even have to transform the functional form from a quadratic type of relationship.
2. *Homoscedasticity* represents a state where the error variance is constants across the values of the independent variables. The main issue with getting the homoscedasticity assumption wrong is that p-values may be smaller than anticipated, leading to an underestimation of the error variance that in turn produces inaccurate F-values and t-values. Solutions to detect and treat the lack of homoscedasticity include the use of weighted least squares or the use of White's standard errors if using a statistical software.
3. *No multicollinearity*, meaning that risk analysts would need independent variables that are not related to each other. The main issue with getting the assumption wrong about no multicollinearity is that the coefficients and standard errors of related variables may be unreliable. Solutions to detect and treat multicollinearity include evaluating the correlation between the independent variables, calculating the variance inflation factors, and potentially removing one or more of the affected, least relevant independent variables.
4. *Exogeneity*, which includes a large class of issues including reverse causality and the omitted variable bias. Exogeneity means that each independent variable (*Xn*) is uncorrelated to the dependent variable (*Y*). The main issue with getting the assumption wrong about exogeneity is that the regression model can only be used to make predictions; however, the model cannot infer causation. Solutions to detect and treat exogeneity include checking the correlation among independent variables, or using intuition, which is prone to bias.

There are other assumptions that are also validated during the development of the linear regression model, which include checking for independent error terms (i.e., autocorrelation), and verifying the normality of the error terms. Other potential challenging points that analysts need to diagnose include identifying outliers, influential observations, leverage points, and determining regression model over-fitting [17]. Risk analysts might be using parametric models

already developed such as the RAND model found on RP 43R-08: Parametric Modelling [18]. Given that the RAND model yields results in a mean predicted value, risk analysts must use methods to translate the mean to a full distribution of values. This may introduce additional bias and noise. Additional cognitive biases that risk analysts may face when working with parametric models include planning fallacy, optimism bias, confirmation bias, the curse of knowledge bias, law of the instrument, and anchoring bias.

Instead of trusting parametric models at face value, risk analysts must inquire and formulate a series of questions if they are not creating their own parametric models. Documenting the use of an existing model or the development process of a new parametric model is paramount given that parametric models are based on a combination of historical data and expert judgment. At a minimum, risk analysts should ask the following questions to ensure that they are working with a valid parametric model:

- What historical data was used to develop the parametric model?
- Do risk analysts have access to the historical data?
- Is the historical data relevant and representative of my project?
- How was the data cleansed and normalized?
- Which assumptions were tested and validated during the parametric model development?
- When was the last time that the parametric model was calibrated? How often is it calibrated?
- What biases could influence the SMEs and risk analysts who are providing inputs (e.g., choosing *parameters* and *adjustments* defined by the model)?

Qualitative Risk Analysis

For decades, risk analysts have relied on qualitative risk assessments (QLRA) to prioritize risks after qualifying their probabilities of occurrence and potential impacts [19]. Once all identified risks are qualified, risk analysts focus their attention on the most significant risks to define the treatment and post-treatment response efforts. The main steps of a QLRA are risk identification, development of risk thresholds to create a probability and impact matrix, and ranking the risks based on their ratings (i.e., product of their probabilities of occurrence times their impacts). When it comes to risk identification for major projects, the author recommends the use of a minimum of three risk identification strategies to avoid group think, confirmation, hindsight, and uniqueness biases. These risk identification strategies include individual interviews, check lists, the Delphi method, and the Crawford slip method.

While most risk analysts use risk registers in their daily risk management practices and are familiar with risk thresholds, some risk analysts do not use thresholds and rely on the SMEs' judgements to identify critical risks that will be included in the QRA. This is another source of bias and noise. Regarding the risk register and the probability and impact matrix, some researchers and practitioners have found several inconsistencies and limitations with its use, namely subjective thresholds; ambiguous inputs and outputs of probability and impacts; range

compression; and suboptimal resource allocation [20]. There are other issues pertaining to how experts interpret verbal labels for probabilities such as frequent, likely, and rare. Any rating or threshold designed to qualify project objectives, its risks' probabilities, and their impacts will have inherent bias and noise. The following table shows an example of a rating scale based on 5 x 5 levels and the thresholds determined for the risk ratings.

Probability						
	Frequent	0.9	1.8	2.7	3.6	4.5
	Occasional	0.7	1.4	2.1	2.8	3.5
	Likely	0.5	1	1.5	2	2.5
	Improbable	0.3	0.6	0.9	1.2	1.5
	Rare	0.1	0.2	0.3	0.4	0.5
		Negligible	Minor	Major	Hazardous	Catastrophic
		Risk Impact				

Table 1–Probability and Impact Matrix with Risk Thresholds

Risk and Uncertainty Quantification Using Monte Carlo Simulation

During the past 30 years Monte-Carlo simulation (MCS) has become the preferred statistical sampling technique used to perform cost and schedule risk analyses for capital projects. MCS has made it easy to present estimates of possible outcomes as distributions and has displaced historical project databases; MCS relies on quick risk inputs from trusted SMEs and places the focus on subjective probabilities. On the other hand, this is one of the main drawbacks of MCS since it uses inputs based on subjective expert judgements that ultimately create biased forecasts. Practitioners have identified the main biases that are prevalent in the development of cost estimates and project schedules and have written ways to recognize and address optimism bias and planning fallacy [21] [22]. MCS is one of the most subjective risk analysis approaches to perform QRA, whether risk analysts use it to perform a standalone cost risk analysis, a schedule risk analysis or integrated risk analysis.

On major projects, the main challenge that many risk analysts face is selecting probability distribution functions (PDF) to assign to the critical risk impact estimates; these PDF may be assigned to cost line items, activity durations, or other variables deemed important by the project team. There are over a dozen PDFs that risk analysts can choose from, and they include the uniform, triangular, trigen, normal, beta, lognormal, and binomial distributions. RP 66R-11: Probability Distribution Functions [23], lists common PDFs used in cost and schedule risk analysis and it highlights their advantages and disadvantages. These PDFs are very different from each other because they either concentrate or spread the probability mass around the mode. Having a good understanding about how each PDF works is key to being able to represent the inputs provided by the SMEs; for instance, risk analysts must decide whether they will use a discrete or continuous PDF, unimodal or bimodal, skewed or symmetric.

Risk analysts need to be aware not only of the SMEs' preconceptions but also their own given that optimism, confirmation, overconfidence, group think, anchoring and uniqueness biases are more prevalent at the time of choosing the appropriate PDF. Practitioners are uncovering new biases related to how they plan and schedule capital projects. For instance, early-dates bias

occurs in schedule risk analysis, and it happens when activities are simulated on early starts in every iteration, disregarding the impact of delayed or late starts due to impact of float use [24]. PDFs will become the inputs to the MCS that will influence the QRA results and forecasts, and ultimately guide the decision makers. Therefore, it is vital to improve the reliability and accuracy of subjective judgments.

Other Decision Methods and Indicators Susceptible to Bias

There are additional decision tools and parameters that risk analysts have at their disposal to evaluate problems and make forecasts. Below are some examples of these tools and parameters that enjoy widespread use or are gaining traction, and which are prone to bias and noise based on the author's experience:

- Scenario Analysis; planners use this technique to examine different impacts of positive and negative events. It usually requires decision makers to identify a worst-case, expected case, and best-case scenario where heuristics are used in the form of percentage of improvement or failure. Initially, planners calculate the expected case, and then they apply percentages to the expected case that range between -20% to -30% reduction for a worst-case scenario, to 30% to 50% improvement for a best-case scenario. Scenario analysis is susceptible to base rate, optimism, power, anchoring, and availability biases.
- Cost-Benefit Analysis; usually performed early during the scoping phase to assess different alternatives and to choose the most appropriate one in terms of cost and return. It may include a full life-cycle cost analysis but most often the cost-benefit analysis is performed up to the asset's first year of operations. Cost-benefit analysis is prone to optimism, commitment, power, base rate, anchoring, and availability biases.
- Escalation; it is common practice to use industry-specific price indices that are affected by location, market conditions, cash flow duration, and risk analyst's adjustments that are susceptible to substitution, base rate, availability biases.
- Correlation; this parameter is very subjective and risk analysts use it constantly to model the relationships between risk and activity pairs. In practice correlation values ranges from 30% to 100% positive correlation with 70% to 80% being commonly used; observational data from the author's experience show that negative correlation is rarely used. The main issue with the use of correlation coefficients is that there are not many sources of historical data, leaving risk analysts with their own subjective judgments to choose a coefficient. Correlation is prone to group think, availability, anchoring, and base rate bias.
- Systems Dynamics Modeling; this method is well established in construction claim analysis and is starting to gain momentum in the risk management practice, where risk analysts use it to model systemic risks and non-linear cost impacts. While this paper will not cover this method in detail in terms of steps to create and run the system dynamics' model, readers can review an introduction to non-linear probabilistic modeling paper written by Dr. Raydugin [25]. System dynamics is susceptible to optimism bias, anchoring, strategic representation, and group think.

Calibration Assessments to Validate Subjective Probabilities and Impact Ranges

Expert judgment informs every aspect of the risk assessment process, from identifying and qualifying risks to quantifying their likelihood of occurrence and their potential impacts, to suggesting the amount of cost and schedule contingency. As presented in this paper, it is well documented that bias and noise affect the inputs in the risk models and might skew the results. Despite this, calibration of expert judgments within the construction industry remains nonexistent. There are several structured elicitation protocols (e.g., classical model, IDEA) to collect expert judgments formally and to help minimize the influence of biases and noise.

The author has used the *classical model* for several years to perform calibration assessments on risk analyses for major capital projects. This is a protocol that is relatively simple to apply and has been in use for over 30 years. One of the benefits is that it is convenient for use in remote elicitation during virtual risk workshops as well as face-to-face and hybrid risk workshops. There are four (4) basic steps to integrate a calibration assessment using the classical model to any QRA method; these are expert identification and selection, expert training, expert elicitation, and aggregation of expert judgments. Figure 3 illustrates the basic steps and their timing of execution during a risk analysis workshop.

Pre-workshop		Risk Workshop	Post-workshop
1) Identify, contact and brief experts on the calibration assessment.	**2)** Perform the 1st calibration exercise (i.e., benchmark test) before any training. This usually occurs during the kickoff meeting. Share the results with SMEs as anonymous responses. Training should follow the 1st calibration exercise.	**3)** Perform the 2nd calibration exercise during the risk workshop; usually before defining the thresholds or the QLRA. Then proceed to collect the probabilities of each risk in the risk register and impacts ranges using PDFs.	**4)** Aggregate the cost and schedule impact ranges only of the calibrated SMEs.

Figure 3–Steps to Perform a Calibration Assessment during a Risk Analysis Workshop

Training of the SMEs includes basic understanding of the biases that affect judgment and guidelines on how to reduce them, how the different PDFs work, and estimation of probabilities and confidence intervals for uncertainty ranges. The risk analyst distributes 10 to 15 predetermined seed questions using a survey platform to facilitate the collection of responses and to track who has provided the answers. The author suggests performing at least one (1) calibration assessment per risk analysis, and ideally two (2) calibration assessments so that the risk analysts can test probabilities and range uncertainty. It takes 5 to 10 additional hours to include the calibration assessment to any cost or schedule risk analysis.

After the calibration assessments are performed, SMEs provide their QRA inputs on probabilities and uncertainty ranges. Regarding the ranges, SMEs are asked to provide their 80% confidence intervals. Risk analysts then aggregate their responses; some approaches combine all experts'

responses, while other aggregate the responses from either the top three or top five calibrated SMEs; the author uses the latter.

Lessons from Calibration Assessment Results of 14 Quantitative Risk Analyses

The author collected the data, responses, and results from 14 recent calibration assessments performed by the author. This is the largest known database of calibration assessments specific to construction published that covers projects in the aviation, transportation, transit and pharmaceutical industries. The database contains assessments of risk probabilities and uncertainty ranges, it has more than 3,500 questions and responses from over 240 participants. Below is a breakdown of the database composition for questions concerning probabilities when using expert judgment:

- 68% of participants are either overconfident (53%) or underconfident (15%).
- 32% of participants are calibrated, meaning that they were correct 80 percent of the time.
- For the same set of questions, participants from the private sector were slightly more calibrated than participants from the public sector (i.e., 34% vs. 30%). However, participants from the public sector were less overconfident than their private sector counterparts (51% vs 56%). The author found that respondents from the private sector were not statistically different from public sector respondents in terms of calibration (Chi-Square = 1.37, df = 2, p = .5051).

Inputs are captured by asking SMEs to provide answers to questions unknown to them; SMEs need to provide their confidence, from a range of 50% to 100% confidence, that their answers are correct. The figure below shows the calibration assessment results from one of the 14 projects; the responses to the binary questions that asked for a true or false answer show that team members were 72% overconfident; 11% underconfident; 17% calibrated.

Respondent	Question Score	Calibration Score	Calibration
Respondent 1	9 / 15	11.4	Overconfident
Respondent 2	8 / 15	13.1	Overconfident
Respondent 3	10 / 15	7.7	Underconfident
Respondent 4	8 / 15	8.4	Calibrated
Respondent 5	9 / 15	12.0	Overconfident
Respondent 6	7 / 15	9.1	Overconfident
Respondent 7	6 / 15	10.0	Overconfident
Respondent 8	10 / 15	9.9	Calibrated
Respondent 9	11 / 15	13.4	Overconfident
Respondent 10	8 / 15	11.4	Overconfident
Respondent 11	10 / 15	10.1	Calibrated
Respondent 12	10 / 15	12.3	Overconfident
Respondent 13	7 / 15	11.3	Overconfident
Respondent 14	7 / 15	9.7	Overconfident
Respondent 15	11 / 15	7.5	Underconfident
Respondent 16	8 / 15	11.0	Overconfident
Respondent 17	11 / 15	12.3	Overconfident
Respondent 18	9 / 15	13.5	Overconfident

Figure 4–Calibration Assessment for Binary Questions (True or False)

For questions related to uncertainty ranges, which are usually captured as three-point estimates, the results shown in Figure 4 for calibration question No. 5 indicate that there is overconfidence and availability bias in the predictive judgments provided for by the respondents; the answers are mostly on one side of the true value rather than the other. The results also confirm that the mean of their responses (i.e., estimates) is often much lower than the true value. While group think was not confirmed since the SMEs answered the questions independently, it does happen during the QRA while collecting the probability of occurrence and the ranges for cost and schedule impacts. Respondents were asked to provide their optimistic, most likely, and pessimistic inputs per each question, and to use an 80 percent confidence interval for the ranges. Overall, judgements were more accurate when the true value of the responses was below 1,000 units as shown on calibration question No. 3. There was a higher level of bias when the answers involved units larger than 1,000 as shown in calibration question No. 5.

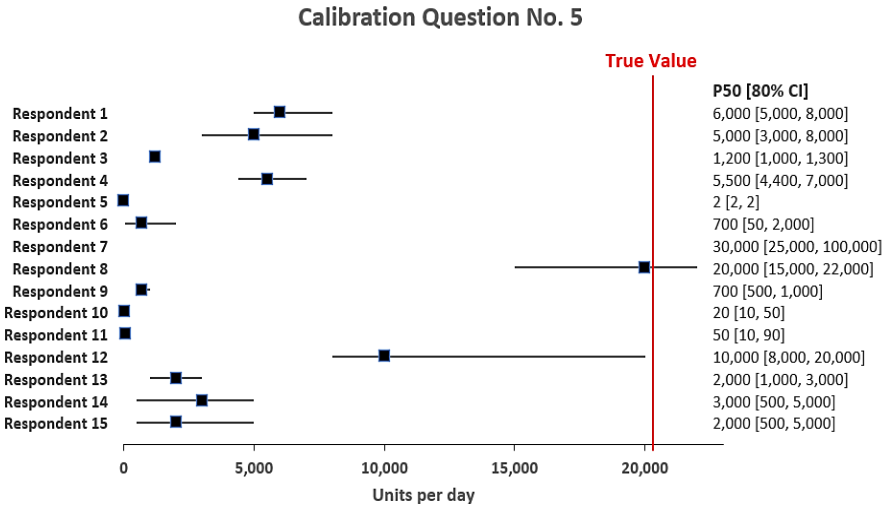

Figure 5–Calibration Assessment for Uncertainty Ranges - Results for Question No. 5

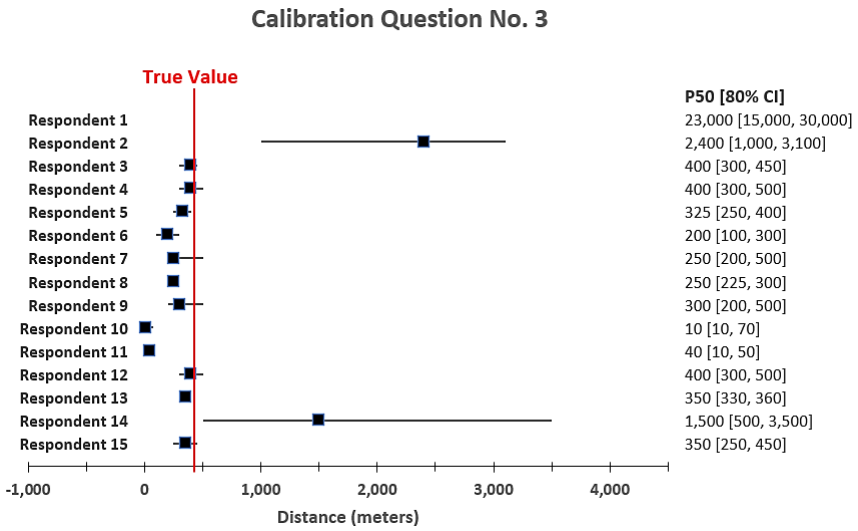

Figure 6–Calibration Assessment for Uncertainty Ranges - Results for Question No. 3

Conclusion

Data collection, evaluation and usage should be a main goal of any organization; historical data should drive decision making. In the absence of data, where risk analysts rely heavily on SME's and their subjective judgments, structured elicitation and calibration assessments are the tools that should be applied for major projects to ensure that decision makers use inputs based on reliable judgments.

Perhaps the most common omission in risk management practice is failing to perform structured expert judgement elicitations and calibration assessments. Research shows that bias and noise affect both SMEs and risk analysts alike. Most conclusions drawn and recommendations made at the time of decision making are based on judgments with unknown true answers. Given this postulate, a main goal of risk analysts to improve outputs is to improve the inputs: i.e., to improve the reliability, transparency, and defensibility of the collected judgments.

With processes, skills and knowledge, and resources in place, calibration assessments for expert judgments are relatively simple to implement and research has proven that they improve judgment forecasts of QRA inputs. The author encourages others to take a conscious look at how they currently collect subjective probabilities from SMEs, and advocates that a serious effort be made to calibrate these experts' inputs where appropriate to improve cost and schedule forecasts.

References

1. B. Flyvbjerg, C.-k. Hon and W. H. Fok, "Reference Class Forecasting for Hong Kong's Major Roadworks Projects," *Proceedings of the Institution of Civil Engineers,* vol. 169, no. Issue CE6, pp. pp. 17-24, 2016.
2. B. Mellers, E. Stone, T. Murray, A. Minster, N. Rohrbaugh, M. Bishop, E. Chen, J. Baker, Y. Hou, M. Horowitz, L. Ungar and P. Tetlock, Identifying and cultivating superforecasters as a method of improving probabilistic predictions, Perspectives on Psychological Science, 2015.
3. A. R. Colson and R. M. Cooke, "Expert Elicitation: Using the Classical Model to Validate Experts' Judgments," *Review of Environmental Economics and Policy,* vol. 12, no. 1, pp. 113-132, 2018.
4. AACE International, Recommended Practice No. 42R-08, Risk Analysis and Contingency Determination Using Parametric Estimating, Morgantown, WV: AACE International, Latest revision.
5. AACE International, Recommended Practice No. 40R-08, Contingency Estimating – General Principles, Morgantown, WV: AACE International, Latest revision.
6. A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science,* vol. 185, no. 4157, pp. 1124-31, 1974.
7. C. R. Sunstein, D. Kahneman and O. Sibony, Noise: A Flaw in Human Judgment, London: William Collins, 2021.
8. B. Flyvbjerg, Top Ten Behavioral Biases in Project Management: An Overview, Sage, 2021.

9. B. Benson, 1 September 2016. [Online]. Available: https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18. [Accessed 3 December 2021].
10. J. K. Hollmann, Project Risk Quantification, Sugarland, TX: Probabilistic Publishing, 2016, p. 94.
11. R. H. Thaler, "Misbehaving: How economics became behavioral," *Allen Lane,* p. 295, 2015.
12. R. Cooke and L. Goossens, "Procedures Guide for Structured Expert Judgments," Commission of European Communities, Delft, 2000.
13. C. Fox, "Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias Toward the Ignorance Prior," *Management Science,* vol. 51, no. 9, pp. 1417-1432, 2005.
14. AACE International, Recommended Practice No. 122R-22, Quantitative Risk Analysis Maturity Model, Morgantown, WV: AACE International, 2022.
15. AACE International, Recommended Practice No. 85R-14, Use of Decision Trees in Decision Making, Morgantwon: AACE International, Latest revision.
16. B. Flyvbjerg and D. Bester, "The Cost-Benefit Fallacy: Why Cost-Benefit Analysis Is Broken and How to Fix It.," *Journal of Benefit-Cost Analysis,* vol. 12, no. 3, pp. 395-419, 2021.
17. H. Abu-Abed, X. Guo, R. Kok, P. Lindsay and J. Tousignant-Barnes, Lessons Learned in Developing Cost Estimating Relationships, Morgantwon, WV, 2016.
18. AACE International, Recommended Practice No. 43R-08: Risk Analysis and Contingency Determination Using Parametric Estimating – Example Models as Applied for the Process Industries, Morgantown, WV: AACE International, Latest revision.
19. AACE International, Recommended Practice No. 62R-11, Risk Assessment: Identification and Qualitative Analysis, Morgantown, Morgantown, WV: AACE International, Latest revision.
20. P. Thomas, R. B. Bratvold and E. J. Bickel, "The Risk of Using Risk Matrices," *SPE Econ & Mgmt,* vol. 6, pp. 56-66, 2014.
21. J. A. Valdahl and S. A. Katt, The Planning Fallacy and Its Effect on Realistic Project Schedules, Morgantown, WV: AACE International Transactions, 2015.
22. John K. Hollmann, Estimate Validation and Bias Assessment: Ratio-to-Driver Method, Morgantwon, WV: AACE International Transactions, 2019.
23. AACE International, Recommended Practice No. 66R-11, Selecting Probability Distribution Functions for User in Cost and Schedule Risk Simulation Models, Morgantown, WV: AACE International, Latest revision.
24. G. Ponce de Leon and V. Puri, Removing the Early-Dates Bias in CPM Risk Analysis, Morgantwon, WV: AACE International Transactions, 2016.
25. Y. Raydugin, Non-Linear Probabilistic (Monte Carlo) Modeling of Systemic Risks, Morgantown, WV: AACE International Transactions, 2018.

Francisco Cruz Moreno, PE
PMA Consultants
fcruz@pmaconsultants.com