# MEASURING AND IMPROVING THE RELIABILITY OF PROBABILISTIC ASSESSMENTS IN PETROLEUM ENGINEERING

A Dissertation

by

MALIK KHALID M ALARFAJ

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Duane A. McVay |
| Co-Chair of Committee, | Mahmoud El-Halwagi |
| Committee Members, | Eduardo Gildin |
| | George W. Voneiff |
| Head of Department, | Mark Weichold |

August 2018

Major Subject: Interdisciplinary Engineering

# ABSTRACT

Petroleum industry performance has been consistently below expectations. This underperformance has been attributed in part to the existence of cognitive biases in project evaluation, resulting in poor project valuation and selection. It was demonstrated in the literature that chronic overconfidence and optimism (estimated distributions of project value too narrow and shifted positively), both common in industry, produce substantial disappointment (realized portfolio values less than estimated).

In this work, I aim to evaluate the impact of overconfidence as well as underconfidence (estimated distributions too wide) on portfolio performance, to determine if it is more beneficial to reduce biases and improve calibration or to reduce uncertainty, to provide a simple way of measuring biases from historical assessments, to determine the relationship between the number of probabilistic assessments and the accuracy of these measurements, and to determine guidelines for minimizing biases in new assessments using external adjustment.

I simulated the performance of projects selected in a typical portfolio of O&G projects to determine the effects of biases on portfolio performance and to compare reducing biases against reducing uncertainty. Next, I generated calibration curves for historical probabilistic assessments and used these curves to calculate different reliability measures. Then I generated different numbers of biased assessments and used them to determine the relationship between the number of assessments and the accuracy of the bias measurements. Furthermore, I used the calibration curve to adjust new forecasts and

measured the reliability of the new forecasts after adjustment as a function of the number of historical assessments and other parameters.

This research demonstrates that underconfidence is just as detrimental to portfolio performance as overconfidence. Decision error will be minimized and portfolio value will be maximized only when there is no bias in project estimation. Furthermore, I found that reducing biases consistently generates more value than reducing uncertainty. Moreover, this research shows that using more historical assessments to measure biases typically improves the accuracy of the bias measurements. However, even a low number of assessments is enough to detect moderate and extreme biases. Finally, this research shows that production forecasts that were updated frequently over time using newly available data and externally adjusted using the most recent bias measurements were superior in terms of calibration to forecasts that were not updated or externally adjusted.

The methods presented in this work can be used to measure and improve the reliability of probabilistic assessments in many petroleum engineering applications. Implementing these methods will result, over the long run, in the best calibrated assessments. Well-calibrated assessments result in better identification of superior projects and inferior projects, and ultimately, better investment decision making and increased profitability.

# DEDICATION

To those who kept me going

and to the giants on whose shoulders I stand.

# ACKNOWLEDGEMENTS

I would like to thank my committee chair Dr. Duane McVay for all the support, knowledge and guidance he gave me in the last few years, and also my committee co-chair, Dr. Mahmoud El-Halwagi, and my committee members Dr. Eduardo Gildin, and Prof. George Voneiff for their support and guidance.

I would also like to thank my friends and colleagues in Dr. McVay's research group for all the time, support, suggestions, and advices they have given me.

Furthermore, my thanks go to the Society of Petroleum Engineers for the copyright permission to reuse SPE-181430 "Improved Framework for Measuring the Magnitude and Biases in Project Evaluation".

My thanks and appreciation to my sponsor, Saudi Aramco, for giving me the opportunity to pursue my PhD degree and funding my education.

And finally, my forever gratitude to my mom, for encouraging me to follow my curiosity and keeping me in her prayers, my dad, for his unparalleled support, my wife and the love of my life, for bearing with me and taking care of me all those years, and finally my children, for bringing happiness and joy into my life.

# CONTRIBUTERS AND FUNDING SOURCES

## Contributors

This work was supervised by a dissertation committee consisting of Dr. Duane McVay of the Department of Petroleum Engineering, Dr. Mahmoud El-Halwagi of the Department of Chemical Engineering, Dr. Eduardo Gildin, and Prof. George Voneiff of the Department of Petroleum Engineering.

All work for the dissertation was completed independently by the student.

## Funding Sources

There are no outside funding contributions to acknowledge related to the research and compilation of this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

# CHAPTER I

# INTRODUCTION*

Several authors over several decades (Brashear et al. 2001, Capen 1976, Rose 2004) have observed that petroleum industry performance has been consistently below expectations. While this is painfully obvious during the industry downturn beginning in 2014, available evidence suggests that even when the industry is profitable, e.g., during the decade prior to the most recent downturn, it still performs substantially below expectations and its potential (Nandurdikar 2014). Many attribute this underperformance to cognitive biases in project evaluation, resulting in poor project valuation and selection. McVay and Dossary (2014) presented a simplified framework to estimate the cost of underestimating uncertainty. They demonstrated that chronic overconfidence and optimism (estimated distributions of project value too narrow and shifted positively), common in industry, produce substantial disappointment (realized portfolio values less than estimated), also common in industry.

## Status of the Question

While many authors have cited the qualitative benefits of reliably assessing uncertainty, only a few studies tried to assess the impact quantitatively. Welsh et al. (2007) modeled

* Part of this chapter is reprinted with permission from Alarfaj, M. K., and McVay, D. A. 2016. Improved Framework for Measuring the Magnitude and Impact of Biases in Project Evaluation. Presented at the SPE Annual Technical Conference and Exhibition, Dubai, UAE 26-28 September. SPE-181430-MS. https://doi.org/10.2118/181430-MS. Copyrights [2016] by Society of Petroleum Engineers.

the impact of three individual biases commonly found in project evaluation—overconfidence, trust, and availability. They showed that all three biases impact the estimated value of a project and result in a true net present value (NPV) considerably lower than its estimated NPV. However, they performed their analysis on a single-project basis and did not consider the overall impact of biases on a portfolio.

Begg and Bratvold (2008) explored the impact of prediction errors caused by the Optimizer's Curse or selection-mechanism systematic bias on an expected basis at the portfolio level. They found that this bias may not be as substantial as previously thought, especially when considering that the impact of other sources of bias may be significantly larger.

McVay and Dossary (2014) proposed a framework to estimate the value of assessing uncertainty by quantifying the monetary impact of biases on a portfolio of O&G projects. The essence of their framework is that biases that affect judgement and estimation in project evaluation can be rolled into two primary biases: overconfidence (underestimation of uncertainty, where the estimated distribution of an uncertain quantity is too narrow) and directional bias (where the estimated distribution is shifted in the optimistic or pessimistic direction). Their study showed that even moderate levels of overconfidence and optimism, which are common in the industry, could result in as much as 30-35% reduction, on average, from the estimated to the realized portfolio value. However, their framework did not include the effects of underconfidence (overestimation of uncertainty, where the estimated distribution of an uncertain quantity is too wide). While underconfidence is not currently common, as the industry hopefully improves in uncertainty estimation, it is of

interest to assess the impact on portfolio performance of possible overcorrection of overconfidence into underconfidence. Moreover, the authors used estimated distributions that were limited to truncated probability distributions. In practice, estimated distributions could be full distributions such as normal and lognormal distributions. Furthermore, while they showed that reducing the overconfidence bias reduces disappointment and decision error, they did not address which has greater benefit: reducing biases to make better calibrated assessments, or reducing uncertainty by acquiring more information and/or using more complex and detailed models.

Furthermore, none of these studies discussed in detail how to measure or eliminate these biases. McVay and Dossary (2014) suggested that the key to eliminating overconfidence is through a continual process of forecast tracking, lookbacks as actual values become available, checking calibration by comparing actual values to forecasts, and then using this calibration information to adjust new probabilistic assessments. Capen (1976) demonstrated how to use calibration results to externally adjust forecasts. For example, knowing from lookbacks and calibration that forecast P10-P90 ranges were too narrow, i.e., actually P30-P70 ranges, he simply plotted the forecast values versus the calibrated P30-P70 probabilities on probability paper (normal or lognormal) and extended the ranges to revised P10-P90 values.

Fondren et al. (2013) demonstrated how a database tracking system was used to externally adjust shale-gas probabilistic production forecasts to improve their reliability. To measure calibration, they used calibration plots in which the frequency of outcomes is plotted against the assessed probability of outcomes. They then implemented a

methodology similar to the one suggested by Capen (1976) to externally correct these forecasts. Landman and Goddard (2002) used model output statistics, a multiple linear regression technique, to recalibrate rainfall forecasts for extreme seasons over southern Africa using predictor values from a general circulation model and historical record of the predictand (regional rainfall indices). Piani et al. (2010) assumed that both normalized observed and simulated (estimated) distributions are well approximated by a gamma distribution and used a transfer function that can be derived graphically to correct the simulated distributions. This is similar to using calibration plots to externally adjust assessments; however, the latter can be considered more general since it is not restricted to a specific distribution and the CDFs do not need to be normalized. Mandel and Barnes (2014) used Karmarker's transformation, which utilizes a tuning parameter to improve the calibration of forecasts in strategic intelligence applications. Turner et al. (2014) used a combination of forecast aggregation and recalibration (adjustment) using a linear-in-log-odds function to generate a less-biased forecast. There is very little, if anything, in the literature that addresses the accuracy of these measures of reliability, or of biases, as a function of the number of assessments available.

**Research Objectives**

In my research, I aim to:

- evaluate the impact of confidence bias (including over and underconfidence) and directional bias on portfolio performance and determine which has greater benefit: reducing biases to make better calibrated assessments or reducing uncertainty by acquiring more information and/or using more complex models.

- determine the relationship between the number of probabilistic assessments and the accuracy of bias measurements.
- determine guidelines for minimizing biases in new assessments using external adjustment.

**Dissertation Outline**

In Chapter II, I generalized the McVay and Dossary (2014) framework to include underconfidence in addition to overconfidence. I also generalized it to include full estimated distributions (e.g., normal or lognormal), in addition to the truncated distributions used in the original framework. Using the generalized framework, I simulated the performance of projects selected in a typical portfolio of O&G projects to determine the effects of the confidence bias (including over and underconfidence) and directional bias (including positive and negative) on portfolio performance and to compare reducing biases against reducing uncertainty. Finally, I showed a simple method for measuring confidence and directional biases from calibration curves.

In Chapter II, I measured the reliability of probabilistic assessments by calculating and estimating the coverage rate, calibration score, confidence, and directional biases in biased probabilistic assessments that were generated using the generalized framework developed in Chapter II. I also used the generalized framework to generate different numbers of biased assessments and then determined the relationship between the number of assessments and the accuracy of the bias measurements. Next, I used the calibration curve to adjust new forecasts, and I measured the reliability of the new forecasts after adjustment as a function of the number of historical assessments and other parameters. I complete

Chapter III with a case study that compares different options for updating production forecasts and recommend the option that was superior to the others in terms of calibration.

Finally, in Chapter IV, I summarize the conclusions of these chapters and suggest future work.

# CHAPTER II

# IMPROVED FRAMEWORK FOR MEASURING THE MAGNITUDE AND IMPACT OF BIASES IN PROJECT EVALUATION*

## Overview

Several authors over several decades (Capen 1976; Brashear et al. 2001; Rose 2004) have observed that petroleum industry performance has been consistently below expectations. While this is painfully obvious during the industry downturn beginning in 2014, available evidence suggests that even when the industry is profitable, e.g., during the decade prior to the most recent downturn, it still performs substantially below expectations and its potential (Nandurdikar 2014). Many attribute this underperformance to cognitive biases in project evaluation, resulting in poor project valuation and selection. McVay and Dossary (2014) presented a simplified framework to estimate the cost of underestimating uncertainty. They demonstrated that chronic overconfidence and optimism (estimated distributions of project value too narrow and shifted positively), common in industry, produce substantial disappointment (realized portfolio values less than estimated), also common in industry.

In this work, we generalized their framework to include full estimated distributions (e.g., normal or lognormal), instead of the truncated distributions they employed. In

* Part of this chapter is reprinted with permission from Alarfaj, M. K., and McVay, D. A. 2016. Improved Framework for Measuring the Magnitude and Impact of Biases in Project Evaluation. Presented at the SPE Annual Technical Conference and Exhibition, Dubai, UAE 26-28 September. SPE-181430-MS. https://doi.org/10.2118/181430-MS. Copyrights [2016] by Society of Petroleum Engineers.

addition, we extended their framework to model underconfidence (estimated distributions too wide), and demonstrate that underconfidence is just as detrimental to portfolio performance as overconfidence. Decision error will be minimized and portfolio value will be maximized only when there is no bias in project estimation—i.e., neither overconfidence nor underconfidence and neither optimism nor pessimism. We compared the value gained from reducing biases to that from reducing uncertainty and found that reducing biases consistently generates more value than reducing uncertainty.

Using either framework, operators can quantitatively measure biases—overconfidence, underconfidence, optimism and pessimism—from lookbacks (comparing actual performance to probabilistic forecasts) and calibration plots. Once aware of the direction and magnitude of biases, operators have means for eliminating these biases in new forecasts through a combination of internal adjustment of uncertainty assessments, via training or ongoing feedback, and external adjustment of assessments using measurements of bias from calibration results.

**Introduction**

The industry has suffered massive losses in the recent oil price downturn. Xu and Bell (2016) reported that a sample of 59 US-based oil and gas producers and refiners posted combined net losses of nearly $102.9 billion in 2015 compared with net income of nearly $86.5 billion in 2014. Haynes and Boone (2018) reported that, as of March 2018, 144 North American oil and gas producers filed for bankruptcy since the beginning of 2015 with approximately $90.2 billion in cumulative secured and unsecured debt.

A cumulative body of evidence suggests that the severity of these losses can be attributed at least in part to unreliable uncertainty assessment caused by systematic biases. Literature review indicates that the difficulty and importance of assessing uncertainty were recognized early on. In his seminal work over 40 years ago, Capen (1976) warned about the difficulty of assessing uncertainty. He conducted several experiments with petroleum engineers and showed that they are chronically overconfident. When asked to produce 90% confidence intervals, they produced intervals that corresponded to a 32% confidence interval on average. He concluded that people tend to be a lot prouder (more confident) of their probability ranges than they should be. He reported that, even when people have been warned, probability ranges tend to be too small; they do slightly better but still cannot bring themselves to make their probability ranges wide enough. Finally, he warned against the negative consequences of poorly quantifying uncertainty.

Unfortunately, it seems that the industry has realized little improvement in its ability to reliably assess uncertainty and perform in line with expectations. Industry performance in the last decade of the twentieth century was dismal. Brashear et al. (2001) noted that the average return of the largest U.S.-based E&P companies in the 1990s was around 7% despite using project-hurdle rates generally of 15% or more. Furthermore, Rose (2004) reported that exploration departments of most E&P companies delivered only about half of the new reserves they promised. In the Norwegian sector of the North Sea, all of the active participants delivered only 38% of the expected reserves.

The industry's financial performance may have improved in the years prior to the current oil slump, but this likely happened because of high oil prices rather than systematic

improvement in uncertainty estimation. It is possible that high oil prices may have caused the industry to relax and make even worse project-selection decisions. Indeed, Merrow (2012) reported that, since 2003, the success rate for petroleum megaprojects—those that exceed one billion USD—declined from 50% to 22%, while the success rate for non-petroleum megaprojects stayed constant at around 50% for the same period. Nandurdikar (2014) claimed the improvement in the industry's financial performance was mainly because of unexpected high oil prices. In reality, the estimated ultimate recovery (EUR) updated 2 years after startup fell outside the 80% confidence range of the original EUR estimates 40% of the time—twice the expected rate and usually on the negative side. In other words, the updated EURs were usually less than promised. Moreover, he noted that most businesses did not recognize the value erosion because the actual oil price was higher than assumed at sanction. Therefore, it appeared that their financial results were better than expected, while in reality they left much on the table.

These reports indicate that the industry continues to perform below expectations. Why does it continue to underperform? Brashear et al. (2001) argued that use of evaluation methods that do not account for the full range of uncertainty contributed to the industry's underperformance. Rose (2004) attributed it mainly to chronic biases in estimating key evaluation parameters that control project evaluations. McVay and Dossary (2014) hypothesized that part of the reason the industry continues to underestimate uncertainty is lack of appreciation of the monetary (quantitative) impact of biases on industry performance.

While many authors have cited the qualitative benefits of reliably assessing uncertainty, only a few studies (Welsh et al. 2007; Begg and Bratvold 2008; Hdadou and McVay 2014; McVay and Dossary 2014) tried to assess the impact quantitatively. To address this problem, McVay and Dossary (2014) proposed a framework to estimate the value of assessing uncertainty. Their model estimated the monetary impact of overconfidence (underestimation of uncertainty, where the estimated distribution of an uncertain quantity is too narrow) and directional bias (where the estimated distribution is shifted in the optimistic or pessimistic direction). Their study showed that even moderate levels of overconfidence and optimism, which are common in the industry, could result in as much as 30-35% reduction, on average, from the estimated to the realized portfolio value.

However, the McVay and Dossary framework did not include the effects of underconfidence (overestimation of uncertainty, where the estimated distribution of an uncertain quantity is too wide). Although underconfidence is not currently common, as the industry hopefully improves in uncertainty estimation, it is of interest to assess the impact on portfolio performance of possible overcorrection of overconfidence into underconfidence. Furthermore, the authors used estimated distributions that were limited to truncated probability distributions. While they applied various levels of biases to simulated portfolios similar to those available to a large E&P company, they did not attempt to measure biases quantitatively from actual probabilistic assessments post-development. In this work, we propose a new framework for modeling both overconfidence and underconfidence in combination with directional bias, and which is not limited to truncated estimated distributions. In the remainder of this paper, we compare

results from the new framework to the previous framework, and we show also the effects of underconfidence on portfolio performance. We end by demonstrating how to measure over/underconfidence and directional bias from probabilistic assessments using calibration plots, and how to adjust new assessments using these measured values.

**Previous Framework**

We build on McVay and Dossary's (2014) framework for modeling the impact of biases. The essence of their framework is that biases that affect judgement and estimation tend to affect the following:

- The uncertainty or variability of the estimate (usually in the direction of overconfidence, or underestimation of uncertainty).
- The central tendency of the estimate (usually in the direction of optimism)
- Or, both the uncertainty and the central tendency of the estimate.

Thus, all biases can be rolled into two primary biases: overconfidence and directional bias (optimism or pessimism). Overconfidence is the failure to consider all the possible outcomes. Optimism can manifest when one ignores or fails to consider possible negative outcomes or gives them less weight than equally probable positive outcomes. On the other hand, pessimism occurs when one ignores or fails to consider possible positive outcomes or gives them less weight than equally probable negative outcomes.

Suppose that you were asked to provide an estimate for an unknown quantity, for example, project value. You can provide a single value as your estimated project value. You can also define your uncertainty about this value by providing a standard deviation (SD) or specifying a complete distribution. In this work, we call this the estimated project-

value distribution. Such distributions result from typical limited-resources assessments and would include biases typically present in O&G project evaluations. These biases make the estimated project-value distribution different from your "true" project-value distribution. The "true" value distribution as defined by Smith and Winkler (2006) is the distribution that would result from an unlimited-resources assessment. In other words, the true project-value distribution would be obtained if you had unlimited time, money, and computational ability. McVay and Dossary (2014) clarified that these unlimited resources can only be used to analyze existing data and cannot be used to obtain further data. Begg et al. (2014) stated that uncertainty is related to the estimator's state of information and that the state of information is particular to a person (or a company). Thus, the true project-value distribution is personal; i.e., different estimators can have different knowledge and assessment processes and, thus, can have different "true" but valid unlimited-resources project-value distributions. McVay and Dossary (2014) clarified that, ultimately, "true" project-value distributions are those that are "reliable," or perfectly calibrated. By reliable they mean that over a large number of similar estimations, the frequencies of the outcomes would correspond to their assigned probabilities. For example, events or assessments that has been assigned 10% probability should occur 10% of the time, those that has been assigned a 50% probability should occur 50% of the time, and those that has been assigned 90% probability should occur 90% of the time. Calibration is discussed further in later sections.

McVay and Dossary introduced two parameters to define the relationship between the true and estimated project-value distributions. The overconfidence parameter was defined

as a parameter that ranges from 0.0 to 1.0 and specifies the fraction of the true distribution not sampled by the estimated distribution in the limited-resources assessment. Therefore, a value of 0.0 denotes that the entire true distribution was sampled, and no biases are present. On the other hand, an overconfidence value greater than 0.0 denotes that only a subset of the true distribution is sampled. This results in an estimated distribution that is narrower than the true distribution.

The directional bias parameter was defined as a parameter that ranges from -1.0 to 1.0 and it specifies the location of the estimated distribution relative to the true distribution. A directional bias value of -1 means that only the lowest possible outcomes of the true distribution were considered; i.e., the estimated distribution is shifted to the left of the true distribution. On the other hand, a directional bias value of +1 means that only the highest possible outcomes of the true distribution were considered; i.e., the estimated distribution is shifted to the right of the true distribution **(Fig. 2.1**). In the McVay and Dossary model, there can be no directional bias if there is no overconfidence because, in this situation, the estimated distribution is the same as the true distribution.

McVay and Dossary (2014) did not clearly distinguish between directional bias (DB) and optimism-pessimism bias (OPB). A bias in the positive direction could mean optimism or pessimism depending on the parameter. For example, for a value-based parameter such as Net Present Value (NPV), a positive DB value is considered optimism because expecting a greater value than reality is of benefit to the estimator. On the other hand, for a cost-based parameter such as Capital Expenditure (CapEx), a negative DB is considered optimism because expecting lower cost than reality is of benefit to the

estimator. While it is possible for some uncertain parameters to have no optimism-pessimism bias associated with directional bias, virtually all uncertain parameters affecting petroleum project evaluation will have associated optimism-pessimism bias.

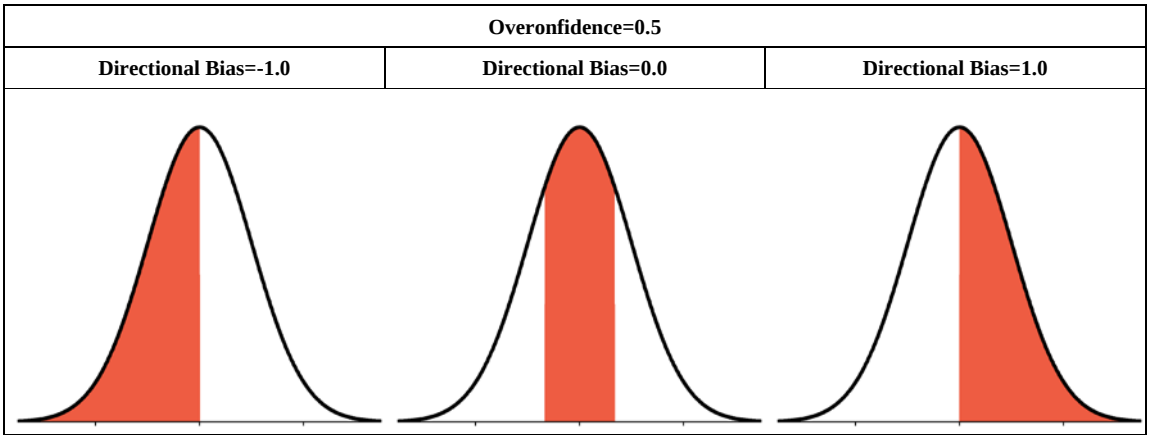


**Fig. 2.1—In the McVay and Dossary (2014) framework, the overconfidence parameter specifies the fraction of the true distribution (black curve) not sampled by the estimated distribution (red area) and the directional bias specifies the location of the estimated distribution relative to the true distribution.**

*Modeling overconfidence and directional bias in previous framework*

To model the estimated distribution, the true distribution is simply truncated at the tails. **Fig. 2.2** shows the true and estimated probability distribution functions (PDF) assuming a normal distribution for the true. The truncated PDF $f_e$ represents the estimated distribution while the full PDF $f_t$ represents the true distribution. The cumulative distribution function (CDF) corresponding to $f_t$ is $F_t$ (**Fig. 2.3**).

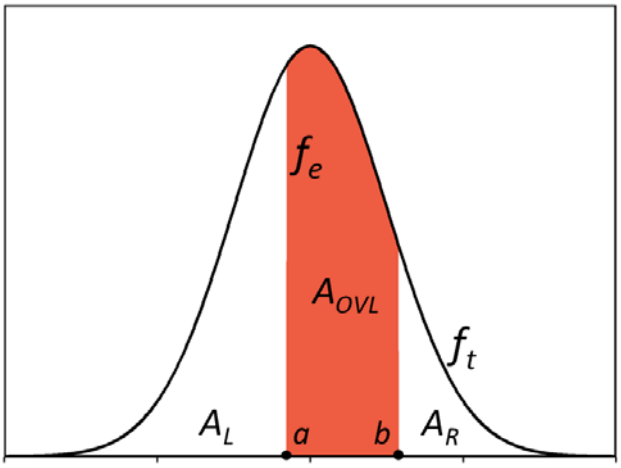


**Fig. 2.2—True and estimated distributions for a directional bias value of 0.5 and overconfidence value of 0.5 assuming a normal true distribution.**

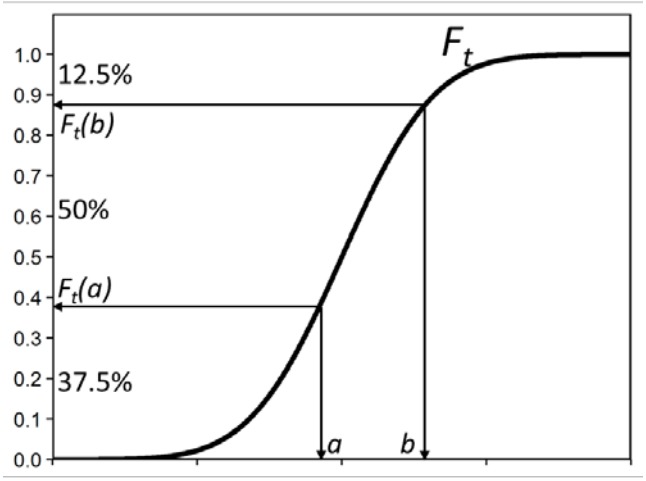


**Fig. 2.3—CDF of the true distribution for a directional bias value of 0.5 and overconfidence value of 0.5 assuming a normal true distribution.**

Mathematically, overconfidence can be calculated as the sum of the true-distribution areas not included in the estimated distribution (the sum of the unshaded areas under $f_t$), while the directional bias can be calculated from the ratio of the left or right unshaded area to the sum of the unshaded areas.

Let *a* and *b* denote the truncation points for the estimated distribution (Fig. 2.2). Also, let $A_L$ denote the unshaded area on the left and $A_R$ denote the unshaded area on the right. Then, the confidence bias parameter for overconfident estimated distributions can be calculated as follows:

$$CB_{OC} = A_L + A_R = F_t(a) + [1 - F_t(b)] = 1 + F_t(a) - F_t(b) \quad (2.1)$$

The directional bias parameter in the presence of overconfidence can be calculated using either the left or right unshaded area as follows:

$$DB_{OC} = 2\left(\frac{A_L}{CB_{OC}}\right) - 1 = 1 - 2\left(\frac{A_R}{CB_{OC}}\right) \quad (2.2)$$

$$DB_{OC} = 2\left(\frac{F_t(a)}{1+F_t(a)-F_t(b)}\right) - 1 = 1 - 2\left(\frac{1-F_t(b)}{1+F_t(a)-F_t(b)}\right) \quad (2.3)$$

*Modeling underconfidence and directional bias using truncated distributions*

McVay and Dossary (2014) did not include underconfidence in their framework. Because it is possible to have underconfidence and because there is potential value in considering the effects of underconfidence, we extended their truncated-estimated-distribution model to include it. For underconfidence, we simply flipped the distributions so that the full distribution is the estimated distribution and the truncated distribution is the true distribution. **Fig. 2.4** shows the true and estimated PDFs assuming a normal distribution for the estimated. This time, the truncated PDF $f_t$ represents the true distribution while the full PDF $f_e$ represents the estimated distribution. The CDF corresponding to $f_e$ is $F_e$ (**Fig. 2.5**).

Underconfidence ranges from -1 for complete underconfidence (no information about the uncertain quantity) to 0 for no underconfidence, which is the same as 0 overconfidence (the estimated distribution is the same as the true distribution).

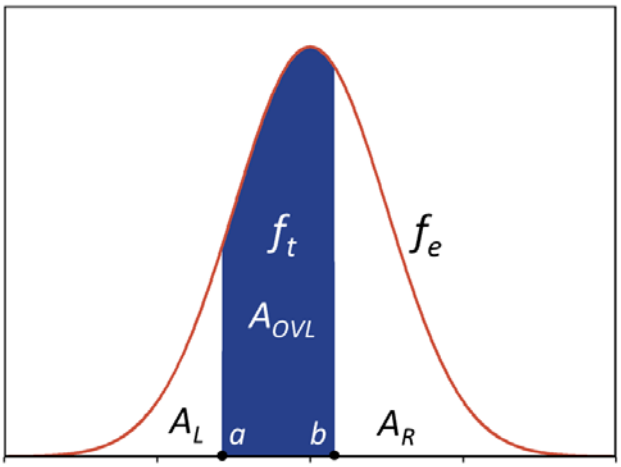


**Fig. 2.4—True and estimated distributions for a directional bias value of 0.5 and underconfidence value of -0.5 assuming a normal estimated distribution.**

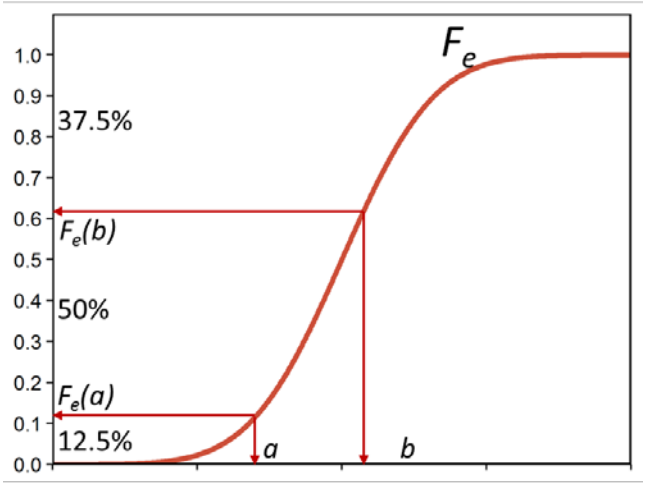


**Fig. 2.5—CDF of the estimated distribution for a directional bias value of 0.5 and underconfidence value of -0.5 assuming a normal estimated distribution.**

Mathematically, underconfidence can be calculated as the negative of the sum of areas not included in the true distribution (the sum of the unshaded areas under $f_e$), while the directional bias can be calculated from the ratio of the left or right unshaded area to the sum of the unshaded areas.

Let *a* and *b* denote the truncation points for the true distribution (Fig. 2.4). Also, let $A_L$ denote the unshaded area on the left and $A_R$ denote the unshaded area on the right. Then, the confidence bias parameter for underconfident estimated distributions can be calculated as follows:

$$CB_{UC} = -(A_L + A_R) = -\{F_e(a) + [1 - F_e(b)]\} = F_e(b) - F_e(a) - 1 \quad (2.4)$$

For underconfidence, we change the directional bias equation to:

$$DB_{UC} = 1 + 2\left(\frac{A_L}{CB_{UC}}\right) = -2\left(\frac{A_R}{CB_{UC}}\right) - 1 \quad (2.5)$$

$$DB_{UC} = 1 + 2\left(\frac{F_e(a)}{F_e(b) - F_e(a) - 1}\right) = -2\left(\frac{1 - F_e(b)}{F_e(b) - F_e(a) - 1}\right) - 1 \quad (2.6)$$

As with overconfidence, a negative DB value means that the estimated distribution is shifted to the left relative to the true distribution. Conversely, a positive DB value means that the estimated distribution is shifted to the right relative to the true distribution.

**Generalized Framework**

The truncated estimated distribution used in the previous framework provides an easy way to visualize and understand over/underconfidence and directional bias. It also simplifies the mathematical computations. However, in practice, we typically do not work with truncated distributions in project evaluation or probabilistic estimates in general. We need a framework that can handle the kinds of probabilistic distributions we commonly use in

assessments. Thus, we generalized the McVay and Dossary (2014) framework to accept both truncated and full distributions for the estimated distribution in the case of overconfidence, and both truncated and full distributions for the true distribution in the case of underconfidence. The new framework does not change the definitions of directional bias and over/underconfidence, but rather introduces a new way of calculating them that allows both truncated and full distributions. It also combines the overconfidence and underconfidence portions of the confidence bias (CB) parameter so that CB values from -1 to 0 denote underconfidence and values from 0 to 1 denote overconfidence.

*Modeling overconfidence in new framework*

In Fig. 2.2, if the estimated distribution starts at point *a* and ends at point *b*, the confidence bias parameter for overconfident distributions is equal to the cumulative area under the true distribution from -∞ to *a* plus the area from *b* to ∞. However, this model cannot be applied to estimated distributions that are not bounded, such as the normal distribution, which ranges from -∞ to ∞ (**Fig. 2.6**) and the lognormal distribution which ranges from 0 to ∞. Therefore, we propose a slightly more general definition for calculating the areas for overconfident distributions. Overconfidence is the area under the true distribution PDF that is not included in the area under the estimated distribution PDF (Fig. 2.6). Let $f_t$ denote a PDF of the true distribution and $f_e$ a PDF of the estimated distribution. The shaded area in Fig. 2.6 is the area under both PDFs. It is called the overlapping coefficient ($A_{OVL}$) and it can be calculated as follows (Bradley 2006):

$$A_{OVL} = \int_{-\infty}^{\infty} \min[f_t(x), f_e(x)]\, dx \quad (2.7)$$

Since $A_{OVL}$ is equal to the shaded area, we can define the overconfidence portion of the confidence bias parameter using $A_{OVL}$ as:

$$CB_{OC} = 1 - A_{OVL} \quad (2.8)$$

Directional bias is calculated the same as before for overconfident estimates:

$$DB_{OC} = 2\left(\frac{A_L}{CB_{OC}}\right) - 1 = 1 - 2\left(\frac{A_R}{CB_{OC}}\right) \quad (2.9)$$

For overconfident estimates:

$$A_L = \int_{-\infty}^{Mo_{f_e}} \max[f_t(x) - f_e(x), 0]\, dx \quad (2.10)$$

$$A_R = \int_{Mo_{f_e}}^{\infty} \max[f_t(x) - f_e(x), 0]\, dx \quad (2.11)$$

$$Mo_{f_e} = \text{Mode}(f_e) \quad (2.12)$$

Note that these definitions will not change CB and DB parameter values for truncated distributions. **Fig. 2.7** shows that the area under the true normal distribution $f_t$ that is not included in the estimated truncated-normal distribution $f_e$ is the same as the cumulative area under the true distribution from -∞ to *a* plus the area from *b* to ∞**.** In other words, calculating CB and DB parameters for a truncated estimated distribution using the generalized framework equations will produce the same values as the McVay and Dossary (2014) equations. Therefore, the generalized framework is backward compatible with the previous framework.

**Fig. 2.8** shows the relationship between the estimated distribution and the true distribution as a function of overconfidence and directional bias parameters using both truncated and full estimated distributions, for a true standard-normal distribution with mean of 0 and SD of 1. There are no differences in the estimated expected value (EV) for

full versus truncated distributions when the DB is zero because of the symmetry. The differences in estimated EV between full and truncated distributions increase as the DB value becomes more extreme (both in the positive and the negative directions) because of the increased difference in distribution shapes at the extremes.

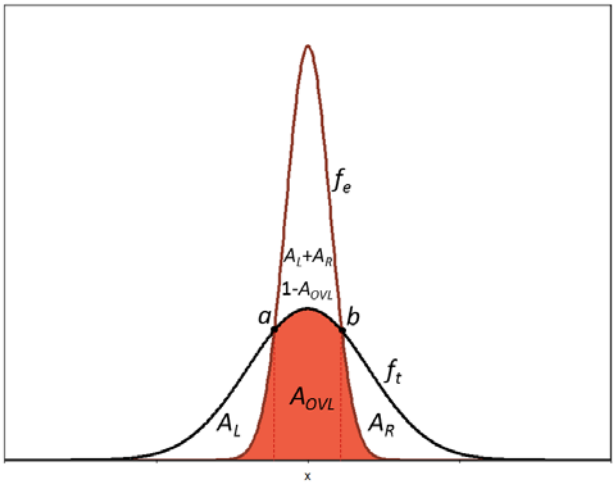


**Fig. 2.6—Under the new framework, overconfidence is calculated from the overlapping area between the true distribution $f_t$ and the estimated distribution $f_e$.**

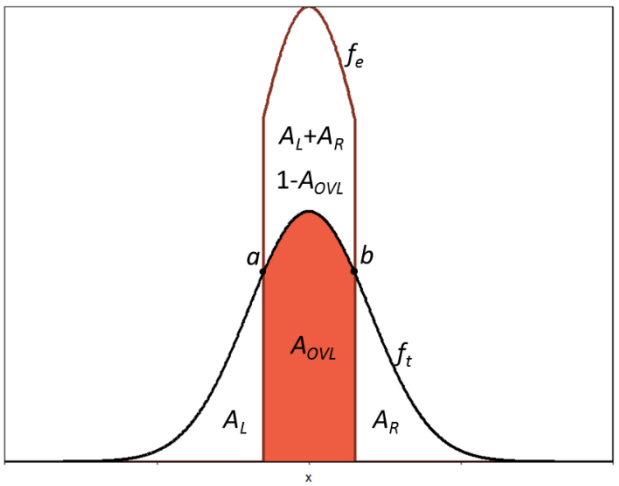


**Fig. 2.7—The overlapping area between the true distribution $f_t$ and the truncated estimated distribution $f_e$ is the same as the shaded area in the previous framework.**

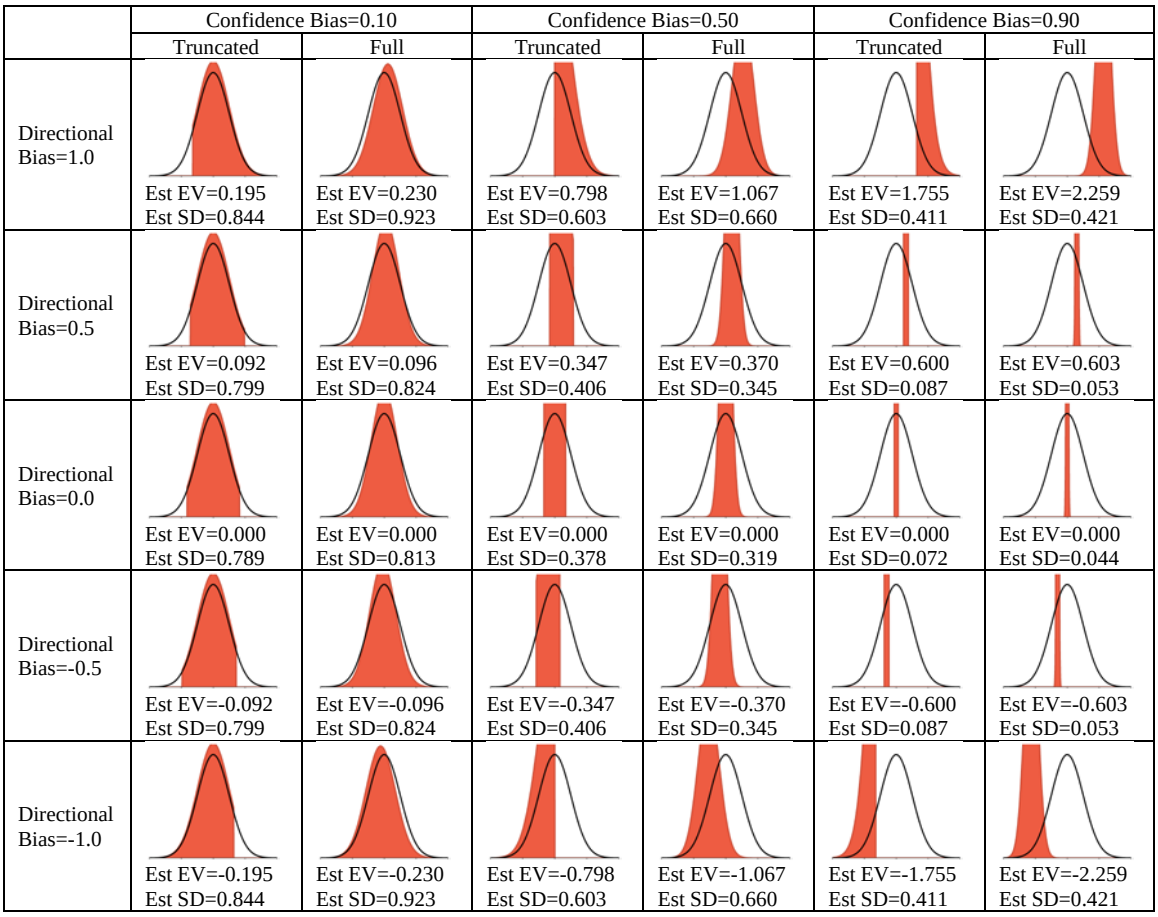


**Fig. 2.8—Relationship between the estimated (red curve) and the true (black curve) distributions in the generalized framework as a function of confidence bias and directional bias parameters. The "Truncated" columns show the relationship between a truncated estimated distribution and a full true distribution similar to the previous framework while the "Full" columns show the relationships between two full distributions [adapted from McVay and Dossary (2014)].**

*Modeling underconfidence in new framework*

Modeling the underconfident portion of CB will follow similar principles to modeling the overconfident portion. Also, although we have not shown them, there will be similar but inverted relationships between true and estimated distributions for underconfidence as was shown for overconfidence in Fig. 2.8. For underconfident portion of CB, the estimated distribution $f_e$ is wider than the true distribution $f_t$ (**Fig. 2.9**). We start by calculating the

overlapping coefficient ($A_{OVL}$) just as we did previously. Then, the underconfidence portion of the confidence bias parameter is defined as:

$$CB_{UC} = A_{OVL} - 1 \quad (2.13)$$

The directional bias parameter is calculated the same way we did in the truncated underconfidence model, that is:

$$DB_{UC} = 1 + 2\left(\frac{A_L}{CB_{UC}}\right) = -2\left(\frac{A_R}{CB_{UC}}\right) - 1 \quad (2.14)$$

For underconfident estimates:

$$A_L = \int_{-\infty}^{Mo_{f_t}} \max[f_e(x) - f_t(x), 0]\, dx \quad (2.15)$$

$$A_R = \int_{Mo_{f_t}}^{\infty} \max[f_e(x) - f_t(x), 0]\, dx \quad (2.16)$$

$$Mo_{f_t} = \text{Mode}(f_t) \quad (2.17)$$

**Fig. 2.10** shows an underconfident estimated distribution with a positive directional bias $f_{ep}$, and an underconfident estimated distribution with a negative directional bias $f_{en}$ relative to the true distribution $f_t$.

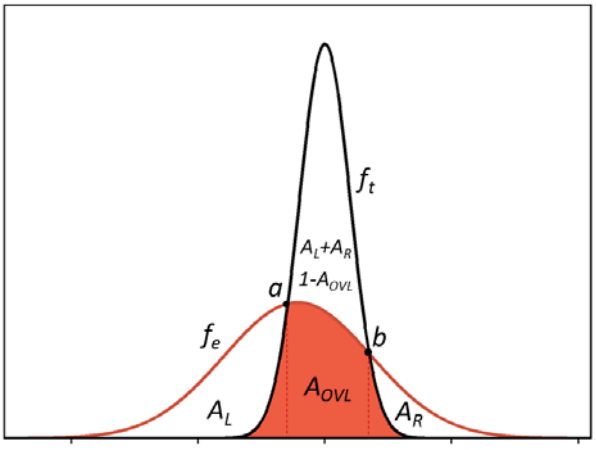


**Fig. 2.9—Underconfidence is calculated from the overlapping area between the true distribution $f_t$ and the estimated distribution $f_e$.**

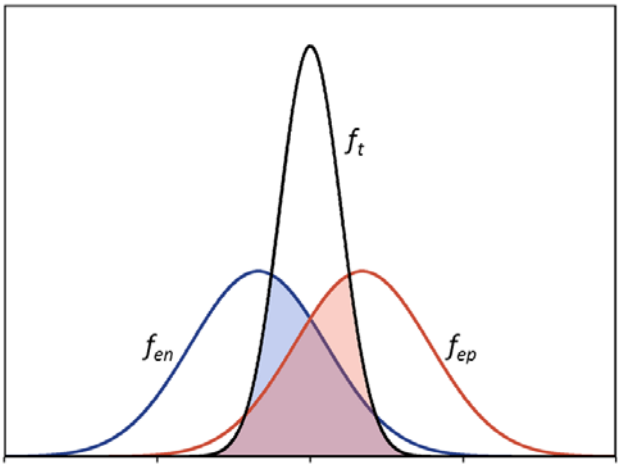


**Fig. 2.10—The relationship between the true distribution $f_t$ and an underconfident estimated distribution with positive DB $f_{ep}$ and an underconfident estimated distribution with negative DB $f_{en}$.**

*Generalized framework summary*

In summary, a confidence bias parameter can be defined as follows:

$$CB = \begin{cases} 1 - A_{OVL}, & for\ Overconfidence \\ A_{OVL} - 1, & for\ Underconfidence \end{cases} \quad \text{(2.18)}$$

and the directional bias parameter is:

$$DB = \begin{cases} 2\left(\frac{A_L}{CB}\right) - 1 = 1 - 2\left(\frac{A_R}{CB}\right), & for\ Overconfidence \\ 1 + 2\left(\frac{A_L}{CB}\right) = -2\left(\frac{A_R}{CB}\right) - 1, & for\ Underconfidence \end{cases} \quad \text{(2.19)}$$

Although we used normal distributions to demonstrate the relationship between the true and the estimated distributions in Fig. 2.8, these equations would also apply to lognormal distributions and potentially any unimodal distribution that can be defined by a continuous PDF.

We remind the reader of the difference between directional bias (DB) and optimism-pessimism bias (OPB). Just like with overconfident distributions, with underconfident

distributions, a value-based parameter such as Net Present Value (NPV) would have a positive DB value for optimistic estimates and a cost-based parameter such as Capital Expenditure (CapEx) would have a negative DB for optimistic estimates (**Table** 2.1).

| Parameter | Positive DB | Negative DB |
|---|---|---|
| Value-based parameter | Optimism | Pessimism |
| Cost-based parameter | Pessimism | Optimism |

Table 2.1—The relationship between directional bias and optimism-pessimism bias.

### Modeling the Impact of Biases

We started by modeling the McVay and Dossary (2014) project selection experiments. The most significant change we made was to use full lognormal distributions to represent estimated distributions, instead of the truncated lognormal distributions they used. Each experiment began by generating a pool of 100 projects typical of those available to a large O&G company. For each project, we generated a true Capital-Expenditure (CapEx) distribution and a true Present-Value-of-Operating-Cash-Flow (PVOCF) distribution. We assumed that the CapEx and PVOCF distributions are lognormal and independent (uncorrelated). The means of the PVOCF and the CapEx distributions were sampled from global distributions with the parameters in **Table 2.2**. The standard deviations of the CapEx and the PVOCF distributions were generated by multiplying the sampled means by a fraction sampled from a PERT distribution with minimum 0.3, mode 0.8, and maximum 1.3.

| Parameter | Global Distribution | Mean | Standard Deviation | Shift |
|---|---|---|---|---|
| PVOCF mean | Shifted lognormal | 750 MM | 750 MM | 300 MM |
| CapEx mean | Shifted lognormal | 600 MM | 600 MM | 100 MM |

Table 2.2—The PVOCF and CapEx means are sampled from a shifted lognormal distribution [from McVay and Dossary (2014)].

Next, for each project, we applied CB and OPB to the distributions of true CapEx and PVOCF to obtain the distributions of estimated CapEx and PVOCF. Similar to McVay and Dossary, we applied the same amount of bias to both CapEx and PVOCF. For example, in cases with CB of 0.5, we have applied CB=0.5 to both CapEx and PVOCF. We also applied OPB equally to both distributions using the relationship between DB and OPB explained in Table 2.1. From the CapEx and PVOCF distributions, we calculated Net Present Value (NPV=PVOCF-CapEx) and Investment Efficiency (IE=NPV/CapEx) distributions for use in project selection, for both the true and estimated distributions.

Next, we conducted unconstrained-budget and constrained-budget project selections from the pool of 100 projects. For the unconstrained-budget scenario, we selected projects that had an estimated expected NPV (EV) > 0. For the constrained-budget scenario, projects were ranked in terms of decreasing estimated expected IE. Top projects were successively selected until a CapEx budget of $5 billion was exhausted. For the last project selected, we took an appropriate percentage of the project to exactly fill the CapEx budget. For each scenario (unconstrained and constrained budgets), we calculated the estimated portfolio EV by adding the estimated EVs of the individual projects selected based on

their estimated EV or E(IE). We also calculated the realized portfolio EV by adding the *true* EVs of the individual projects selected based on their estimated EV or E(IE). Finally, we calculated the best-possible portfolio EV by adding the *true* EVs of projects selected based on their *true* EV and E(IE) rather than estimated EV and E(IE), i.e., projects that estimators would have selected if they were unbiased.

For each experiment, we calculated the portfolio's *expected disappointment* (ED) and *expected decision error* (EDE). Expected disappointment was defined as the estimated portfolio EV minus the realized portfolio EV. It can be positive (disappointment) or negative (pleasant surprise). It was calculated as a percentage of the estimated portfolio EV as follows:

$$ED\%E = \frac{estimated\ portfolio\ EV - realized\ portfolio\ EV}{estimated\ portfolio\ EV} \times 100\% \quad (2.20)$$

Expected decision error was defined as the best-possible portfolio EV minus the realized portfolio EV; this is the portion of disappointment that results from selecting the wrong projects. It was also calculated as a percentage of the estimated portfolio EV as follows:

$$EDE\%E = \frac{best\text{-}possible\ portfolio\ EV - realized\ portfolio\ EV}{estimated\ portfolio\ EV} \times 100\% \quad (2.21)$$

To help illustrate the impact of biases on portfolio EV, we introduce the portfolio EV attainment, which is the realized portfolio EV as a percentage of the best-possible portfolio EV:

$$EVA\%BP = \frac{realized\ portfolio\ EV}{best\text{-}possible\ portfolio\ EV} \times 100\% \quad (2.22)$$

Finally, we repeated this process in a Monte Carlo simulation to determine the expected

values of ED%E, EDE%E, and EVA%BP over thousands of different 100-project pools. For simplicity, we left the parameter names the same instead of adding another expectation operator.

*Overconfidence*

**Figs. 2.11** and **2.12** show ED%E for the unconstrained and constrained budget cases using the new framework. As noted by McVay and Dossary (2014), ED%E increases monotonically as optimism increases. With pessimism, the estimator experiences negative expected disappointment (post-decision pleasant surprise). However, this pleasant surprise does not come without a cost. Although a pessimistic estimator realizes more EV than estimated (Figs. 2.11 and 2.12), the pessimism combined with overconfidence results in reduced value from the best possible (**Figs. 2.13 and 2.14**) because of decision error (**Figs. 2.15 and 2.16)**. That is, the estimator makes incorrect project selections because of the biases and, as a result, the realized portfolio value is lower than the best-possible value (i.e., with projects selected using the true, unbiased distributions). The results also show that the realized EV is maximized, and expected disappointment and expected decision error are minimized, when CB=0.

The results of our simulations assuming overconfidence were close to McVay and Dossary (2014), even though we used different distribution shapes (full instead of truncated) for the estimated CapEx and PVOCF. **Figs. 2.17 and 2.18** show the differences in ED%E and EVA%BP between using full and truncated estimated distributions for a confidence bias value of 0.5 for both the constrained and unconstrained budget scenarios. The differences are more pronounced at extreme values of directional bias. However, it is

obvious from the plots that the differences are not significant in either scenario. The similarity of results from the previous framework and our framework are as expected. Projects were selected based on their EVs, and Fig. 2.8 shows that the EVs for different values of CB and OPB are similar in both frameworks except at extreme values of OPB. The interested reader can compare the ED%E and EDE%E results to those of the previous framework by comparing Figs. 2.11, 2.12, 2.15 and 2.16 in this paper to Figs. 3, 4, 7 and 8 in McVay and Dossary (2014).

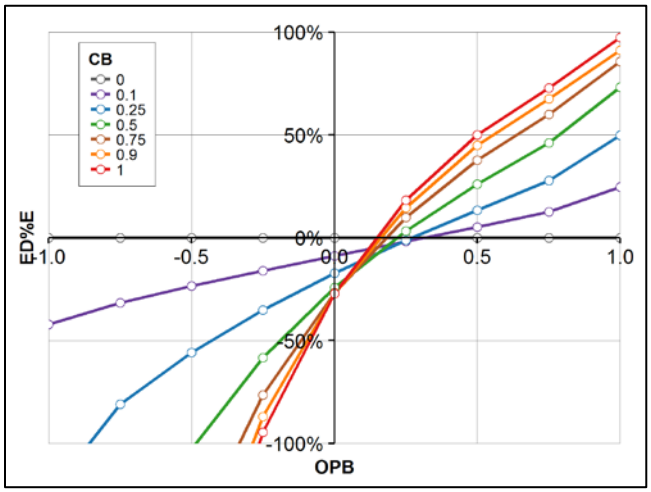


**Fig. 2.11—Expected disappointment for the unconstrained budget scenario and overconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

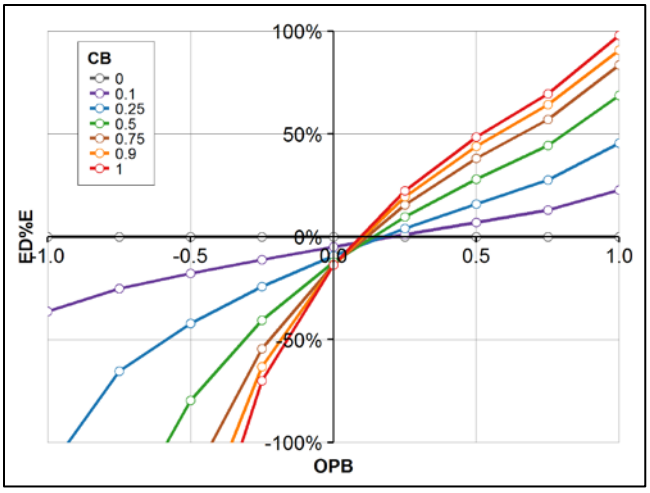


**Fig. 2.12—Expected disappointment for the constrained budget scenario and overconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

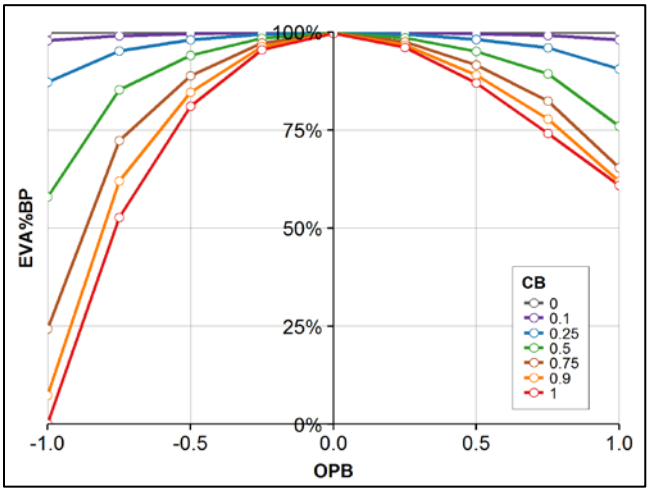


**Fig. 2.13—Expected value attainment for the unconstrained budget scenario and overconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

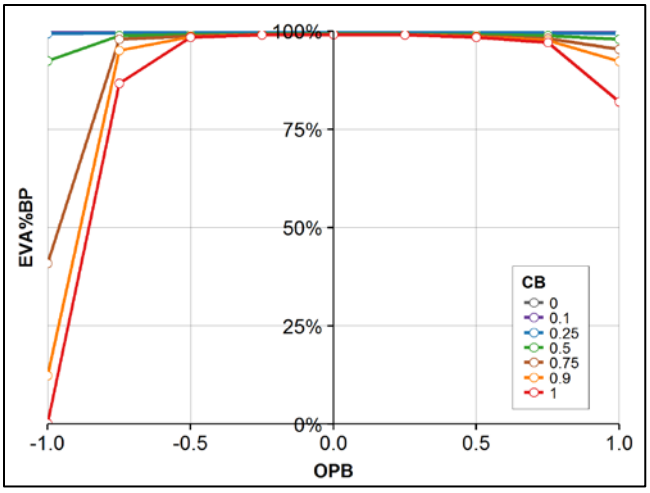


**Fig. 2.14—Expected value attainment for the constrained budget scenario and overconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

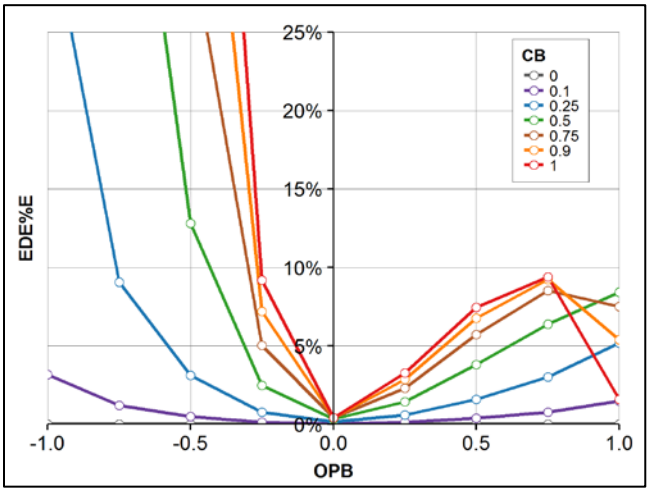


**Fig. 2.15—Expected decision error for the unconstrained budget scenario and overconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

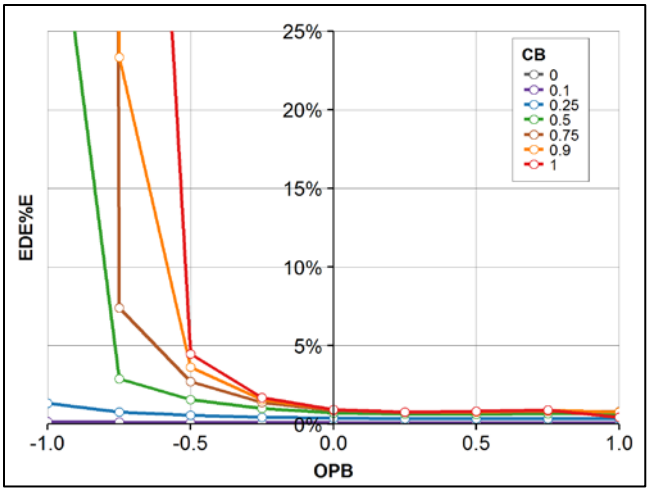


**Fig. 2.16—Expected decision error for the constrained budget scenario and overconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

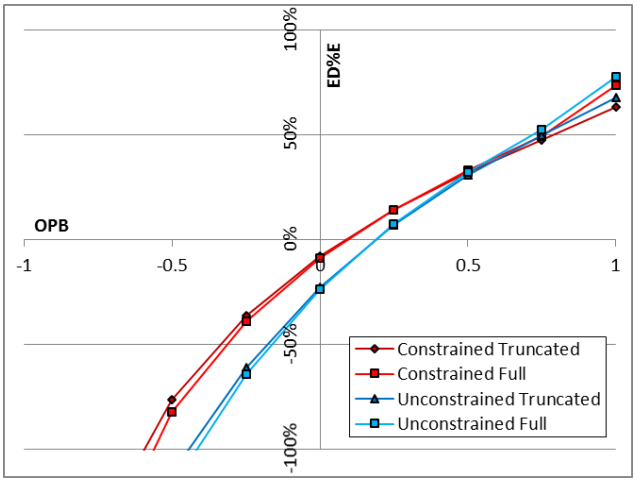


**Fig. 2.17—A comparison of ED%E resulting from using full and truncated estimated distributions for an overconfidence value of 0.5 with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

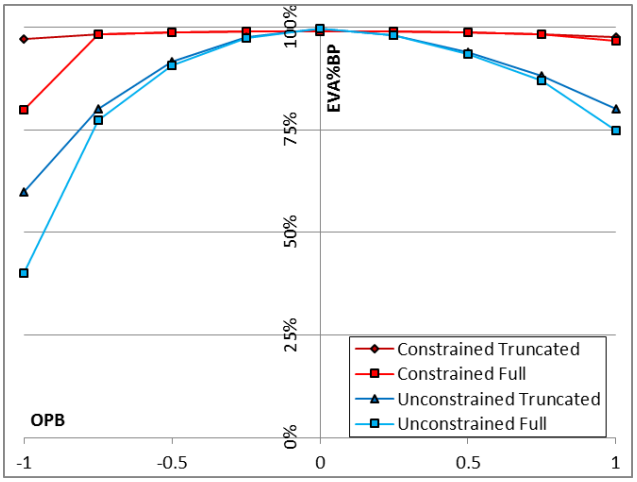


**Fig. 2.18—A comparison of EVA%BP resulting from using full and truncated estimated distributions for an overconfidence value of 0.5 with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

Figs. 2.11 and 2.12 show only the expected value of the portfolio percentage disappointment and not the range of possible portfolio outcomes. **Figs. 2.19 and 2.20** show that, with zero CB and OPB, it is possible to be disappointed or pleasantly surprised with an individual portfolio, while on average over many portfolios the estimator experiences zero percentage disappointment. Figs 2.19 and 2.20 also show that it is possible to experience zero or negative disappointment for an individual portfolio with 0.5 CB and 0.5 OPB. However, it is more likely that the estimator will be disappointed on average over many portfolios.

Biases affect not only the expected value of the estimated portfolio NPV, but also the uncertainty (or risk) of the estimated portfolio, as reflected in the SD. **Fig. 2.21** shows the best-possible, realized, and estimated portfolio NPV distributions for a CB value of 0.5 and OPB values of -0.5, 0.0, and 0.5, respectively. These figure show that the estimated

portfolio distribution is narrower than that of the realized and best-possible. That is, the estimator is underestimating the portfolio uncertainty (or risk) because of biases.

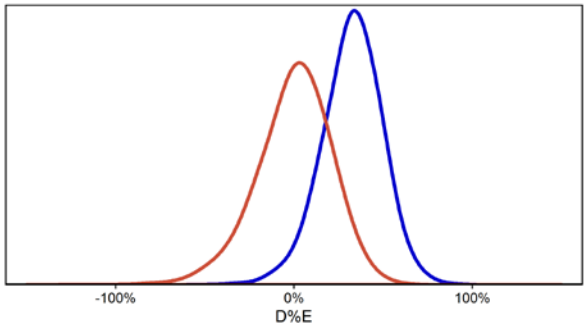


**Fig. 2.19—Distribution of portfolio percentage disappointment for 0 CB and 0 OPB (red curve) and 0.5 CB and 0.5 OPB (blue curve) for the unconstrained budget scenario.**

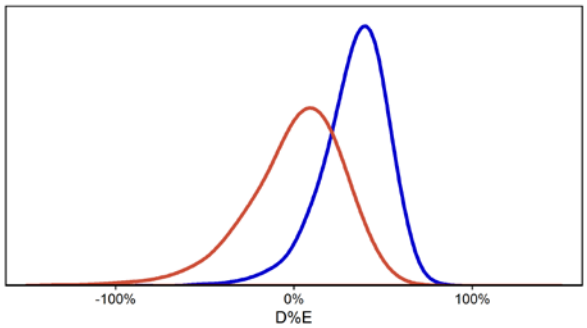


**Fig. 2.20—Distribution of portfolio percentage disappointment for 0 CB and 0 OPB (red curve) and 0.5 CB and 0.5 OPB (blue curve) for the constrained budget scenario.**

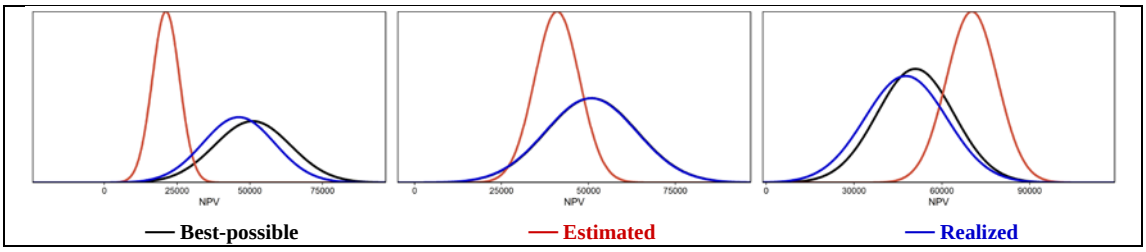


**Fig. 2.21—Best-possible, realized, and estimated portfolio NPV distributions for a CB value of 0.5 and OPB values of-0.5 (left), 0.0 (middle), and 0.5 (right).**

*Underconfidence*

While underconfidence is rare, it is possible. However, unlike overconfidence, it is unlikely to be highly underconfident because this results in extremely wide ranges. Thus, we limited our simulations of underconfidence to low-to-moderate levels of underconfidence.

ED%E with underconfidence (**Figs. 2.22 and 2.23**) is similar in character to ED%E with overconfidence (Figs. 2.11 and 2.12); ED%E increases monotonically as optimism increases. However, the OPB crossover point between disappointment and pleasant surprise occurs at OPB of about -0.25 to -0.75 (pessimism) for underconfidence (Figs. 2.22 and 2.23) versus OPB of about 0.1 to 0.2 (optimism) for overconfidence (Figs. 2.11 and 2.12). We attribute this asymmetric behavior to the use of lognormal distributions for CapEx and PVOCF. Similarly, the EVA%BP behavior with underconfidence (**Figs. 2.24 and 2.25**) is similar to its behavior with overconfidence (Figs. 2.13 and 2.14), at least for the unconstrained budget case. One noticeable difference is that the EVA%BP goes to nearly 100% at 0 OPB for all CB in the case of overconfidence but not in the case of underconfidence. This is because the estimated expected NPV remains close to the true expected NPV (within 5-10%) regardless of the overconfidence level, while the estimated expected NPV with underconfidence varies significantly because of the extreme width (or more precisely, dispersion) of the estimated distributions (remember that the standard deviation of the estimated distribution approaches $\infty$ as CB approaches -1). The extreme width of lognormal estimated distributions results in significant differences in estimated EV, which causes more projects with positive NPV to be rejected and more projects with

negative NPV to be accepted, which increases decision error and lowers value attainment. Behavior is complicated further for the constrained-budget scenario; as the magnitude of underconfidence increases, CapEx distributions become so large that they exhaust the budget limit quickly and, as a result, fewer projects are selected.

Finally, just like EDE%E with overconfidence (Figs. 2.15 and 2.16), the EDE%E with underconfidence (**Figs. 2.26 and 2.27**) decreases as OPB increases from -1 to 0 and largely increases as OPB increases from 0 to 1 for the unconstrained budget case, and largely decreases as OPB increases from -1 to 1 for the constrained budget case.

In summary, just as with overconfidence, underconfidence in combination with optimism-pessimism bias results in decision error (incorrect project selections), disappointment, and reduced portfolio value. Just as pessimism is not the remedy for optimism, underconfidence is not the remedy for overconfidence. An operator will make the best decisions, eliminate disappoint, and maximize portfolio value when completely unbiased—i.e., when neither overconfident nor underconfident, and neither optimistic nor pessimistic (**Figs. 2.28 and 2.29**). These figures show that expected disappointment is minimum and expected value attainment is maximum near 0 CB and 0 OPB.

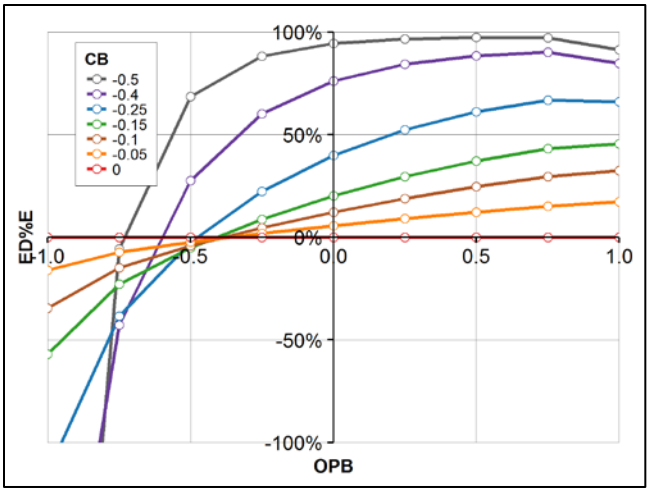


**Fig. 2.22—Expected disappointment for the unconstrained budget scenario and underconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

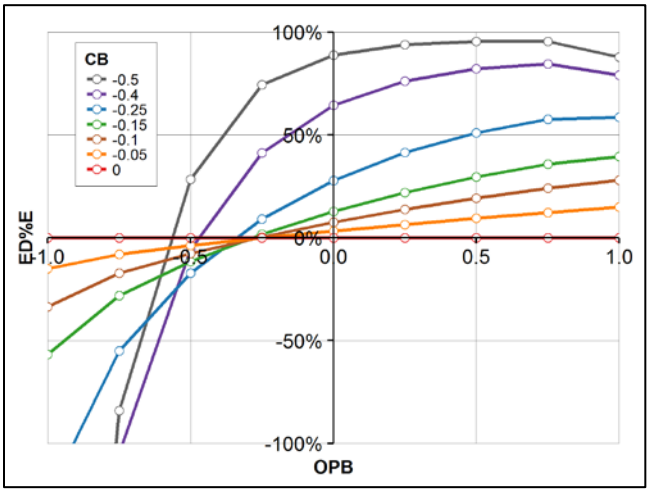


**Fig. 2.23—Expected disappointment for the constrained budget scenario and underconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

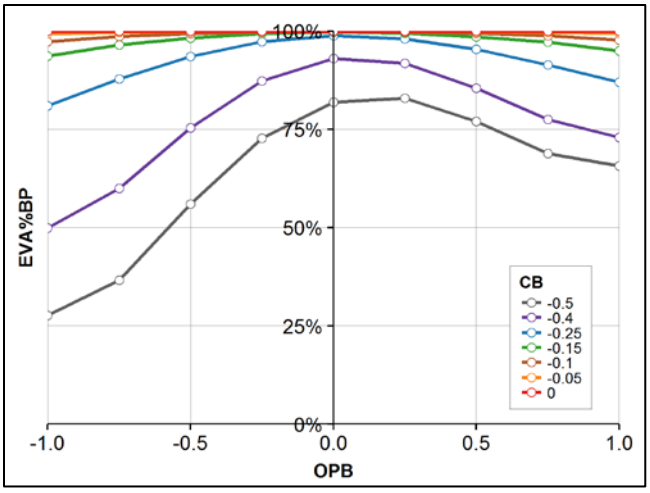


**Fig. 2.24—Expected value attainment for the unconstrained budget scenario and underconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

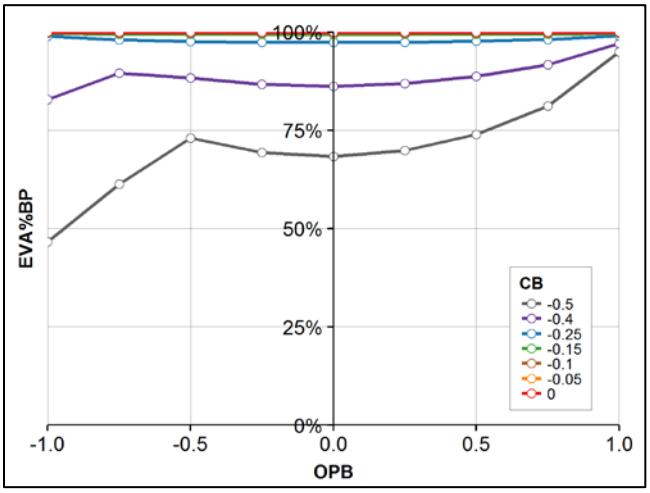


**Fig. 2.25—Expected value attainment for the constrained budget scenario and underconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

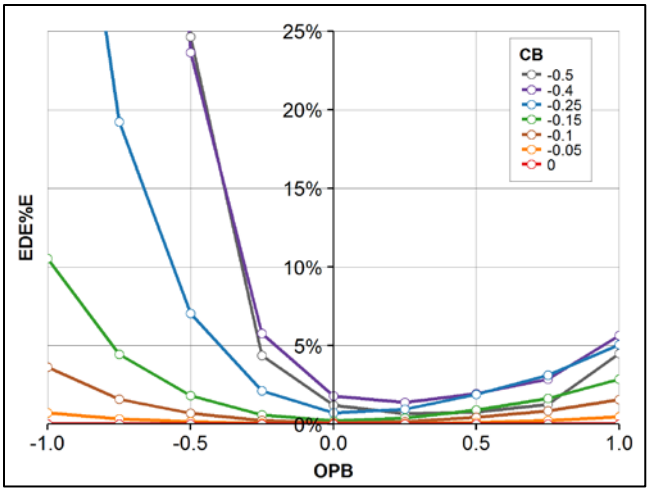


**Fig. 2.26—Expected decision error for the unconstrained budget scenario and underconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

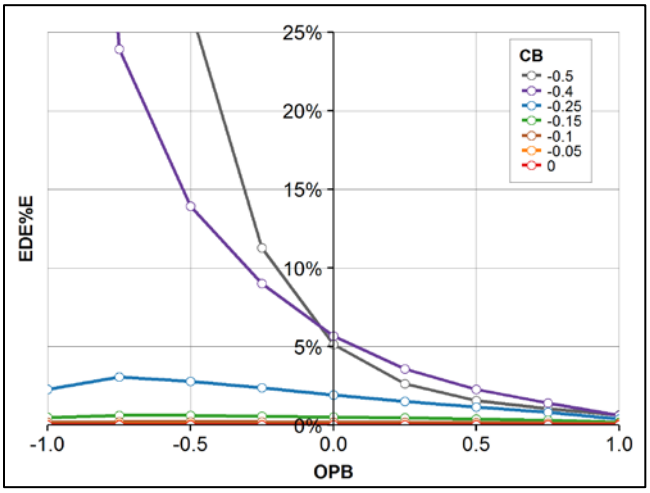


**Fig. 2.27—Expected decision error for the constrained budget scenario and underconfidence using the new framework with OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

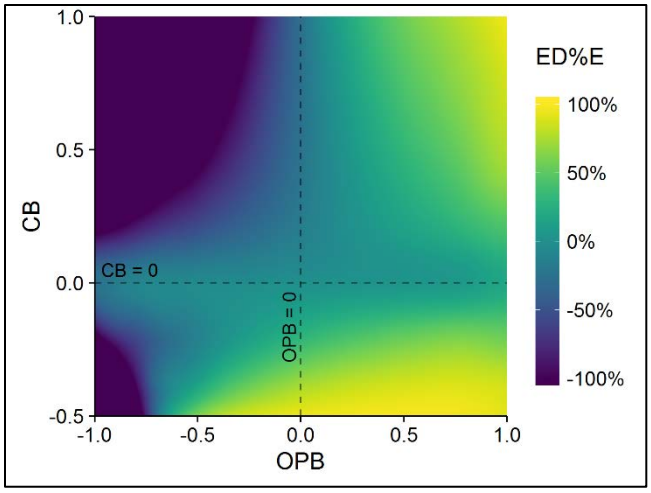


**Fig. 2.28—Expected disappointment using unconstrained budget scenario and CB from -0.5 (moderate underconfidence) to 1 (complete overconfidence), and OPB from -1 (complete pessimism) to 1 (complete optimism).**

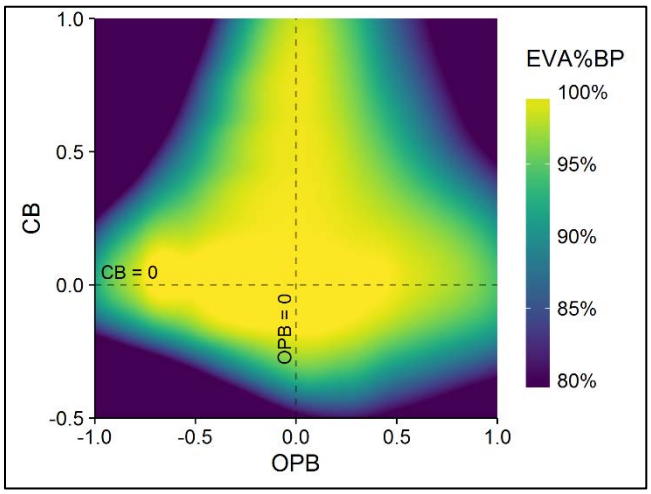


**Fig. 2.29—Expected value attainment using unconstrained budget scenario and CB from -0.5 (moderate underconfidence) to 1 (complete overconfidence), and OPB from -1 (complete pessimism) to 1 (complete optimism).**

## Eliminate biases or reduce uncertainty?

In the previous section, we showed the detrimental effects of biases on portfolio performance. It follows that reducing these biases should subsequently reduce

disappointment and improve value attainment. In the industry, however, the typical path taken when faced with uncertainty is to try to reduce it by acquiring more information and/or by using more complex and detailed analysis methods. Considering this, a question arises: when making decisions under uncertainty, would we gain more benefit if we try to reduce biases and make better calibrated assessments, or try to reduce uncertainty by acquiring more information and/or using more complex models? In this section, we seek to provide insights into the answer to that question.

We do that by comparing the results of our previous simulations (which show the impact of reducing biases) to a number of scenarios which have reduced uncertainty in the estimated distributions. We assume that acquiring more information or using more complex models will reduce uncertainty (represented by the SD) in the true PVOCF distribution (leaving the CapEx distribution unchanged). In this analysis, we assumed that no additional biases were added to the estimate in the uncertainty reduction process; i.e., we assumed that CB and OPB remained the same when the true SD decreased. We believe that this is a conservative assumption and that it is possible, if not likely, that biases will increase as true SD decreases (e.g., as a result of anchoring to the previous biased estimate) and the benefits of reducing uncertainty may be eroded. To keep the same global portfolio properties after reducing uncertainty, we sampled the PVOCF mean for the reduced-uncertainty distribution ($f_{t\text{-}reduced}$) from the original true PVOCF distribution ($f_t$) that we used in the previous section. Next, we reduced the standard deviation of the new distribution ($f_{t\text{-}reduced}$) by a specified percentage in each scenario. Reduction in uncertainty ranged from 10% to 90% (**Fig. 2.30**). This simulates a case where acquiring more

information or increasing analysis/model complexity moves the expected value of the new PVOCF distribution ($f_{t\text{-}reduced}$) closer to the true value and reduces the uncertainty about it. After we calculate the reduced-uncertainty true PVOCF distribution ($f_{t\text{-}reduced}$) properties, we calculate the corresponding estimated distributions ($f_e$ and $f_{e\text{-}reduced}$) by applying biases, as described in the previous section (**Fig. 2.31**).

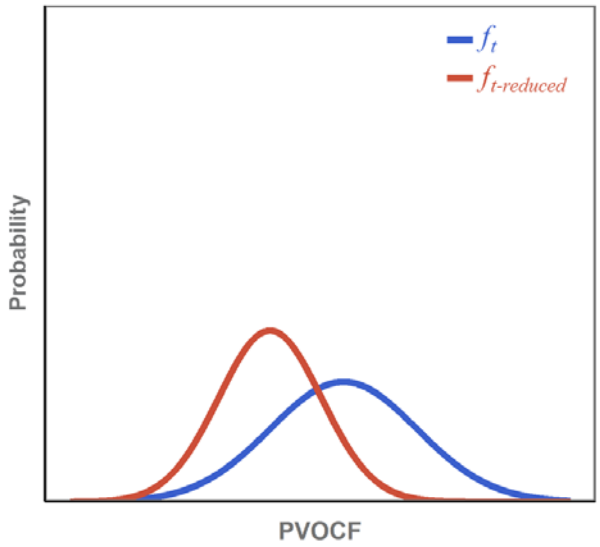


**Fig. 2.30—The original true PVOCF distribution (ft) is sampled from the global distribution (Table 2.2). Then, the reduced uncertainty PVOCF distribution (ft-reduced) mean is randomly sampled from ft. Finally, the standard deviation of ft-reduced is reduced by a specific percentage for each scenario (25% in this figure).**

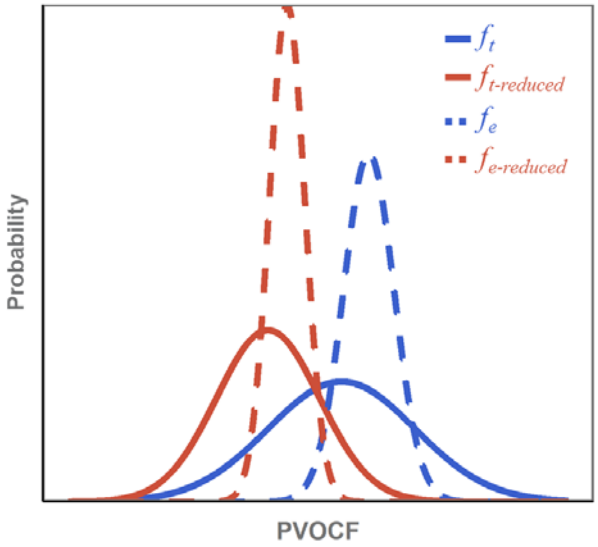


**Fig. 2.31—The estimated distributions ($f_e$, and $f_{e\text{-reduced}}$) corresponding to each true distribution is calculated using the generalized framework. In this figure, both estimated distributions have a CB = 0.5 and OPB = 0.5 relative to their corresponding true distributions.**

**Figs. 2.32 and 2.33** show the changes in ED%E and EVA%BP, respectively, as functions of the amount of uncertainty reduction. Reducing uncertainty reduces expected disappointment and increases expected portfolio value attainment, as expected. However, the improvement is insignificant with moderate overconfidence and optimism-bias levels (around 0.5 CB and 0.5 OPB), which are apparently common in industry. That is, the value attained from reducing uncertainty with moderate levels of biases is small given global portfolio properties assumed in this work. Comparing Figs. 2.32 and 2.33 to Figs. 2.11 and 2.13, it is apparent that reduction in disappointment and improvement in value attainment is significantly greater if operators focus on reducing biases rather than reducing uncertainty. Indeed, disappointment can be decreased to zero and value attainment can reach 100% with elimination of biases, but not by reducing uncertainty

(except by reduction to zero uncertainty, which is unachievable in practice). Similar ED&E reductions and EVA%BP improvements are observed at other CB levels; they are less pronounced at lower CB levels (**Figs. 2.34 and 2.35** for CB=0.1) and more amplified at higher CB levels (**Figs. 2.36 and 2.37** for CB=0.9), but they are still less than the improvements from reducing biases. While this analysis is based on a single set of assumed global portfolio properties and would benefit from further research, it appears that there is more to gain from eliminating biases than from reducing uncertainty. This by no means suggests that operators should not attempt to reduce uncertainty (if value-of-information calculations demonstrate there is benefit in doing so). However, these results do suggest that, to improve decision making, operators should focus first on eliminating biases and second on reducing uncertainty. As demonstrated in the previous section, operators will make the best decisions and maximize portfolio value only when they are completely unbiased in project evaluation.

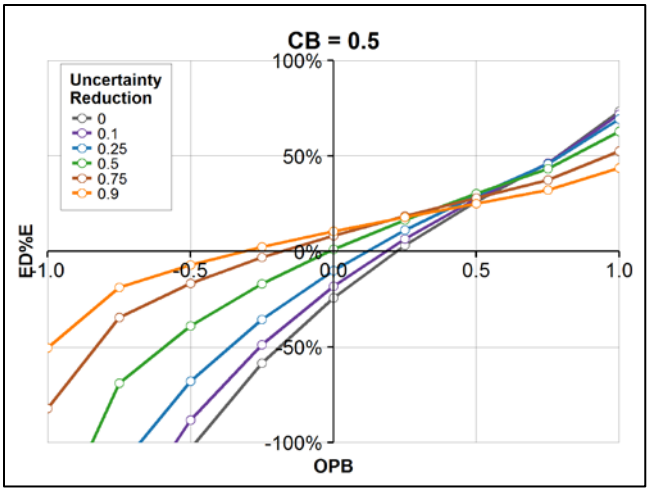


**Fig. 2.32—Expected disappointment as a function of uncertainty reduction using unconstrained budget scenario with CB equals 0.5, and OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

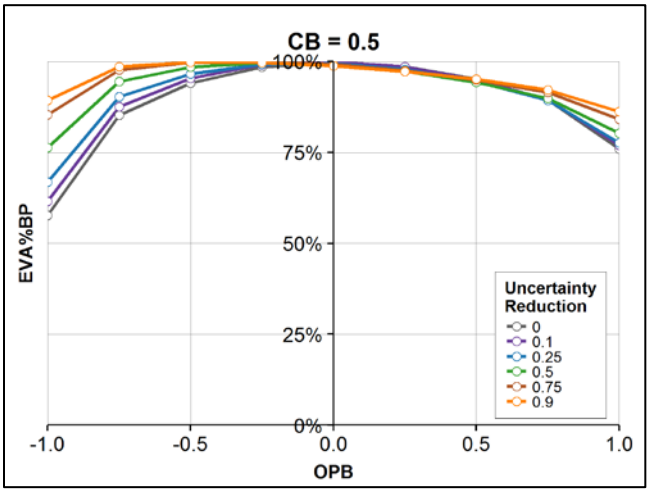


**Fig. 2.33—Expected value attainment as a function of uncertainty reduction using unconstrained budget scenario with CB equals 0.5, and OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

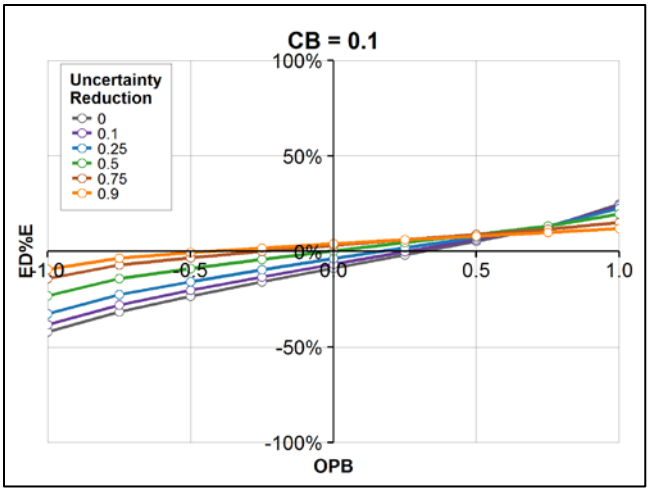


**Fig. 2.34—Expected disappointment as a function of uncertainty reduction using unconstrained budget scenario with CB equals 0.1, and OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

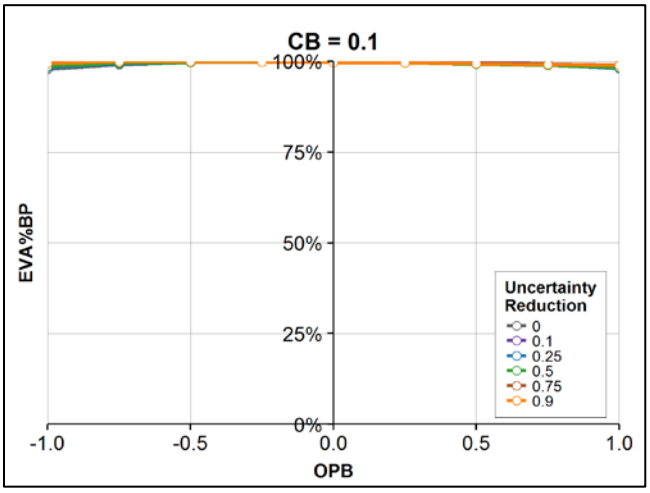


**Fig. 2.35—Expected value attainment as a function of uncertainty reduction using unconstrained budget scenario with CB equals 0.1, and OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

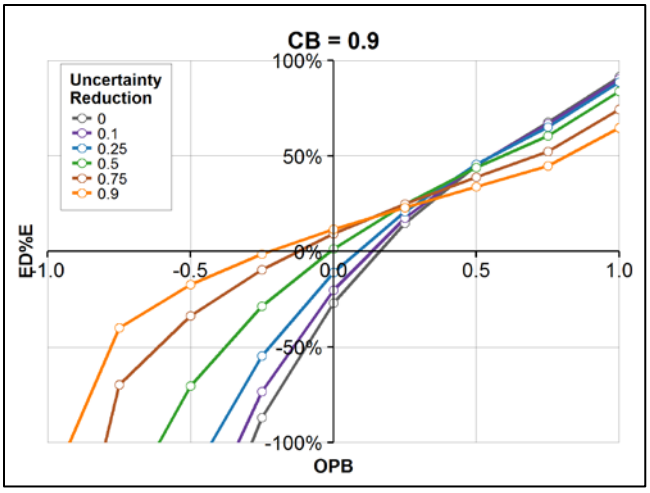


**Fig. 2.36—Expected disappointment as a function of uncertainty reduction using unconstrained budget scenario with CB equals 0.9, and OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

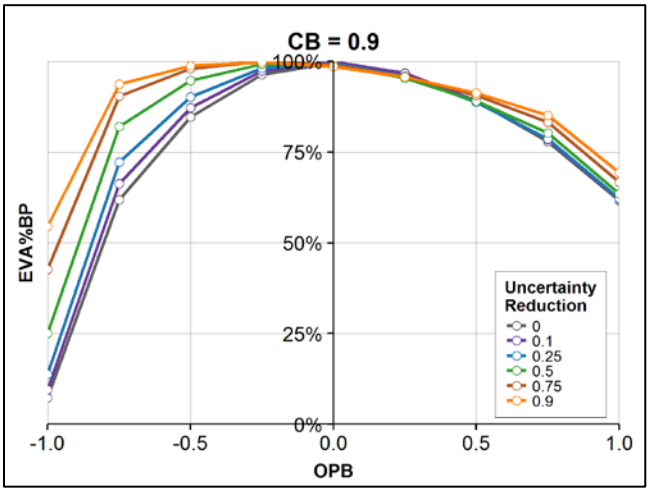


**Fig. 2.37—Expected value attainment as a function of uncertainty reduction using unconstrained budget scenario with CB equals 0.9, and OPB ranging from -1 (complete pessimism) to +1 (complete optimism).**

## Measuring Confidence and Directional Biases

In previous sections, we showed that both overconfidence and underconfidence, particularly in combination with optimism and pessimism, can produce decision error and

disappointment, and significantly erode portfolio value. Considerable petroleum industry evidence suggests that chronic overconfidence and optimism have resulted in long periods of large disappointment and significant underperformance of the industry. This suggests that operators should seek with all diligence to eliminate these biases from forecasts and assessments. How can operators eliminate biases? First, they need to be aware that they are biased. This requires recording and tracking of probabilistic forecasts and lookbacks to compare actual performance to forecasts. With this information they can then measure the direction and magnitude of their biases and the reliability of their forecasts using calibration plots. In a calibration plot (**Fig. 2.38**), the frequency of outcomes—the proportion of estimates that became true (labeled "Proportion Correct" in the plot) —is plotted against the assessed probability of outcomes. The proportion correct can be calculated by dividing the number of true forecast assessments assigned a specific probability (e.g., the number of times the actual value is less than or equal to the P10 value) by the total number of forecasts assigned that same probability (e.g., the total number of P10 assessments). We are using a cumulative probability convention where the P10 is the low number and the P90 is the high number. A group of probabilistic forecasts are probabilistically *reliable* if the actual values are less than or equal to the P10 estimates about 10% of the time, are less than or equal to the P90 estimates about 90% of the time, and the same for all other cumulative probabilities. Thus, reliable probabilistic forecasts that quantify the "true" uncertainty will fall on the unit-slope line on a calibration plot. Fig. 2.38 shows the calibration for a set of forecasts that are both overconfident (slope is less than 1 because the distributions are too narrow) and directionally biased in the positive

direction (shifted upward, relative to the unit-slope line). Overconfidence increases as the slope decreases, and directional bias increases as the upward shift increases. This figure represents optimism bias if the forecasts are all value-based forecasts; optimistic cost-based forecasts would be shifted downward (Table 2.1). Thus, calibration plots provide a quick, visual indication of the types and degrees of the major types of bias—overconfidence, underconfidence, optimism and pessimism.

*Measuring CB and DB using the previous framework*

In addition to visual indication, we can estimate these biases quantitatively with calibration plots. One of the advantages of the McVay and Dossary (2014) framework is that it provides convenient interpretation of these plots. If the true distributions are normal or lognormal and the estimated distributions are truncations of these distributions, then the calibration curve in the extreme (very large number of forecasts) will be a straight line (e.g., Fig. 2.38). Furthermore, one can easily derive the relationships between the slope and intercept of the line and the confidence and directional bias parameter values. Let *m* be the slope of the calibration line and *a* its intercept. For overconfident forecasts—slope less than 1—the confidence bias parameter is:

$$CB_{OC} = 1 - m \quad (2.23)$$

and the directional bias parameter is:

$$DB_{OC} = \frac{2a}{1-m} - 1 = \frac{2a}{CB_{OC}} - 1 \quad (2.24)$$

Using the truncated framework described earlier, we generated a very large number of estimated distributions using CB = 0.4 and DB = 0.3. Then, we created the calibration plot in Fig. 2.38 for this set of forecasts by randomly sampling from the true distributions to

get the actual values. The calibration line has slope of 0.6 and intercept of 0.26. Substituting these values into Eqs. 2.23 and 2.24, we back calculate the input confidence and directional biases values used to generate this set of forecasts:

$$CB_{OC} = 1 - m = 1 - 0.6 = 0.4$$
$$DB_{OC} = \frac{2a}{CB_{OC}} - 1 = \frac{2 \times 0.26}{0.4} - 1 = 0.3$$

In practice, we will have only one actual value for each of a finite number of probabilistic forecasts and the estimated distributions will typically consist of a finite number of percentiles, e.g., P10, P50 and P90. **Fig. 2.39** shows a calibration plot we created using 30 probabilistic forecasts with truncated estimated distributions and one actual value—sampled from the true distribution—for each forecast. The estimated distributions were generated using CB = 0.4 and DB = 0.3. Calculating the slope and intercept from the least-squares best fit of the calibration curve in Fig. 2.39 and substituting the values into Eqs. 2.23 and 2.24, we get:

$$CB_{OC} = 1 - m = 1 - 0.5682 = 0.4318$$
$$DB_{OC} = \frac{2a}{CB_{OC}} - 1 = \frac{2 \times 0.283}{0.4318} - 1 = 0.31$$

These are good approximations of the input CB and DB parameter values. Of particular interest is the number of forecasts required to obtain reasonable estimates of the calculated CB and DB parameter values, or more generally, the relationship between number of forecasts and accuracy of the calculated parameter values. This should be investigated in future work.

Similar relationships between calibration-line properties and bias-parameter values can be derived for underconfidence as well, because underconfidence also produces straight-line calibration plots under the McVay and Dossary framework. Again, let *m* be the slope

of the calibration line and $a$ its intercept. For underconfident forecasts—slope greater than 1—the confidence parameter is:

$$CB_{UC} = \frac{1}{m} - 1 \quad (2.25)$$

and the directional bias parameter is:

$$DB_{UC} = 1 - \frac{2a}{1-m} \quad (2.26)$$

We created the calibration plot in **Fig. 2.40** using CB = -0.4 and DB = 0.3 assuming a very large set of forecasts. Fitting a trend line yields a slope of 1.6667 and intercept of -0.2333. Substituting these values into Eqs. 2.25 and 2.26, we back calculate the input confidence and directional bias values:

$$CB_{UC} = \frac{1}{m} - 1 = \frac{1}{1.6667} - 1 = -0.4$$
$$DB_{UC} = 1 - \frac{2a}{1-m} = 1 - \frac{2 \times -0.2333}{1 - 1.6667} = 0.3$$

**Fig. 2.41** shows a calibration plot we created using 30 truncated estimated forecast distributions and one actual value—sampled from the true distribution—for each forecast. The estimated distributions were generated using CB = -0.4 and DB = 0.3. Calculating the slope and intercept from the least-squares best fit of the calibration curve in Fig. 2.41 and substituting the values in Eqs. 2.25 and 2.26, we get:

$$CB_{UC} = \frac{1}{m} - 1 = \frac{1}{1.6154} - 1 = -0.3810$$
$$DB_{UC} = 1 - \frac{2a}{1-m} = 1 - \frac{2 \times -0.2205}{1 - 1.6154} = 0.2834$$

These are good approximations of the input CB and DB parameter values. Note that in this exercise we used P20, P40, and P60 instead of the usual P10, P50, and P90. This is because, in this underconfident case, the proportion correct for P10 is 0 and the proportion correct for P90 is 1, which are uninformative about the actual slope of the calibration curve

(see Fig. 2.40). If underconfidence was common (it is not), it would be a challenge to measure moderate to extreme underconfident assessments (CB < -0.1), since many assessors provide only P10, P50, and P90 values. Assessors should generate values at more cumulative probabilities (e.g., at P20, P30, P40, P60, P70, and P80) if they are worried that their assessments may cross over into the underconfidence region.

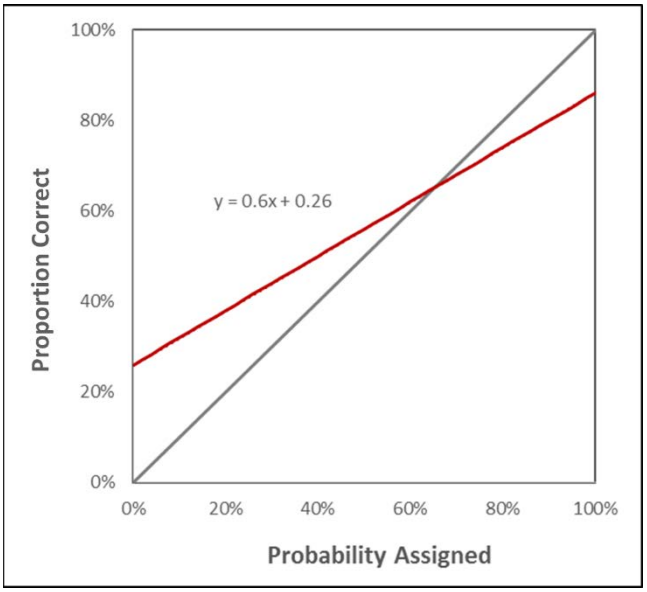


**Fig. 2.38—Calibration plot using McVay and Dossary (2014) framework and assuming known true and estimated distributions (created using CB = 0.4 and DB = 0.3).**

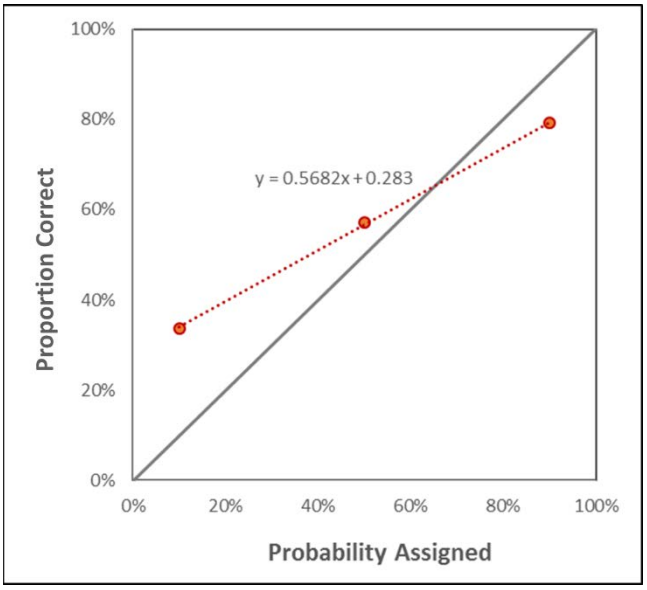


**Fig. 2.39—Calibration plot using McVay and Dossary (2014) framework and assuming 30 forecasts and one actual value for each forecast (created using CB = 0.4 and DB = 0.3).**

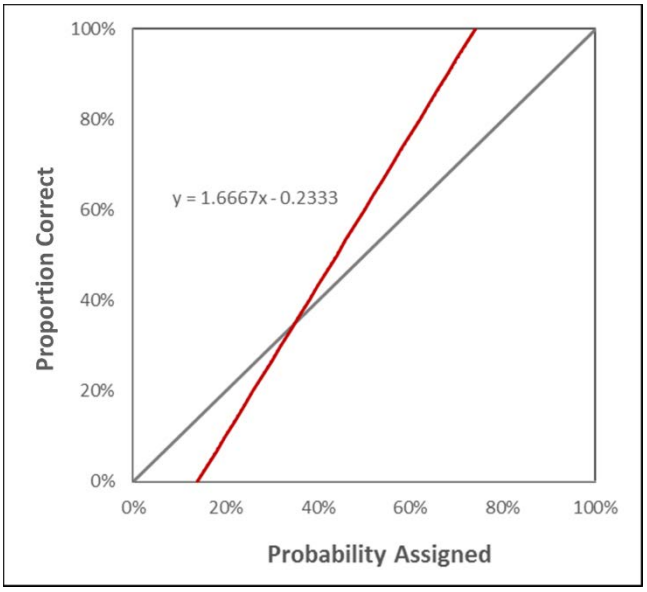


**Fig. 2.40—Calibration plot using McVay and Dossary (2014) framework and assuming known true and estimated distributions (created using CB = -0.4 and DB = 0.3).**

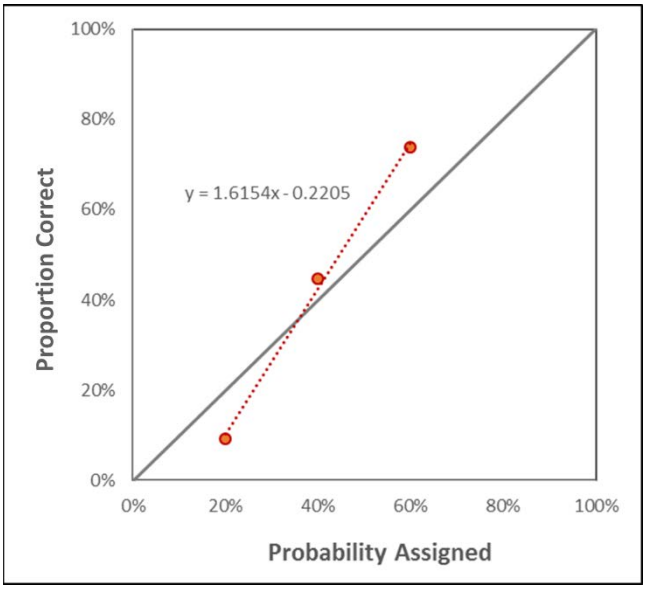


**Fig. 2.41—Calibration plot using McVay and Dossary (2014) framework and assuming 30 forecasts and one actual value for each forecast (created using CB = -0.4 and DB = 0.3).**

*Measuring CB and DB using the new general framework*

Under the new general framework, both estimated and true distributions can be full distributions, such as normal or lognormal. In this situation, calibration plots will generally result in curves rather than straight lines (**Fig. 2.42**), and calculating confidence and directional bias parameter values is more challenging.

In an effort to devise a calculation method, we first assumed that we know the true and the estimated distributions. If both distributions are either normal or lognormal, then the calibration curve in the extreme (very large number of forecasts), will produce an S-shaped curve similar to Fig. 2.42. We calculate a least-squares best fit of that curve. If the number of forecasts at each cumulative probability is not the same, then we should use a weighted least-square method where each point on the calibration plot is weighted by the number

of forecasts used to generate this point. Let *m* be the slope and *a* be the intercept of the best fit. The CB and DB parameter values can be approximated using Eqs. 2.23 and 2.24, since the calibration curve exhibits overconfidence.

For example, we created the calibration plot in Fig. 2.42 using CB = 0.4 and DB = 0.3. Fitting a least-squares best-fit line yields a slope of 0.5594 and intercept of 0.2856. Substituting these values into Eqs. 2.23 and 2.24 yields:

$$CB_{OC} = 1 - m = 1 - 0.5594 = 0.4406$$
$$DB_{OC} = \frac{2a}{CB_{OC}} - 1 = \frac{2 \times 0.2856}{0.4406} - 1 = 0.2964$$

which are good approximations for the input CB and DB values of 0.4 and 0.3.

Again, in practice we will have a finite number of assessments, with only one actual value for each assessment and a finite number of percentiles for the estimated distributions, e.g., P10, P50 and P90. We repeated the exercise assuming these conditions with 30 probabilistic forecasts, again using CB of 0.4 and DB of 0.3, which resulted in the calibration chart in **Fig. 2.43**. Finding a best-fit straight line and substituting the slope and intercept into Eqs. 2.23 and 2.24 yields:

$$CB_{OC} = 1 - m = 1 - 0.5645 = 0.4355$$
$$DB_{OC} = \frac{2a}{CB_{OC}} - 1 = \frac{2 \times 0.2876}{0.4355} - 1 = 0.32$$

Similarly, one can apply the same methodology to underconfident assessments, should this ever become a problem. We created the calibration plot in **Fig. 2.44** using CB = -0.4 and DB = 0.3. Fitting a trend line yields a slope of 1.6156 and intercept of -0.2092. Substituting these values into Eqs. 2.25 and 2.26, we back calculate the confidence and directional bias values:

$$CB_{UC} = \frac{1}{m} - 1 = \frac{1}{1.6156} - 1 = -0.3810$$
$$DB_{UC} = 1 - \frac{2a}{1-m} = 1 - \frac{2 \times -0.2092}{1 - 1.6156} = 0.3203$$

**Fig. 2.45** shows a calibration plot we created using 30 forecast distributions, with three percentiles, and one actual value—sampled from the true distribution—for each forecast. The estimated distributions were generated using a CB of -0.4 and DB of 0.3. Calculating the slope and intercept from the best fit of the calibration curve in Fig. 2.45 and substituting the values into Eqs. 2.25 and 2.26 yields:

$$CB_{UC} = \frac{1}{m} - 1 = \frac{1}{1.6071} - 1 = -0.3778$$
$$DB_{UC} = 1 - \frac{2a}{1-m} = 1 - \frac{2 \times -0.1964}{1 - 1.6071} = 0.3530$$

Note that in this exercise we used the proportion correct P20, P40, and P60 instead of the usual P10, P50, and P90 for the same reasons mentioned previously.

In all cases, we back calculated reasonable approximations to the input CB and DB parameter values. We do not believe it is critical to know the exact values of confidence and directional biases. Remember that one can measure these bias values only for groups of forecasts, not for individual forecasts. It is only necessary to know the directions (overconfidence versus underconfidence, optimism versus pessimism) and approximate magnitudes of the biases to make beneficial use. Lookbacks, calibration and quantification of biases is so seldom practiced that any reasonable approximation of these biases will be quite useful and valuable if the measurements are used to reduce or eliminate biases in new assessments. Capen (1976) made similar points about the value of approximate adjustment methods in his paper.

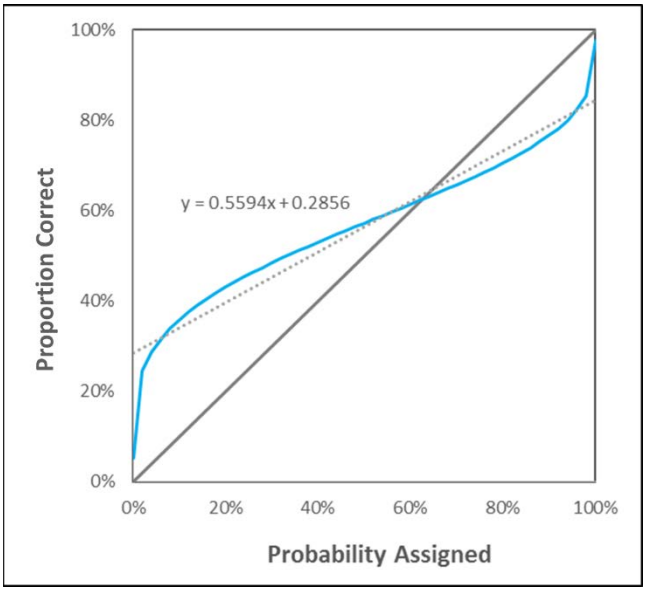


**Fig. 2.42—Calibration plot using the generalized framework and assuming known true and estimated distributions (created using CB = 0.4 and DB = 0.3).**

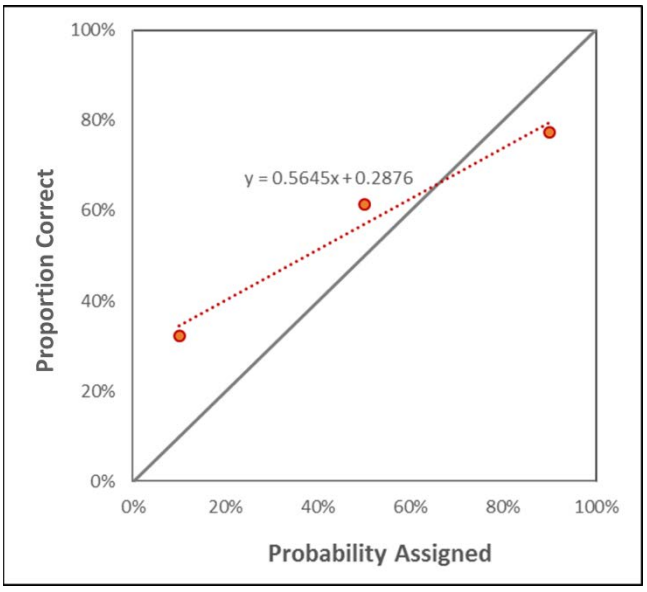


**Fig. 2.43—Calibration plot using the generalized framework and assuming 30 forecasts and one actual value for each forecast (created using CB = 0.4 and DB = 0.3).**

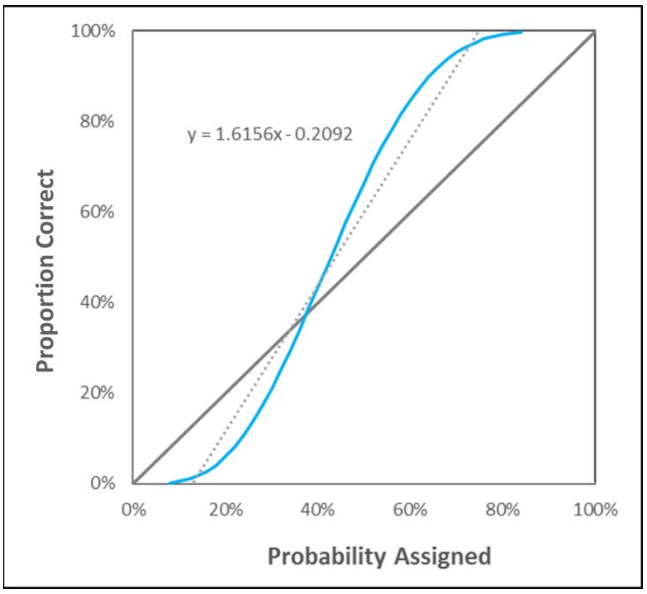


**Fig. 2.44—Calibration plot using the generalized framework and assuming known true and estimated distributions (created using CB = -0.4 and DB = 0.3).**

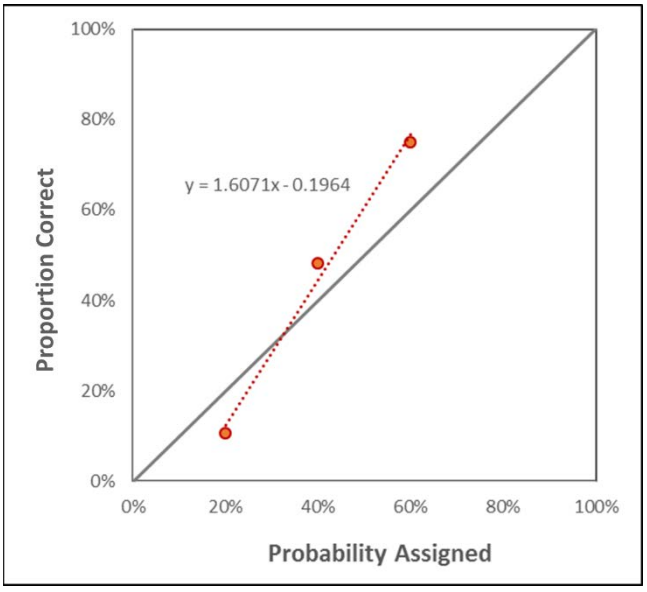


**Fig. 2.45—Calibration plot using the generalized framework and assuming 30 forecasts and one actual value for each forecast (created using CB = -0.4 and DB = 0.3).**

**Eliminating Biases and Improving Probabilistic Forecasts**

In the previous section, we showed that lookbacks and calibration plots can be used to detect and quantify the directions and magnitudes of biases. How does one make use of this information to eliminate biases? There are a variety of ways, including training versus monitoring processes, as well as internal versus external adjustment methods.

The first process is training, particularly of individuals involved in making probabilistic assessments. Training often takes place in a formal training setting, and usually involves making a series of probabilistic estimates of quantities uncertain to the estimator, but for which the actual values are known. The assessments do not necessarily have to be related to one's work or even the petroleum industry in general. Assessing uncertainty is a different skill from petroleum engineering, geology, astrophysics, or any other skill. Because it is a separate skill, it has to be learned. Because it is a separate skill, it can be learned using questions on any topic. For an example of an uncertainty assessment training quiz, see the 10-question quiz in Capen (1976). The advantage of these types of training questions is that immediate feedback can be provided. The trainee completes a series of assessments, the actual values are revealed, calibration plots are generated, and biases are identified and measured. Usually, the primary problem is overconfidence; ranges and distributions are too narrow because much of our uncertainty comes from options or outcomes that we do not consider (unknown unknowns). Another round of assessments is then made. Knowing the directions and degrees of biases, the trainee in some cases may learn on his or her own how to adjust assessments accordingly to eliminate the biases. Several rounds of training may be required to achieve perfect calibration. If not, it may be

necessary to introduce de-biasing techniques and other mental tricks to help trainees become well calibrated. Hubbard (2014) dedicated a whole chapter of his book to calibration training. Instead of a one-time training exercise, training can also be distributed over time. Fondren et al. (2013) reported on improvement in calibration of students resulting from training distributed over the course of a semester. One can also train oneself without a formal training program. Capen (1976) suggested setting up a personal program where one makes some predictions about the future every month, assigns probabilities to the predictions, and checks the results religiously. The process is the same—compare actual to forecasted, check the calibration, measure bias parameter values quantitatively if desired, and then make appropriate adjustments in new assessments. Note that training processes involve internal adjustment of probabilities. That is, with training, one is adjusting one's internal ability to assess uncertainty.

While training is beneficial and we recommend it to become well calibrated, we do not believe one-time training is sufficient. Old bias habits can return over time. We believe it is necessary to demonstrate that one remains well calibrated over time. This requires monitoring—a continual process of forecast tracking, lookbacks as actual values become available, checking calibration by comparing actual values to forecasts, quantifying bias directions and magnitudes, and then using this calibration information to adjust new probabilistic assessments. Fondren et al. (2013) presented a relational database system to facilitate tracking probabilistic assessments and checking calibration. If biases are detected, one can then use the calibration results to make adjustments in new forecasts and assessments. Ideally, the individuals or team involved in making the assessments will

internally self-adjust to eliminate the biases detected. This could involve adjusting parameter distributions input to models, or it could involve making adjustments to the probabilistic methodology that is being used, e.g., moving from a scenario-based method to a Monte Carlo method to better assess uncertainty.

If internal adjustment is insufficient or impractical, it may be appropriate to apply external adjustment of probabilistic assessments to eliminate biases. Capen (1976) demonstrated how to use calibration results to externally adjust forecasts. For example, knowing from lookbacks and calibration that forecast P10-P90 ranges were too narrow, i.e., actually P30-P70 ranges, he simply plotted the forecast values versus the calibrated P30-P70 probabilities on probability paper (normal or lognormal) and extended the ranges to revised P10-P90 values. Fondren et al. (2013) demonstrated how a forecast-tracking system was used to externally adjust shale-gas probabilistic production forecasts to improve their reliability. External adjustment of probabilistic forecasts can be made by the team making the forecasts, or by other parties. For example, if management has historical calibration evidence demonstrating that a particular asset team is biased, e.g., overconfident and optimistic, it can use this calibration information to externally adjust probabilistic forecasts from the asset team prior to making decisions. Eliminating biases and moving to consistent generation of well-calibrated probabilistic forecasts may require a combination of internal and external adjustment of forecasts. If forecast tracking, lookbacks, calibration, and adjustment are maintained as a continual process over time, it is expected that less external adjustment will be required with time as individuals and asset teams learn to adjust for their biases internally.

In a previous section, we showed that eliminating biases was more beneficial than reducing uncertainty for improving decision making and maximizing portfolio value. However, in our analysis we did not consider the relative costs of eliminating biases versus reducing uncertainty. There will likely be some cost associated with changing corporate culture, and possibly changing incentive structures, to include a focus on bias reduction and reliable uncertainty assessment. Once established, however, the cost is primarily that of bookkeeping—tracking forecasts and doing periodic lookbacks. We anticipate that the costs of bias reduction will be significantly less than the costs of uncertainty reduction—acquiring additional data and increasing modeling complexity. Furthermore, the costs of reducing uncertainty can increase exponentially as uncertainty is further reduced. Given the relative benefits and relative costs, we suggest that it is more important to focus on reducing biases first, then work on reducing uncertainty if needed.

**Conclusions**

A new generalized framework for quantifying the value of reliable uncertainty assessment (or quantifying the cost of biased estimation) that allows full, non-truncated estimated distributions replicates well the results and conclusions from a previously presented simplified framework that used truncated estimated distributions. Moderate overconfidence and optimism can easily produce average portfolio disappointment (estimated value minus realized value) of 30-35% of estimated portfolio EV or more.

Extension of the new generalized framework to underconfidence demonstrates that underconfidence, in combination with directional bias, is similarly detrimental to portfolio

performance as overconfidence. Thus, as operators seek to eliminate overconfidence bias, they should be wary of overcorrecting into underconfidence.

Gains from reducing uncertainty are small given moderate levels of confidence and directional biases. At higher levels of confidence and directional biases, reducing uncertainty will result in greater reduction in expected disappointment and increase in expected value attainment. However, these improvements are still less than those that result from reducing biases. The lowest expected disappointment and the highest expected value attainment can be achieved only by eliminating biases.

Biases in project estimation—overconfidence, underconfidence, optimism and pessimism—can be detected and quantified by conducting lookbacks (comparing actual performance to probabilistic forecasts) and constructing and analyzing calibration plots. Armed with quantitative measurements of biases, operators can then make efforts to eliminate these biases in new forecasts through a combination of internal adjustment of uncertainty assessments—via assessment training and/or monitoring—and external adjustment of forecasts using measurements of biases from calibration.

# CHAPTER III

# MEASURING AND IMPROVING THE RELIABILITY OF PROBABILISTIC ASSESSMENTS IN PETROLEUM ENGINEERING

## Overview

Previous work on the impact of biases on portfolio optimization showed that decision error will be minimized and portfolio value will be maximized only when there are no biases in project estimation. If operators track probabilistic forecasts and perform lookbacks, biases can be measured from calibration charts. Operators can then use these bias measurements to mitigate and eliminate biases in new estimates. In this work, we aim to determine the relationship between the number of probabilistic assessments and the accuracy of bias measurements, and to determine guidelines for minimizing biases in new assessments using the external adjustment.

We generated calibration curves for historical probabilistic assessments and used these curves to calculate different reliability measures such as the coverage rate, the calibration score, and confidence and directional bias parameters. We used a generalized biases framework presented in previous work to generate different numbers of biased assessments and then determined the relationship between the number of assessments and the accuracy of the bias measurements. Furthermore, we used the calibration curve to adjust new forecasts, and we measured the reliability of the new forecasts after adjustment as a function of the number of historical assessments and other parameters.

Our research indicates that, in general, using more historical assessments to measure biases improves the accuracy of the bias measurements. However, even a low number of assessments (as low as 10) is enough to detect moderate biases. An even lower number of assessments (2 or 3) is enough to detect extreme biases. Furthermore, our research shows that production forecasts that were updated frequently over time using newly available data and externally adjusted using the most recent bias measurements were superior in terms of calibration to forecasts that were not updated or externally adjusted.

The methods presented in this paper can be used to measure and improve the reliability of probabilistic assessments in many petroleum engineering applications. Implementing these methods in a continual process of tracking assessments, looking backs as actual values become available, checking calibration by comparing actual values to forecasts, quantifying biases, and using these bias measurements to improve new assessments will result, over the long run, in the best calibrated assessments. Well calibrated assessments result in a better identification of superior projects and inferior projects, and ultimately, better investment decision making and increased profitability.

## Introduction

A number of authors (Capen 1976; Brashear et al. 2001; Rose 2004) noted the consistent underperformance of the petroleum industry. Much of this underperformance is attributed to cognitive biases that result in uncertainty ranges that are usually too narrow and too optimistic.

McVay and Dossary (2014) proposed a framework to quantify these biases and estimate the value of assessing uncertainty. Their model estimated the monetary impact of

overconfidence, i.e., underestimation of uncertainty, where the estimated distribution of an uncertain quantity is too narrow, and directional bias (DB), where the estimated distribution is shifted in the optimistic or pessimistic direction. Mathematically, they started with the true distribution (the reliable distribution given the assessor's state of information) and truncated it at the tails to model the estimated distribution. The magnitude of and location of the truncation is determined by the overconfidence and directional bias values. The overconfidence parameter ranges from 0.0 to 1.0 and specifies the fraction of the true distribution not sampled by the estimated distribution. Therefore, a value of 0.0 denotes that the entire true distribution was sampled. On the other hand, an overconfidence value greater than 0.0 denotes that only a subset of the true distribution was sampled. The DB parameter ranges from -1.0 to 1.0 and it specifies the location of the estimated distribution relative to the true distribution. A DB value of -1.0 means that only the lowest possible outcomes of the true distribution were considered. On the other hand, a DB value of +1.0 means that only the highest possible outcomes of the true distribution were considered (**Fig. 3.1**).

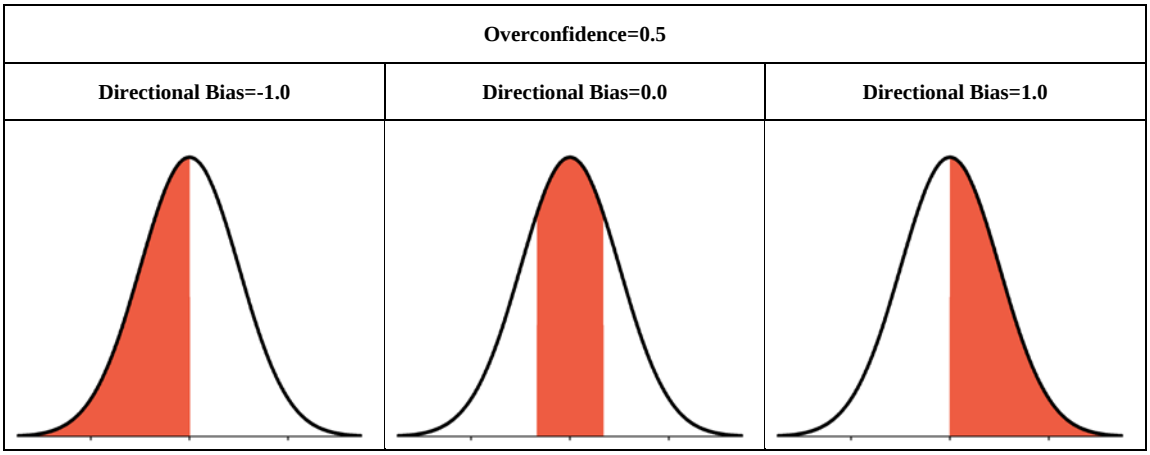


**Fig. 3.1—In McVay and Dossary (2014) framework, the overconfidence parameter specifies the fraction of the true distribution (black curve) not sampled by the estimated distribution (red area) and the directional bias specifies the location of the estimated distribution relative to the true distribution [from Alarfaj and McVay (2018)].**

Alarfaj and McVay (2018) generalized the framework to allow the use of non-truncated estimated distributions and included the effects of underconfidence, i.e., overestimation of uncertainty, where the estimated distribution of an uncertain quantity is too wide. While underconfidence is not currently common, as the industry hopefully improves in uncertainty estimation, it is possible to overcorrect overconfidence into underconfidence. The underconfidence parameter ranges from -1.0 for complete underconfidence (no information about the uncertain quantity) to 0.0 for no underconfidence, which is the same as 0.0 overconfidence. The overconfidence and underconfidence parameters were combined into the confidence bias (CB) parameter that ranges from -1.0 to 1.0, where the negative values indicate underconfidence and positive values indicate overconfidence. Furthermore, Alarfaj and McVay (2018) distinguished between directional bias and optimism-pessimism bias (OPB) by clarifying that a directional bias in the positive direction could mean optimism or pessimism depending on the evaluated parameter. For example, for a value-based parameter such as Net Present Value (NPV), a positive DB

value is considered optimism because expecting a greater value than reality is of benefit to the estimator. On the other hand, for a cost-based parameter such as Capital Expenditure (CapEx), a negative DB is considered optimism because expecting lower cost than reality is of benefit to the estimator.

**Fig. 3.2** shows a summary of the relationship between the expected value (EV) and the standard deviation (SD) of the estimated distribution and the true distribution as a function of confidence and directional bias parameters for a true standard-normal distribution (EV of 0 and SD of 1) using the generalized framework.

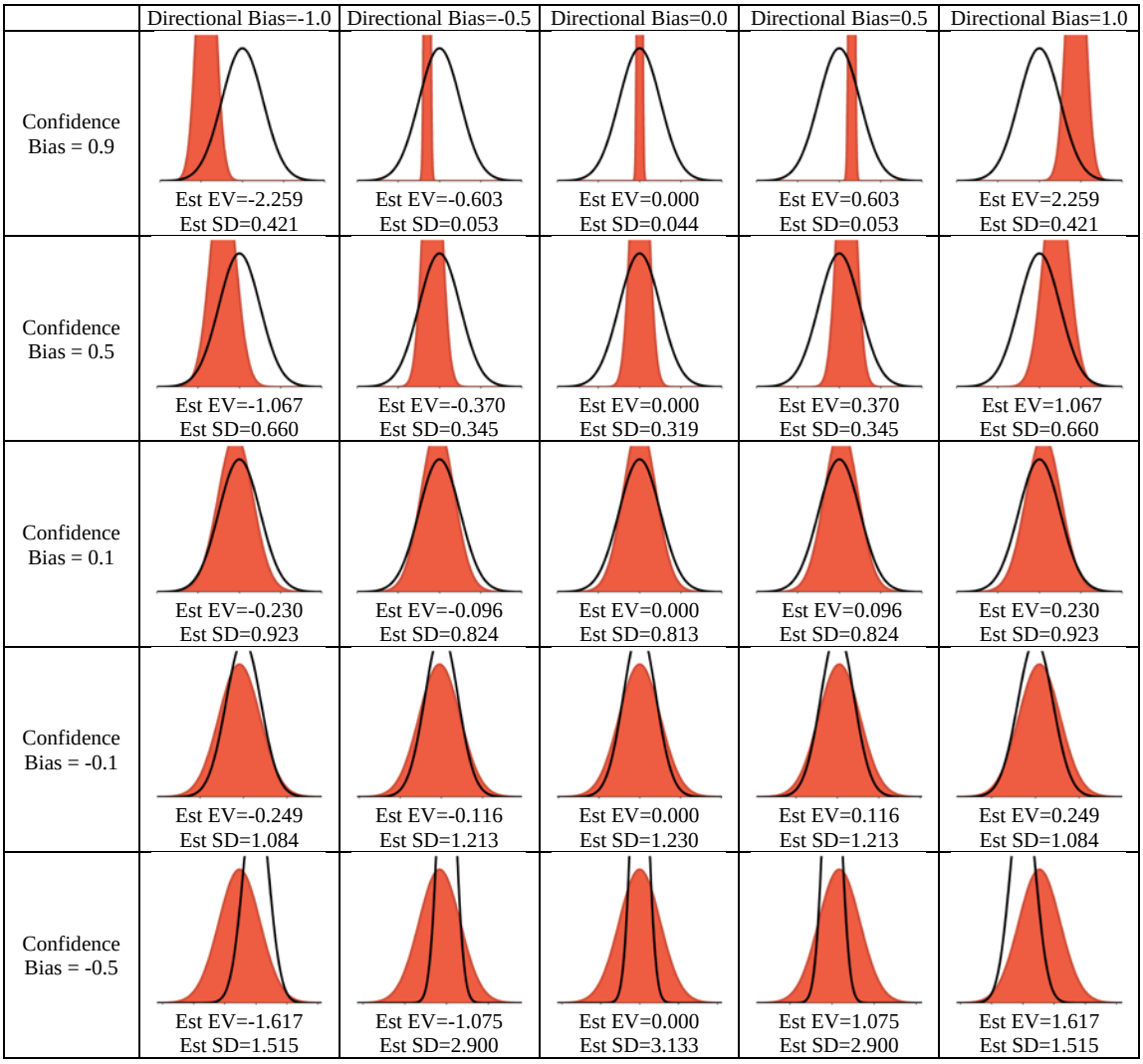


**Fig. 3.2—Relationship between the estimated distribution (red) and the true distribution (black curve) as a function of confidence and directional bias parameters using the generalized framework.**

*Reliability of Probabilistic Assessments Can Be Measured*

An assessor's degree of belief in a proposition is quantified by the probability of it being true (Lichtenstein et al. 1977). In our work, an assessor is a person who is issuing those propositions but could also mean a probabilistic model, or a probabilistic assessment methodology, used to issue these propositions. A forecast is a proposition for a quantity

to be known in the future. An assessment is a group of one or more propositions that are related to the same quantity or event. For example, if an assessor is trying to assess the cost of a gas well and issues forecasts at P10, P50, and P90, then each forecast is a proposition, and all three are part of one assessment.

Probabilistic assessments can be discrete or continuous. Discrete assessments can take only specific values and can be categorical (such as dry hole, oil, or gas) or numeric (such as the number of wells required). The number of possible outcomes may be finite or infinite but the distinguishing feature is that there are no possible outcomes in between. For example, you cannot have seven and a half wells. Probabilistic assessments can also be continuous, i.e., they are not restricted to specific values but can take any value over a continuous range. Continuous assessments are always numeric. Most assessed quantities in the petroleum industry are continuous, including but not limited to porosity, permeability, reserves, oil price, and capital expenditure. Ultimately, all continuous assessments can be represented by a number of binary outcome (true or false) propositions, as we will show in the next section. In this paper, we are focusing on measuring and improving the reliability of continuous assessments.

Suppose that an assessor proposed that the probability of a project's cost being less than or equal to $10 million is 90% (or P90). Suppose that the project's cost turned out to be $15 million. This in itself will tell us little, if anything, about the validity and the reliability of the assessment because the assessor did assign a 10% chance of the project cost exceeding $10 million. In most cases, we cannot evaluate the reliability of such an assessment in isolation (except if the assessor proposed that there is a 0% or a 100%

chance of an event happening). If, however, the assessor assigned a 90% probability that the cost will be less than or equal a certain value in 200 independent projects, and only 7 of those propositions turned out to be true, there must be something wrong with this assessor's assessments. We can conclude then, that these probabilistic assessments were biased and unreliable.

There are a number of methods to measure the reliability of probabilistic assessments. Capen (1976), Gonzalez et al. (2012; 2013), Fondren et al. (2013), and Gong et al. (2014) used the empirical coverage of the central prediction interval—i.e., the coverage rate (CR)—to measure the reliability of probabilistic assessments. The CR indicates the percentage of outcomes falling within a specified prediction interval. For example, if we specify an 80% central prediction interval, then CR will be the percentage of outcomes falling between the P10 and P90 values. A CR value (e.g., 80%) equivalent to the width of the central prediction interval (e.g., 80%) is a necessary but not sufficient condition for a reliable probabilistic assessment. It indicates only that the width of the prediction interval is reliable but it does not tell us anything about the reliability of specific propositions, or percentiles (e.g., P50 or P95).

If for all propositions (e.g., actual ≤ P10, actual ≤ P50), the relative frequency of true propositions is equal to the probability assigned, we can say that the assessor is perfectly calibrated (Lichtenstein and Fischhoff 1977) and his or her assessments are reliable. The degree of calibration can be evaluated by comparing the proportion of true propositions (proportion correct) to the probability assigned. A graph (**Fig. 3.3**) that shows the proportion correct for each probability assigned is called a calibration curve or a

calibration plot (Lichtenstein and Fischhoff 1977). Gonzalez et al. (2012; 2013) used the calibration plot to measure the reliability of the Markov-Chain-Monte-Carlo (MCMC) method using different decline-curve-analysis (DCA) models on 197 Barnett-shale gas wells. Fondren et al. (2013) proposed an assessment-tracking database system and used calibration plots to measure the reliability of probabilistic assessments. Alarfaj and McVay (2018) showed that CB can be estimated from the slope of the least-squares best-fit line of the calibration curve and DB can be estimated from the slope and the intercept of the same line.

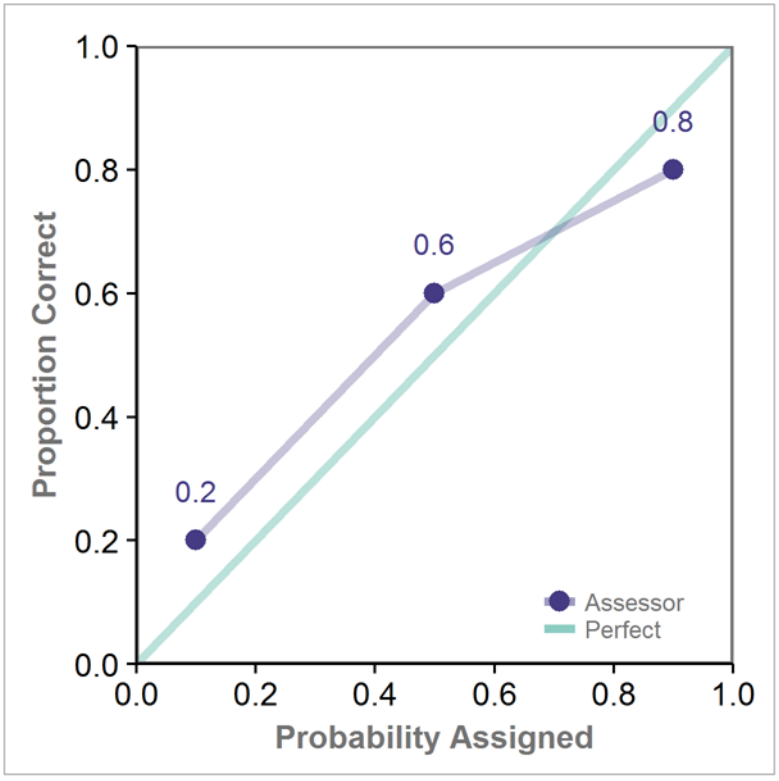


**Fig. 3.3—A calibration curve shows the proportion of correct propositions at each assigned probability.**

Calibration plots provide valuable insights into the reliability of the assessor or the probabilistic assessment method. However, when comparing the reliability of a large number of assessors, or assessor groups, it is often more practical to use a single value rather than the calibration plot. Fondren et al. (2013) used the calibration score (CS), a component of the Brier Score commonly used in weather forecasting, strategic intelligence, and behavioral sciences (Murphy 1973).

All of these measures of reliability, or of biases, require a sufficiently large number of assessments to be evaluated with a high degree of confidence. While grouping assessments is necessary for measuring the reliability of probabilistic assessments, grouping can obscure individual differences in biases among members of the group (Lichtenstein et al. 1977). For example, if you evaluate probabilistic assessments issued by a group of some reliable and some biased assessors, the reliability measure (CR, CS, CB, or DB) will be a composite of the reliability measures of both types of assessors. The evaluator will not be able to distinguish between the reliable and biased assessors. If differences in reliability of group members are suspected, reliability may be assessed by subgroups. However, this may be constrained by the number of assessments available. There will be a trade-off between resolution of the subgrouping and number of assessments available per subgroup, which impacts the accuracy of the reliability measurements as we will discuss in a later section. There is very little, if anything, in the literature that addresses the accuracy of these measures of reliability, or of biases, as a function of the number of assessments available.

*New Assessments Can Be Improved*

Assessors can and should use the results of reliability measurements as feedback to internally adjust their assessments. Often, however, internal adjustments may not be sufficient. Capen (1976) observed that people still gave narrow ranges (albeit slightly improved) even after they were warned that assessments are usually too narrow. In this case, it is beneficial to use reliability measurements results to externally adjust assessments. Capen (1976) demonstrated how to use the CR value to externally adjust forecasts. For example, knowing from look-backs and calibration that forecast P10-P90 ranges were too narrow—CR=0.40 rather than 0.80—and were actually P30-P70 ranges, he simply plotted the forecast values at the P30-P70 probabilities on probability paper (normal or lognormal) and extended the ranges to revised P10-P90 values (**Fig. 3.4**). Capen (1976) did not explicitly state that he assumed that the coverage rate was centrally located, but this assumption can be inferred from the examples he provided. This methodology can improve the coverage rate, but the assumption that the coverage rate is centrally located ignores directional bias and, thus, makes this method (with the centrality assumption) unsuitable for mitigating directional biases.

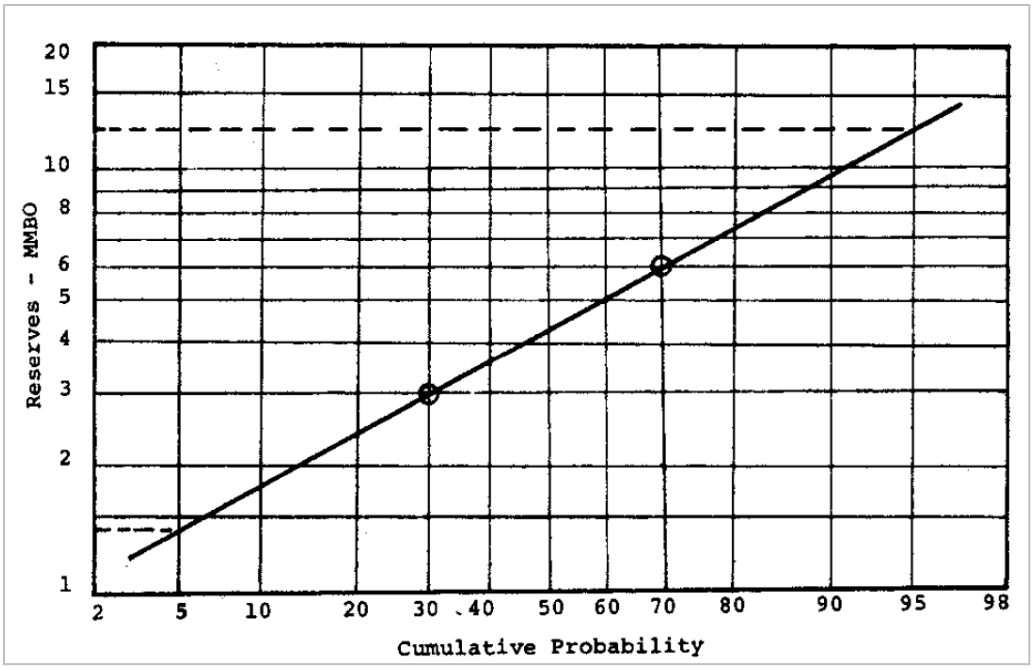


**Fig. 3.4—External adjustment using lognormal probability paper [from Capen (1976)].**

Fondren et al. (2013) used their assessment tracking system to externally adjust shale-gas probabilistic production forecasts to improve their reliability. They implemented a methodology similar to the one suggested by Capen (1976), but which uses calibration curves instead of centrally-located coverage rates to externally adjust these forecasts. However, their work was limited to estimated distributions defined only at P10, P50, and P90, and did not consider calibration curves for fully defined continuous distributions. They used only lognormal distributions to externally adjust new forecasts, which may limit the flexibility of the adjustment process.

In this paper, we will use the calibration plot to measure the reliability of probabilistic assessments and we will show that CR, CS, CB, and DB can be measured or estimated from the calibration plot. Then we will discuss the level of confidence in those reliability

measures and how many assessments are needed to get a specific confidence level for these measures. Next, we will show how to implement external adjustment using calibration curves in a continual process of assessment tracking, look-backs, calibration, and adjustments to improve probabilistic forecasts and reduce biases. Then, we will discuss the accuracy and the limitations of this external adjustment method.

## Measuring the Reliability of Probabilistic Assessments

### *Calibration Plots Are Used to Measure the Reliability of Probabilistic Assessments*

Continuous probabilistic assessments are usually defined in terms of a cumulative distribution function $F$, where the cumulative probability $P$ assigned to a specific outcome value $x_P$ is:

$$P = F(x_P) \quad (3.1)$$

or alternatively $x_P$ is the value which there is a $P$ chance that the observed outcome will be less than or equal to $x_P$,

$$x_P = F^{-1}(P) \quad (3.2)$$

where $F^{-1}$ is the inverse of the cumulative distribution function or the quantile function.

In many instances, $F$ is not fully defined across the probability range but rather is defined at specific cumulative probability values, such as 0.1, 0.5, and 0.9, and the values assigned to them are usually called P10, P50, and P90, respectively. These are called percentiles (or more generally quantiles) where the number corresponds to $P \times 100$. In other cases, such as reserves estimates, these values are presented using the de-cumulative probability notation (where P10 is the high number and P90 is the low number); in this

case the number corresponds to $(1 - P) \times 100$ and P90 means that there is a 90% chance that the actual value will be greater than or equal to the assigned value.

A calibration plot can be constructed by plotting the proportion of true propositions (proportion correct) at each cumulative probability value $P$ versus that probability. If the probabilistic assessments were defined at specific percentiles (e.g., P10, P50, and P90), we evaluate all propositions that have the same assigned cumulative probability together. If on the other hand, the probabilistic assessments were fully defined, then it is more practical to bin the probabilities into a limited number of cumulative probability subintervals and evaluate all propositions that are assigned to each subinterval together. Then we calculate the cumulative probability assigned to that subinterval as the average cumulative probability assigned to all propositions inside that cumulative probability subinterval. The proportion correct ($c_t$) of the $t$'th cumulative probability or subinterval is:

$$c_t(P_t) = \frac{1}{n_{P_t}} \sum_{i=1}^{n_{P_t}} I\,(x_i) \quad \text{.....(3.3)}$$

where $P_t$ is the $t$'th cumulative probability or the average cumulative probability of the $t$'th subinterval, $n_{P_t}$ is the number of propositions defined at the $t$'th cumulative probability or subinterval, $x_i$ is the value of the observed outcome, and the indicator function $I$ is a binary function defined as follows:

$$I(x) \coloneqq \begin{cases} 1, & x \leq F^{-1}(P) \\ 0, & x > F^{-1}(P) \end{cases} \quad \text{.....(3.4)}$$

where $x$ is the value of the observed outcome of the quantity assessed, and $P$ is the cumulative probability assigned to the proposition.

We recommend using points in graphing calibration plots of continuous assessments that are defined by specific percentiles (**Fig. 3.5**) and lines for continuous assessments that are completely defined (**Fig. 3.6**).

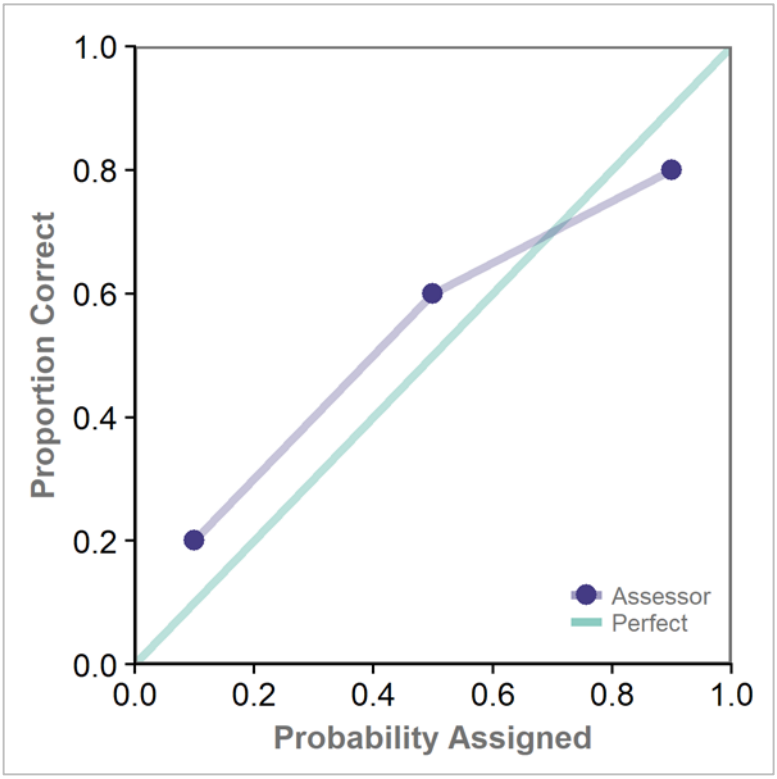


**Fig. 3.5—Calibration plot for continuous assessments defined by the 10th, 50th, and the 90th percentiles.**

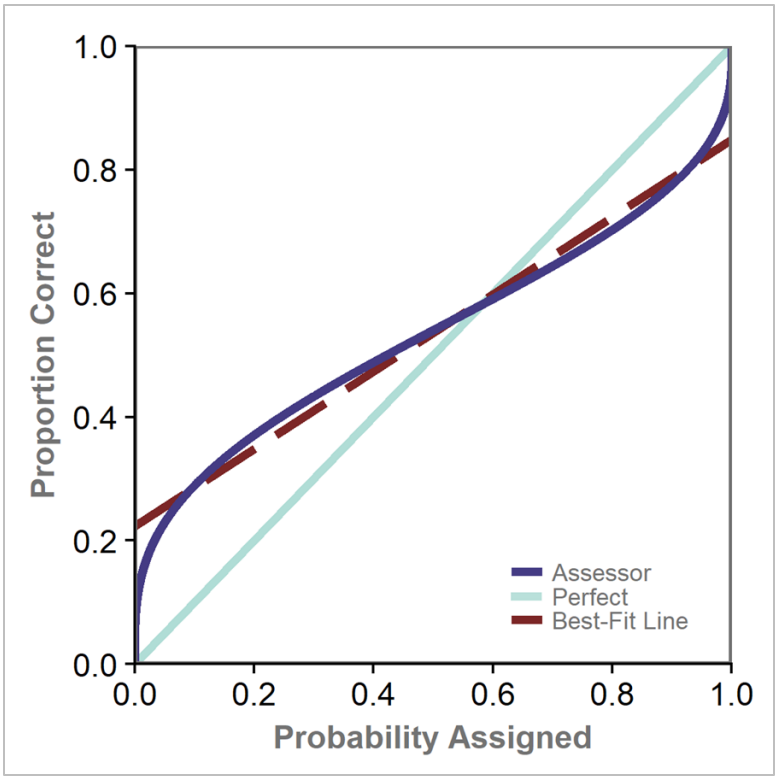


**Fig. 3.6—Calibration plot for continuous assessments that are completely defined over the probability range.**

*Confidence and Directional Biases Can Be Measured from Calibration Plots*

A group of perfect, or completely reliable, probabilistic assessments will fall on the unit-slope line on a calibration plot. That is, propositions that have been assigned a 10% probability will occur 10% of the time, propositions that have been assigned a 50% probability will occur 50% of the time, and those that were assigned 90% will occur 90% of the time. A least-squares best-fit line of the calibration curve with a slope less than 1 indicates that the probabilistic assessments are on average overconfident and have narrower ranges than they should. On the other hand, a slope greater than 1 indicates underconfident probabilistic assessments on average with ranges that are wider than they

should be (**Fig. 3.7**). Probabilistic assessments with positive directional bias will shift the curve upward, while probabilistic assessments with negative directional bias will shift it downward (**Fig. 3.8**).

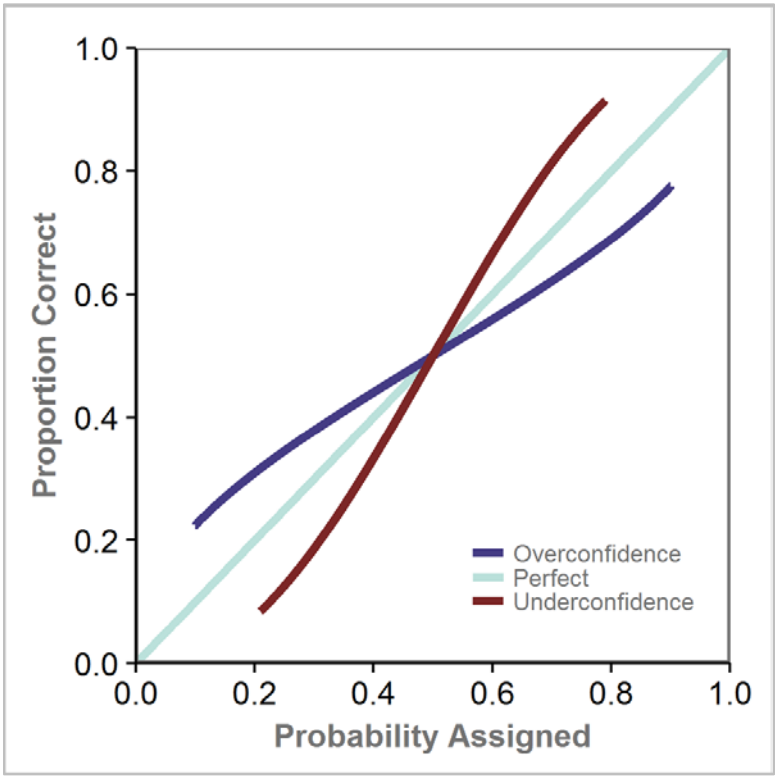


**Fig. 3.7—The calibration curve of reliable probabilistic assessments will fall on the unit-slope line, overconfident assessments will have an average slope less than 1, and underconfident assessments will have an average slope greater than 1 [modified from Gonzalez et al. (2012)].**

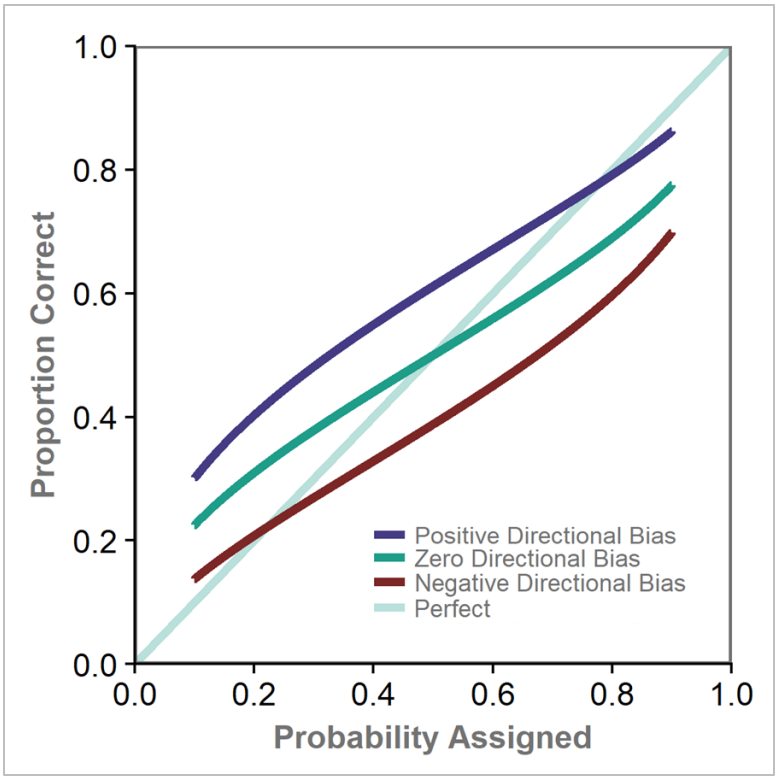


**Fig. 3.8—The calibration curve of reliable probabilistic assessments will fall on the unit-slope line, positively biased assessments will shift it upward, and negatively biased assessments will shift it downward.**

Alarfaj and McVay (2018) demonstrated that the magnitudes of the confidence and directional biases (CB and DB) can be estimated graphically from calibration plots. They noted that the calibration curve is a straight line when assessments consist of truncated estimated distributions. The confidence bias can be calculated from the slope *m* of that line. If the slope is less than 1, then the assessor is overconfident. In this case the confidence and directional bias parameters can be calculated from the slope *m* of calibration curve and its intercept *a* at $P = 0$ as follows:

$$CB_{OC} = 1 - m \quad \text{..... (3.5)}$$

$$DB_{OC} = \frac{2a}{1-m} - 1 \quad (3.6)$$

If on the other hand the slope is greater than 1, then the assessor is underconfident and the confidence and directional bias parameters can be calculated using:

$$CB_{UC} = \frac{1}{m} - 1 \quad (3.7)$$

$$DB_{UC} = 1 - \frac{2a}{1-m} \quad (3.8)$$

They also found that these equations can also reasonably estimate CB and DB using their generalized framework. The estimates can be made using the slope and the intercept of the least-squares best-fit line of the calibration curve. If the number of observations at each cumulative probability is not the same, then a weighted least-square method should be used where each point on the calibration curve is weighted by the number of observations used to generate this point. Because this is a best-fit line of a curve (Fig. 3.6), there will be some information loss. Consequently, these simplified equations will not calculate the exact biases. However, the differences between the measured and the actual biases are not significant. In fact, more variation in the bias measurement will be caused by a low number of assessments than by the information loss. Moreover, because of the information loss, the directional bias equation may result in values that are more than +1 or less than -1. In these cases, we use the limit (+1 or -1) instead of the calculated value.

*The Coverage Rate Indicates the Existence and the Severity of Over or Underconfidence*

A simple measure that can indicate the existence and severity of bias is the empirical coverage of the central prediction interval or the coverage rate (CR). The CR for an 80%

central prediction interval can be calculated directly from the calibration plot by taking the difference between the proportion correct values at $P$=0.9 and $P$=0.1.

We used the generalized framework to generate biased estimated distributions defined at P10, P50, and P90 assuming various CB and DB values. Then we calculated the theoretical (asymptotic) CR for each pair of CB and DB. **Fig. 3.9** shows the relationship between confidence and directional biases and the 80% prediction interval coverage rate for overconfident estimates. The higher the confidence bias, the smaller is the coverage rate. This is expected because overconfidence results in a narrower distribution than it should be. We also note that the directional bias has no significant effect on the coverage rate except at extreme DB values (near -1 or +1).

Finally, **Fig. 3.10** shows that low to moderate underconfidence (negative CB) will have coverage rates higher in values than their corresponding prediction interval. However, probabilistic assessments that have extreme DB (near -1 or +1) and extreme underconfidence (< -0.5) will have coverage rates that are smaller (sometimes much smaller) than their corresponding central prediction interval. This is because most observed outcomes are less than the assigned P10 (in the case of positive directional bias), or greater than the assigned P90 (in the case of negative directional bias).

What do these values mean for the analyst? In general, CR values lower than the assumed prediction interval indicate overconfidence, while CR values higher than the assumed prediction interval indicate underconfidence. Extreme underconfidence coupled with extreme DB is very rare and it is unlikely in practice that a low CR value would indicate underconfidence.

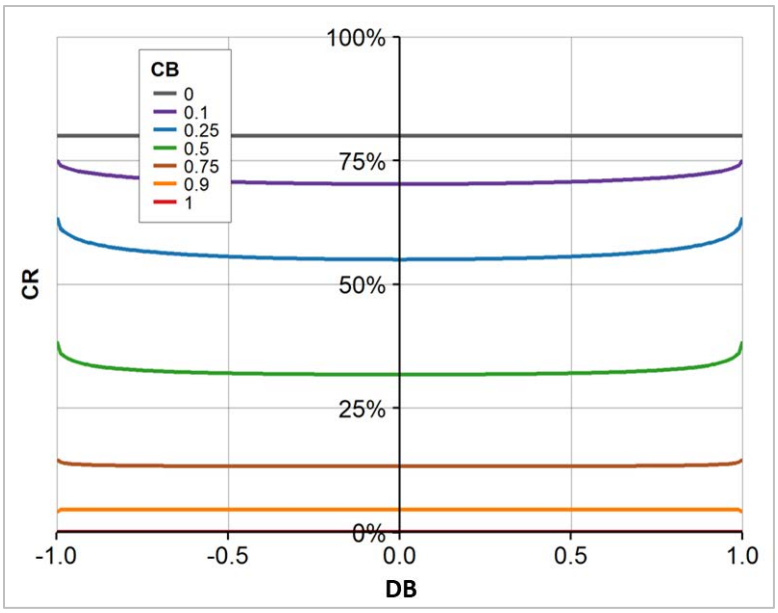


**Fig. 3.9—The relationship between confidence and directional biases and the coverage rate of the 80% central prediction interval assuming overconfidence.**

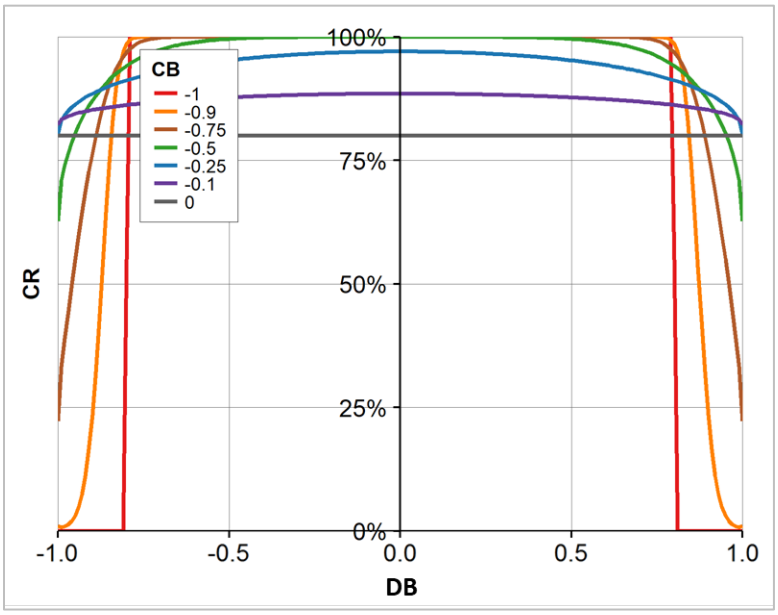


**Fig. 3.10—The relationship between confidence and directional biases and the coverage rate of the 80% central prediction interval assuming underconfidence.**

*Lower Calibration Scores Indicate Lower Biases Overall*

The Brier score was originally developed to assess the reliability of weather forecasts (Brier 1950). It was developed to assess probabilistic assessments tied to discrete, binary outcomes (e.g., rain or no rain). Murphy (1973) decomposed the Brier score into three components: calibration, knowledge, and resolution. The calibration component measures the weighted average of the mean-square difference between the assigned probability and the proportion of correct responses at each probability value or subinterval. The knowledge component is an inverse measure of the event predictability, and it is a property of the evaluated event and not the assessor. The resolution component describes how well an assessor can assign different probability values to different assessments. The knowledge and resolution components are not as significant for continuous assessments as for discrete assessment (Fondren et al. 2013); thus, we will focus on only the calibration component. The calibration score can be calculated using the following equation [adapted from Lichtenstein and Fischhoff (1977)].

$$CS = \frac{1}{N}\sum_{t=1}^{T} n_t (P_t - c_t)^2 \quad \text{..... (3.9)}$$

where $N$ is the total number of propositions, $n_t$ is the number of propositions in the $t$'th cumulative probability or probability subinterval, and $T$ is the number of defined cumulative probabilities or probability subintervals. $P_t$ and $c_t$ were defined previously in the calibration plots subsection. For continuous assessments, the calibration score ranges between 0 and 1/3, where 0 denotes a perfectly calibrated assessor who assigns probabilities corresponding to the actual frequencies of outcomes. That means the lower

the CS, the better is the assessor. Reduction in confidence and directional biases results in reduction of the calibration score.

The calibration score can also be directly calculated from calibration plots if the number of propositions in all assessments is the same. We read the $c_t$ values form the calibration curve at each corresponding assigned cumulative probability $P_t$ and substitute the values into:

$$CS = \frac{1}{T}\sum_{t=1}^{T}(P_t - c_t)^2 \quad (3.10)$$

The calibration score measures the sum of the squared differences between the calibration curve and the unit slope line. Therefore, a calibration score of 0 indicates a calibration curve equal to the unit slope line (a perfect assessor).

Similar to the previous subsection, we used the generalized framework to generate biased estimated distributions defined at P10, P50, and P90 assuming various CB and DB values. Then we calculated the theoretical CS value for each pair of CB and DB. **Fig. 3.11** shows the relationship between confidence and directional biases and the calibration score for overconfident estimates. Any deviation in confidence or directional bias from ideal (0 CB and 0 DB) will cause the calibration score to increase. The effects of directional bias on the calibration score increases as the confidence bias increase. It is noted that the calibration score does not distinguish between positive or negative directional biases; that is, positive and negative directional biases with the same magnitudes will affect the calibration score similarly.

**Fig. 3.12** shows the relationship between confidence and directional biases and the calibration score for underconfident estimates. The shape of the CS values in this plot is

an artifact of evaluating estimated distributions defined only at P10, P50, and P90. For many underconfident distributions, the proportion correct at P10 is 0 and the proportion correct at P90 is 1 regardless of the underconfidence level, which make them uninformative to the calibration score calculation. To overcome this issue, we calculated the CS value for each pair of CB and DB using fully defined estimated distributions. **Fig. 3.13** shows CS curves of the underconfident distributions are very similar to those of overconfident distributions (Fig. 3.12).

In summary, the calibration score does not discriminate underconfidence from overconfidence nor positive from negative directional biases. Similar remarks about the calibration score and over/underconfidence were made by Lichtenstein et al. (1977). This is a disadvantage of using traditional reliability measures such as CS or CR as compared to using CB and DB. Using CS or CR does not provide guidance to assessors on how to internally adjust their assessment models. On the other hand, CB will indicate whether the assessments should be wider or narrower and DB will indicate whether the assessments should be shifted positively or negatively. Ultimately, while it is not possible to distinguish optimism from pessimism or underconfidence from overconfidence using the CS, the lower the CS, the less biased the assessor overall. Furthermore, actions that reduce the CS will also reduce the overall confidence and directional biases.

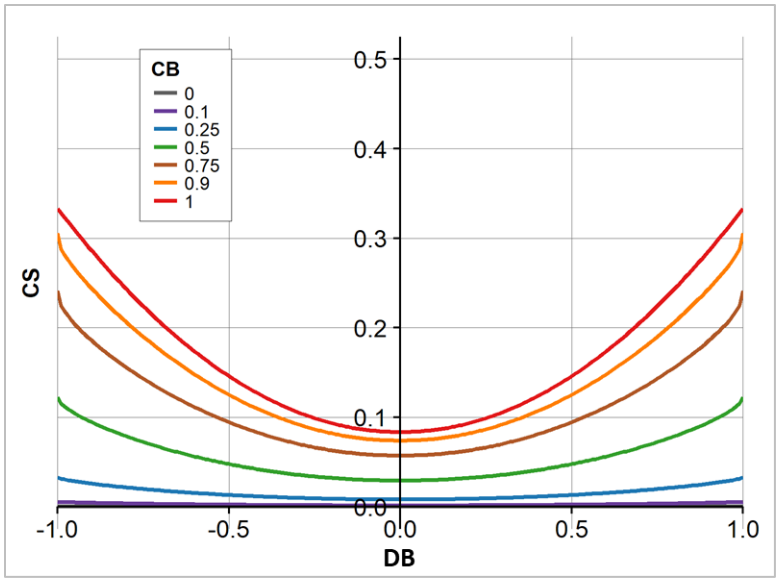


**Fig. 3.11—The relationship between confidence and directional biases and the calibration score assuming overconfident distributions using the generalized framework and estimated distributions defined at P10, P50, and P90 only.**

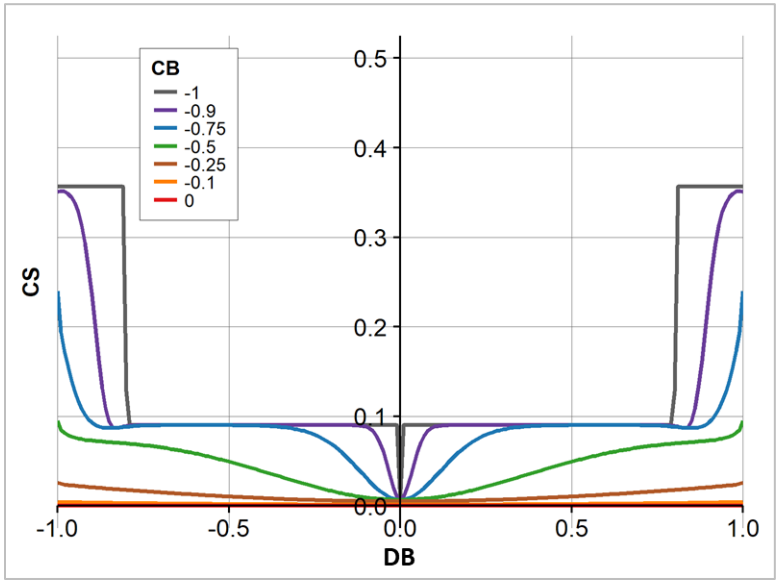


**Fig. 3.12—The relationship between confidence and directional biases and the calibration score assuming underconfident distributions using the generalized framework and estimated distributions defined at P10, P50, and P90 only.**

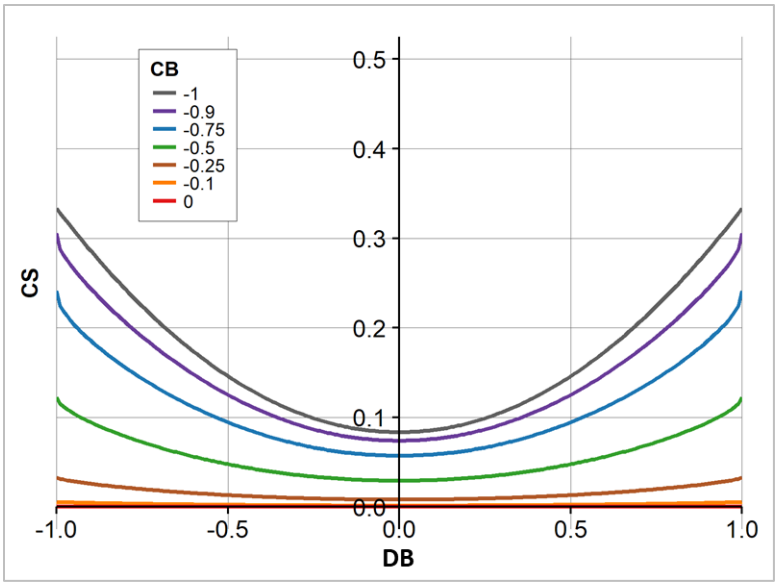


**Fig. 3.13—The relationship between confidence and directional biases and the calibration score assuming underconfident distributions using the generalized framework and fully defined estimated distributions.**

### More Probabilistic Assessments Lead to a More Accurate Bias Measurement

In the previous section, we calculated CB, DB, CR, and CS assuming that we knew both the true and estimated distributions. In reality, however, we know only one actual value for each of a finite number of probabilistic assessments. We can only estimate those reliability measures using pairs of probabilistic assessments and their corresponding actual observations. Therefore, it is of particular interest to learn how many probabilistic assessments are required to obtain reasonable estimates of the calculated reliability measurements. In other words, we are interested in the relationship between number of probabilistic assessments and level of confidence in the calculated parameter values such as CB and DB.

We first generated a number (varied between 2 and 1000) of truncated estimated distributions defined at P10, P50, and P90 assuming moderate confidence and directional biases (CB = 0.5 and DB = 0.5) and extreme confidence and directional biases (CB = 0.9 and DB = 0.9). Next, we sampled one value from each corresponding true distribution and measured the reliability of the estimated distributions (via CS, CB and DB measurements) as explained in the previous section. We repeated these steps 5000 times in a Monte Carlo simulation to determine the confidence intervals for these measurements. We plotted the expected value, the 80% confidence interval as the shaded area for the bias measurement, and the 99% confidence interval as the error bar (**Fig. 3.14**).

Fig. 3.14 shows that the accuracy of the reliability measures increases as the number of evaluated assessments increases. Moreover, it shows that even having low numbers of probabilistic assessments (as low as 10 in the case of moderate confidence and directional biases and as low as 2 or 3 in the case of extreme confidence and directional biases) is enough to give an indication of the existence and direction of confidence and directional biases.

Can assessors tell if they have moderate or extreme biases if they have only 2 or 3 assessments? Assessors can use the second row of Fig. 3.14 in addition to **Fig. 3.15** to qualitatively assess if they have moderate or extreme biases. Suppose that the assessors measured their CB using 2 or 3 assessments and it turned out to be 0.5 or greater. Fig. 3.14 shows that if the assessors' biases were extreme, there is about an 80% chance that their measured CB value is greater than 0.5 and if they were moderate, there is slightly more than 50% chance that the measured CB value is greater than 0.5. Furthermore, Fig. 3.15

shows that there is less than 1% chance that the measured CB value is equal to or greater than 0.5 if assessors have low or no bias. Therefore, in the case of measuring a CB of 0.5 or greater with 2 or 3 assessments, it is likely that they have moderate to extreme biases but they cannot tell if it is moderate or extreme until they evaluate more assessments.

Consider, on the other hand, that the assessors' measured CB value was 0.25 or lower. The assessors can discount the possibility of having extreme biases since there is less than 1 percent chance that they will measure a CB value of 0.25 or lower if extreme biases existed (Fig. 3.14). However, they cannot totally discount the possibility of moderate biases since there is about a 30-35% chance of measuring a CB of 0.25 or lower in the existence of moderate biases (Fig. 3.14). They will need to evaluate more assessments to find out if this measured CB is reflecting actual overconfidence or if they are actually well calibrated.

The directional bias measurement gets less accurate as the confidence bias gets closer to 0.0 (calibrated assessor with no confidence bias). **Fig. 3.16** shows the accuracy of the directional bias measurement of a number of assessors with directional bias of 0 (left column) and 0.9 (right column) and confidence bias of 0.1 (first row), 0.5 (second row), and 0.9 (last row). The plots show that for the assessors with low CB (first row), even a very large number of assessments (1000) was not enough to get a reasonably accurate measure of directional bias. This is in contrast to assessors with moderate to extreme CB (second and last rows). The reason is that at very low confidence bias values, the estimated distribution looks essentially the same regardless of the directional bias value (**Fig. 3.17**). This is also evident in the EVs (Fig. 3.2); the difference between EVs for different DB

scenarios decreases as CB decreases. Thus, this issue has little practical significance at low CB.

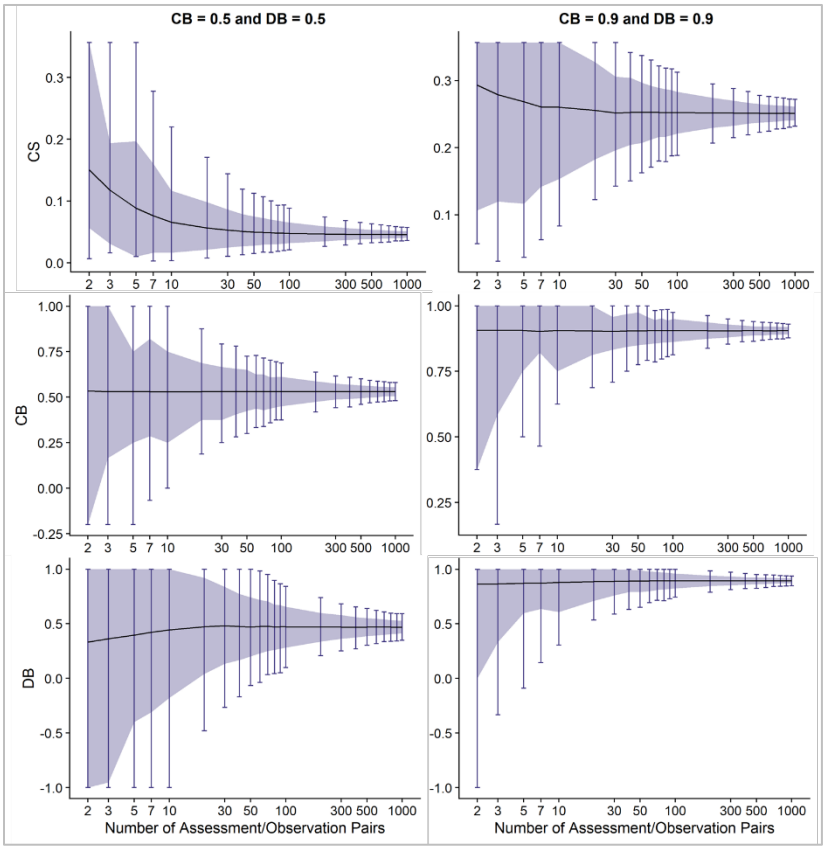


**Fig. 3.14—Confidence level in the CS, CB, and DB measurements with respect to numberof assessment/observation pairs.**

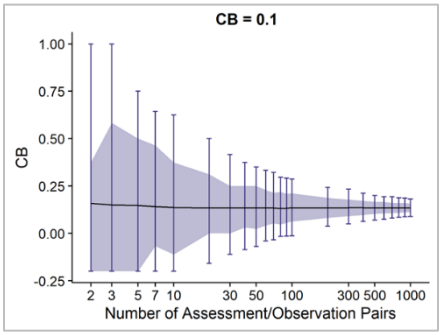


**Fig. 3.15—Confidence level in the CB measurement with respect to number of assessment/observation pairs assuming low confidence bias.**

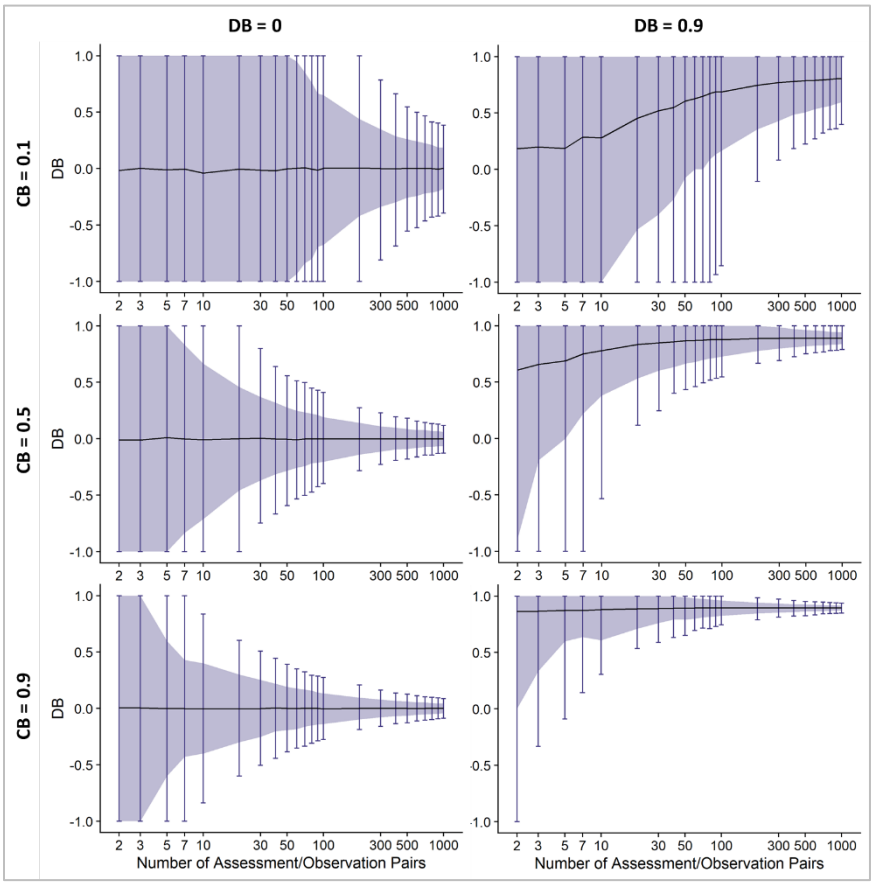


**Fig. 3.16—The confidence in directional bias measurement decreases as the confidence bias gets closer to 0.**

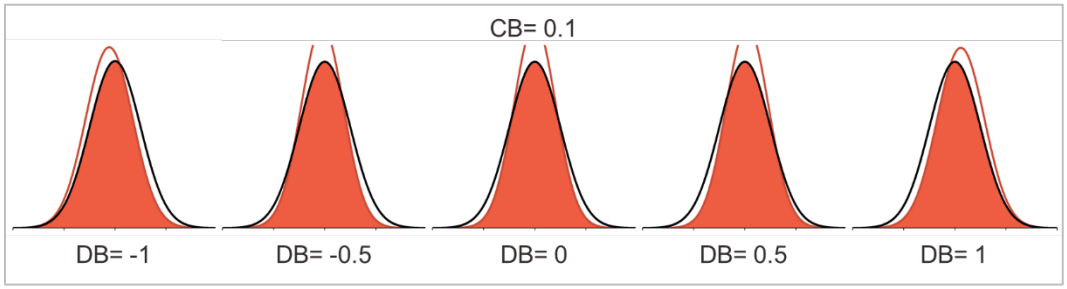


**Fig. 3.17—The estimated distribution looks essentially the same at very low confidence bias values.**

**External Adjustment Improves the Reliability of Probabilistic Assessments**

In the previous sections, we showed that look-backs and calibration plots can be used to detect and quantify the directions and magnitudes of biases. How does one make use of this information to eliminate biases? We discussed generally the process for eliminating biases in Alarfaj and McVay (2018). In summary, eliminating biases starts with training individuals involved in making probabilistic assessments. Next, probabilistic assessments should be monitored in a continual process of assessment tracking, look-backs as actual values become available, checking their calibration by comparing actual outcomes to forecasts and quantifying bias directions and magnitudes, and then using these measurements to improve new probabilistic assessments. If biases are detected, the individuals or the teams making the probabilistic forecasts should try to internally adjust their assessments to mitigate or eliminate biases. However, sometimes internal adjustment is not sufficient. In that case, external adjustment can be applied statistically.

There are a number of ways to externally adjust probabilistic assessments. As we mentioned in the Introduction, Capen (1976) used the coverage rate and probability plots to adjust forecasts. Fondren et al. (2013) expanded upon Capen's (1976) method and used calibration curves instead of the coverage rate. Landman and Goddard (2002) used model output statistics, a multiple linear regression technique, to recalibrate rainfall forecasts for extreme seasons over southern Africa using predictor values from a general circulation model and historical record of the predictand (regional rainfall indices). Piani et al. (2010) assumed that both normalized observed and simulated (estimated) distributions are well approximated by a gamma distribution and used a transfer function that can be derived

graphically to correct the simulated distributions. This is similar to using calibration plots to externally adjust assessment; however, the latter can be considered more general since it is not restricted to a specific distribution and the CDFs do not need to be normalized. Mandel and Barnes (2014) used Karmarker's transformation, which utilizes a tuning parameter to improve the calibration of forecasts in strategic intelligence applications. Turner et al. (2014) used a combination of forecast aggregation and recalibration (adjustment) using a linear-in-log-odds function to generate a less-biased forecast.

In this section, we will explain the statistical adjustment method suggested by Capen (1976) and improved by Fondren et al. (2013). The improved method uses the calibration curve to adjust new assessments made by the assessor. Next, we will show in a case study that combining this method with a continual process of assessment tracking, look-backs, calibration, and external adjustments will improve CS significantly while keeping the uncertainty relatively low.

*External Adjustment Using the Coverage Rate*

Capen (1976) used the coverage rate to externally adjust forecasts. Assume that we have the CR of an assessor, from historical assessments and observations, and further suppose that the assessor has issued a probabilistic assessment assuming an 80% central prediction interval. In other words, the assessor gave the P10 and P90 values. To adjust the assessment for bias, we plot the P10 and P90 values on a probability plot (normal or lognormal) at $P=0.5 \pm CR/2$. To get the adjusted range, we simply draw a line between the values and extend it to $P=0.1$ and $P=0.9$ to read the adjusted P10 and P90 values, respectively. This method assumes that the coverage rate is centrally located and therefore,

one disadvantage of this method is that it is insensitive to directional bias; thus, using this method will not be effective in adjusting for the directional bias of the probabilistic assessment.

To illustrate this method, suppose that an assessor estimated that the NPV of a certain project was between $80 million and $120 million with 80% confidence. Suppose that we have calculated the assessor's calibration curve from look-backs and calibration of historical assessments (**Fig. 3.18**). From the calibration plot, we can calculate the coverage rate:

$$CR = c_t(0.9) - c_t(0.1) = 0.83 - 0.42 = 0.41$$

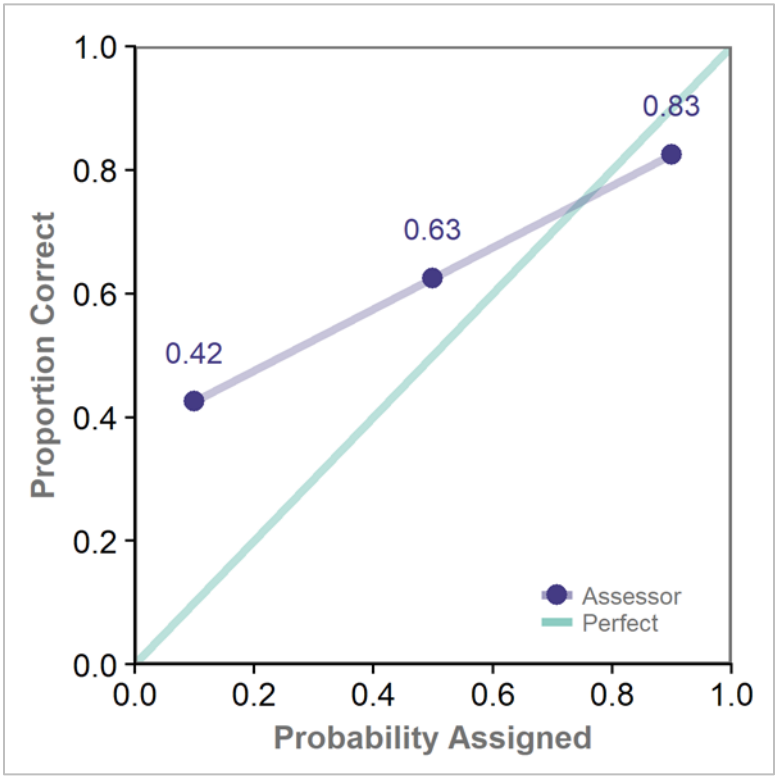


**Fig. 3.18—Calibration curve of the assessor from look-backs and calibration.**

If we assume that the project's NPV is normally distributed, we plot the unadjusted P10 and P90 values on a normal probability plot at $P$=0.5 ± 0.41/2. In other words, we plot \$80 million at $P \approx 0.3$ and \$120 million at $P \approx 0.7$. Finally, we draw a line between the two points and extend the line to read the adjusted P10 and P90 values as $P10_{adj} \approx$ \$52 million and $P90_{adj} \approx$ \$148 million (**Fig. 3.19**). Note that this method does not make use of the P50 value.

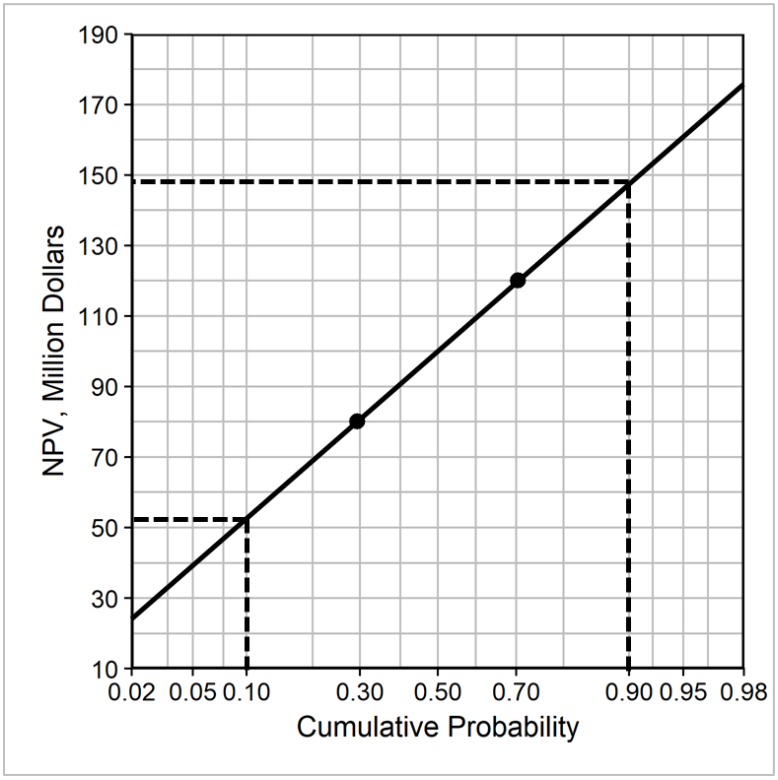


**Fig. 3.19—Graphical demonstration of external adjustment using a centrally located coverage rate and assuming a normal underlying distribution.**

If, on the other hand, we assume that the project's NPV is lognormally distributed, then we plot \$80 million at $P \approx 0.3$ and \$120 million at $P \approx 0.7$ on a lognormal probability plot. Next, we draw a line between the two points and extend the line to read the adjusted P10 and P90 values as $P10_{adj} \approx$ \$60 million and $P90_{adj} \approx$ \$159 million (**Fig. 3.20**).

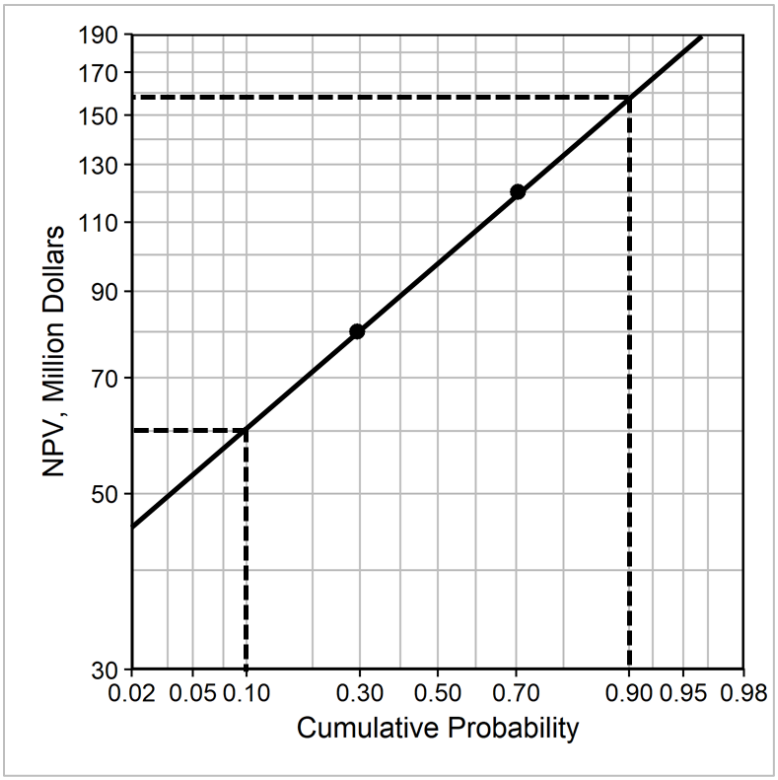


**Fig. 3.20—Graphical demonstration of external adjustment using a centrally located coverage rate and assuming a lognormal underlying distribution.**

Because computational tools are used more commonly than probability paper these days, Dossary (2016) derived the mathematical formulation of the coverage rate external adjustment method assuming that the underlying parameter is normally distributed.

Suppose that we are interested in the 80% prediction interval, then for probabilistic assessments that use the cumulative probability convention, the adjusted P10 and P90 values can be calculated [adapted from Dossary (2016)]:

$$P10_{adj} = \frac{P90+P10-\frac{\mathrm{Erf}^{-1}[PI_{width}]}{\mathrm{Erf}^{-1}[CR]}\cdot(P90-P10)}{2} \quad \text{.....(3.11)}$$

$$P90_{adj} = \frac{P90+P10+\frac{\mathrm{Erf}^{-1}[PI_{width}]}{\mathrm{Erf}^{-1}[CR]}\cdot(P90-P10)}{2} \quad \text{.....(3.12)}$$

where $PI_{width}$ is the width of the prediction interval (0.80 in our case). The inverse error function can be calculated in Microsoft® Excel™ using the following formula:

$$Erf^{-1}(x) = SQRT(GAMMAINV(x, 0.5,1)) \quad \text{.....(3.13)}$$

We extended Dossary's (2016) work and derived the mathematical formulation assuming that the underlying parameter is lognormally distributed. Then, the adjusted P10 and P90 values can be calculated as:

$$P10_{adj} = \exp\left(\frac{\ln(P90)+\ln(P10)-\frac{Erf^{-1}[PI_{width}]}{Erf^{-1}[CR]}\cdot[\ln(P90)-\ln(P10)]}{2}\right) \quad \text{.....(3.14)}$$

$$P90_{adj} = \exp\left(\frac{\ln(P90)+\ln(P10)+\frac{Erf^{-1}[PI_{width}]}{Erf^{-1}[CR]}\cdot[\ln(P90)-\ln(P10)]}{2}\right) \quad \text{.....(3.15)}$$

If we assume that the underlying distribution is normal in the last example, then Eqs. 2.11 and 2.12 can be applied to calculate the adjusted P10 and P90 values directly without using a probability plot:

$$P10_{adj} = \frac{120+80-\frac{\mathrm{Erf}^{-1}[0.8]}{\mathrm{Erf}^{-1}[0.41]}\cdot(120-80)}{2} \approx 52$$

$$P90_{adj} = \frac{120+80+\frac{\mathrm{Erf}^{-1}[0.8]}{\mathrm{Erf}^{-1}[0.41]}\cdot(120-80)}{2} \approx 148$$

in million dollars, which are the same results we got from using the probability plot.

If, on the other hand, we assume that the underlying distribution is lognormal in the last example, then Eqs. 2.142.15 and 2.15 can be applied to calculate the adjusted P10 and P90 values directly without using a probability plot:

$$P10_{adj} = \exp\left(\frac{\ln(120)+\ln(80)-\frac{Erf^{-1}[0.8]}{Erf^{-1}[0.41]}\cdot[\ln(120)-\ln(80)]}{2}\right) \approx 60$$

$$P90_{adj} = \exp\left(\frac{\ln(120)+\ln(80)+\frac{Erf^{-1}[0.8]}{Erf^{-1}[0.41]}\cdot[\ln(120)-\ln(80)]}{2}\right) \approx 159$$

in million dollars, which are again the same results we got using the probability plot.

*External Adjustment Using the Calibration Curve*

Fondren et al. (2013) extended Capen's (1976) work by plotting the P10, P50, and P90 propositions on a lognormal probability plot at the corresponding proportion-correct values (from the calibration curve) and then fitting a lognormal distribution through these three points using least-squares regression. This best-fit lognormal distribution is then used to calculate the adjusted P10, P50, and P90 values. In **Fig. 3.21**, we show that the corresponding graphical technique for this method is very similar to Capen's (1976) but, instead of using a centrally-located coverage rate, Fondren et al. (2013) used the proportion correct values from the calibration curve and fit a least-squares best-fit line to these propositions. Since the calibration curve is sensitive to directional bias, it can be more effective than using the centrally-located coverage rate in adjusting for directional bias. We note that they assumed that the number of observations is similar at each cumulative probability value. If the number of forecasts at each cumulative probability is

not the same, then, similar to measuring biases from the calibration curve, a weighted least-square method should be used where each point on the calibration plot is weighted by the number of forecasts used to generate this point.

For example, consider the assessor from the previous example. However, this time we will use the proportion correct values corresponding to P10 and P90 in the calibration curve. That is, in this example, we are plotting the P10 value ($80 million) at a cumulative probability of 0.42 (instead of 0.40) and the P90 value ($120 million) at a cumulative probability of 0.83 (instead of 0.70). Just like in the previous method, we fit a straight line through these points (assuming a lognormal distribution) and extend the line to read the adjusted values as $P10_{adj} \approx \$55$ million and $P90_{adj} \approx \$135$ million. The adjusted values obtained using this method are less than the adjusted values obtained using the centrally-located coverage rate ($60 million and $159 million) because in this method we accounted for the positive directional bias indicated in the calibration curve.

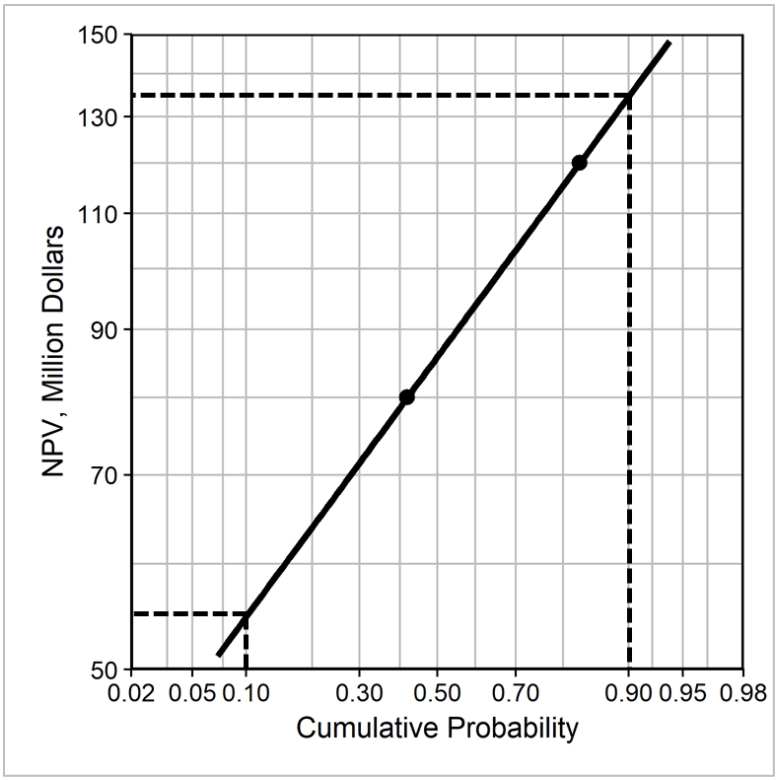


**Fig. 3.21—Graphical demonstration of external adjustment using the calibration curve and assuming a lognormal underlying distribution defined by P10 and P90 only.**

So far, we have used only the P10 and P90 values in our external adjustment procedure. Suppose that the same assessor from the last example has also proposed that the NPV of the project has a P50 value of $100 million. On Fig. 3.21, we add the P50 value at a cumulative probability of 0.63. Finally, we draw a best-fit line through these three points and extend the line to read the adjusted values as $P10_{adj} \approx \$56$ million, $P50_{adj} \approx \$87$ million and $P90_{adj} \approx \$135$ million (**Fig. 3.22**). Note, that in this case, adding the P50 values only slightly changed the adjusted P10 value. However, if the P50 value was more extreme, its effect on the adjusted values would be more pronounced.

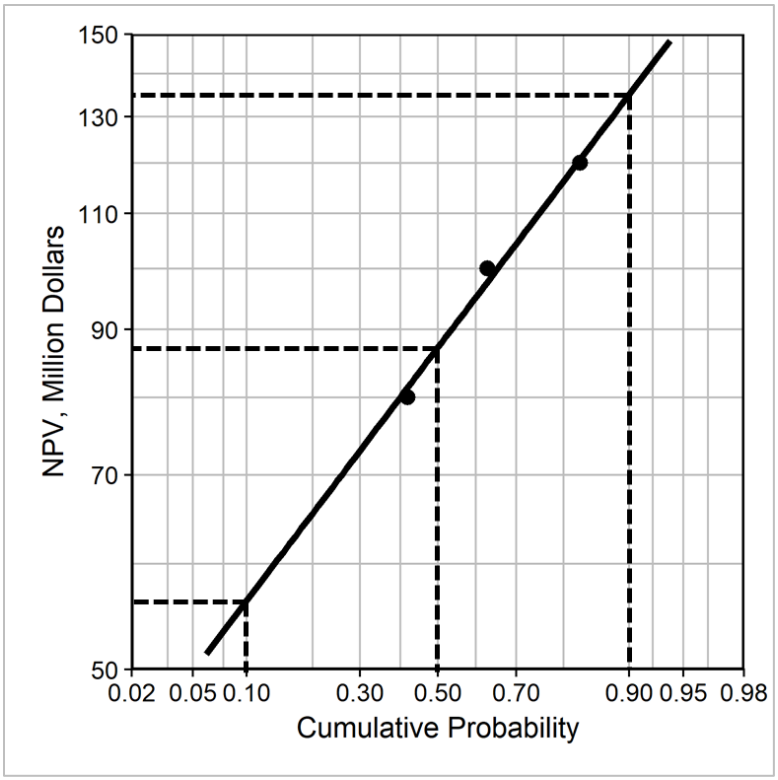


**Fig. 3.22—Graphical demonstration of external adjustment using the calibration curve and assuming a lognormal underlying distribution defined by P10, P50, and P90.**

Of course, the same methodology can be applied on a normal probability plot if the estimated parameter is normally distributed. Furthermore, we can use more *P* values (e.g., P10, P20, P30, …, and P90) if needed and if a calibration curve defined at these percentiles is available.

In mathematical form, our objective is to fit a distribution using least-squares regression. In other words, we find the distribution that will minimize the sum of the squares of the differences between the cumulative probabilities of the proposed values and their corresponding proportion correct values:

$$Minimize \quad \sum_{t=1}^{T}(F(x_t) - c_t)^2 \quad \text{..... (3.16)}$$

where $F$ is the cumulative distribution function of the fitted distribution, $c_t$ is the proportion correct from the calibration curve, and $T$ is the number of defined cumulative probabilities or probability subintervals. $F$ can be normal, lognormal, or any kind of distribution that describes the underlying parameter adequately. If insufficient information is available about the estimated parameter distribution, we suggest using a PERT distribution. A PERT distribution is flexible and can approximate both normal and lognormal distributions. However, unlike normal and lognormal distributions, fitting a PERT distribution will require the estimated distribution to be defined by at least three points (because it is defined by 3 parameters) in comparison to at least two points for the normal and lognormal distributions (because they can be defined by 2 parameters only). Furthermore, the PERT distribution is bounded, unlike normal and lognormal distributions, which may or may not be a desirable trait depending on the quantity assessed. After we find the distribution using least-squares regression, we can use it to calculate the adjusted percentiles.

The example in Fig. 3.22 can be solved numerically by using a solver (such as the Solver add-in in Microsoft® Excel™). We solved for a lognormal distribution such that the unadjusted P10, P50, and P90 values would be closest to its 42$^{nd}$, 63$^{rd}$, and 83$^{rd}$ percentiles, respectively. The minimization of the function in Eq. 2.16 results in a lognormal distribution with a mean of $92.5 million and a standard deviation of $32.5 million. The cumulative distribution function (CDF) of the fitted lognormal function is shown in **Fig. 3.23a**. We used the quantile function (the inverse of the CDF; in Microsoft® Excel™: NORM.INV for the normal distribution, and LOGNORM.INV for the lognormal

distribution), to calculate the adjusted P10, P50, and P90. **Fig. 3.23b** shows that the adjusted P10, P50, and P90 values are $56 million, $87 million, and $135 million, respectively, which are the same results we got graphically. **Table 3.1** shows a summary of the results. For completeness, we added the results of external adjustment using the calibration curve and assuming a normal underlying distribution.

The adjustment process is done on new assessments one assessment at a time and will likely produce adjusted new assessments that are significantly better calibrated than the unadjusted new assessments, particularly if (1) the historical biases are large and, thus, the adjustments are large, and (2) the new assessments are similar in kind to the group of historical assessments that were used to generate the calibration curve. However, improvement in calibration is not automatically assumed. Rather, assessors should apply the same rigorous process of assessment tracking, look-backs, calibration, and quantification of biases to the adjusted new probabilistic assessments going forward.

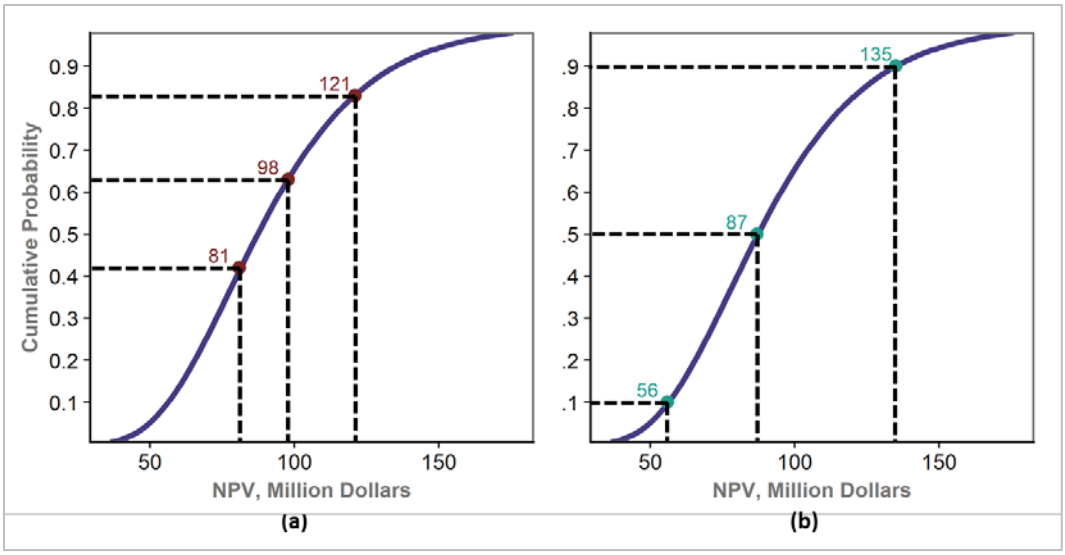


**Fig. 3.23—(a) Least-squares regression is used to fit a lognormal distribution to assessment percentiles that are plotted at probabilities corresponding to their proportion correct value from the calibration curve. (b) New percentiles are calculated from the fitted distribution.**

| Method | Function | Adjusted (Million $) | | |
|---|---|---|---|---|
| | | P10 | P50 | P90 |
| **Centrally located coverage rate** | Normal | 52 | 100 | 148 |
| | Lognormal | 60 | 98 | 159 |
| **Calibration curve** | Normal using P10 and P90 Only | 43 | 87 | 131 |
| | Normal using P10, P50 and P90 | 43 | 88 | 132 |
| | Lognormal using P10 and P90 Only | 55 | 86 | 135 |
| | Lognormal using P10, P50 and P90 | 56 | 87 | 135 |

Table 3.1—A comparison of forecast adjustment methods assuming normal and lognormal distributions. The choice of distribution type for fitting the calibration curve and the new estimates is important.

We note that this method assumes that both the new and historical distributions can be well approximated by the same distribution type (such as normal or lognormal).

Incidentally, many of the processes in petroleum engineering (and nature in general) can be well approximated by a normal or lognormal distribution (Capen 1976). Also, a PERT distribution can roughly approximate both normal and lognormal distributions. Furthermore, this method assumes that the biases measured from historical probabilistic assessments reflect the assessor's overall biases. Using a larger number of historical probabilistic assessments increases the confidence that these measured biases reflect the overall biases of the assessor. The evaluator can use Figs. 3.14-3.16 as a qualitative guide for how many assessments are needed for a specific level of confidence in the measures of biases. Finally, this method assumes that the assessor will have similar levels of biases in his/her historical and new assessments. One reason for why the levels of biases might differ between historical and new assessments is that the assessor is evaluating different types of assessments (such as estimating the ultimate recovery versus estimating the net present value). Ensuring that the historical and new assessments are similar in type should eliminate this as a possible cause of error in the adjustment process. However, there can be other reasons why the levels of biases might differ between historical and new assessments. For example, the assessor could be internally self-adjusting over time with information gained from look-backs and calibration. However, if assessors follow a continual process of assessment tracking, look-backs, calibration, and adjustment, they will eventually incorporate these differences into the external adjustment process.

The examples above show that externally adjusting assessments (in the case of overconfidence, which is more common) typically results in wider distributions (higher SD). We showed in our previous work (Alarfaj and McVay 2018), that there is more to

gain from eliminating biases than from reducing uncertainty. Therefore, the first priority is improving calibration and the second priority is reducing uncertainty (smaller width or SD). That is, if comparing two assessment methods, the evaluator should select the method with better CS if the CS values differ significantly. If the CS values are not significantly different, then the evaluator should pick the one with the lower uncertainty.

**Case Study**

In the previous sections, we showed how to measure calibration of probabilistic assessments and how to externally adjust new assessments to improve their probabilistic reliability. In this section, we will apply the methods described in this paper on probabilistic decline-curve analysis (DCA) from the literature and assess the effectiveness of externally adjusting probabilistic forecasts using calibration curves.

Gonzalez et al. (2012) evaluated the performance of the Markov Chain Monte Carlo (MCMC) method using different DCA models on 197 Barnett shale gas wells. They performed a hindcast study where only a portion of the hindcast period (the time period for which historical data was available) was matched, predictions were made for the remainder of the hindcast period, and comparison was made between the predicted and actual cumulative production at the end of the hindcast period (CPEOH). They varied the portion of historical data that was matched between 6 months and 36 months. They used MCMC coupled with a number of different DCA models. In this case study, we will evaluate the performance of external adjustment using a subset of the DCA models they examined, specifically Arps (Arps) with 5% minimum decline and the power-law (Ilk et al.) models.

*External Adjustment of Long-Term Probabilistic Assessments*

To test the external adjustment procedure, Fondren et al. (2013) evaluated and generated a calibration curve for a set of hindcasts and used it to adjust the same set. This process is somewhat circular because we expect to get nearly perfect calibration when we use the calibration curve to adjust the same assessments that were used to generate it. Indeed, Fondren et al. (2013) showed that the external adjustment process improved the probabilistic assessments markedly and resulted in a nearly perfect calibration curve. Furthermore, they noted that improving the calibration of the same set, other than to verify the calibration/adjustment procedure, is not particularly helpful because it is necessary to have the actual production data to do calibration and, therefore, the adjusted assessments do not add value because we already know the outcomes. The external adjustment process potentially adds value only when we use calibration results of historical probabilistic assessments to adjust new probabilistic assessments.

For a better test of the effectiveness of external adjustment, we picked a set of 197 hindcasts generated using 6 months of production data and randomly divided it into two groups. We used the first group (100 wells) to measure the calibration and generate the calibration curve. Then we used the calibration curve of the first group of hindcasts to externally adjust the second group (97 wells) of hindcasts assuming a PERT distribution. Finally, we measured the calibration of the hindcasts of the second group with and without adjustment. We did this for two distinct sets that were generated using power-law and Arps-with-5%-minimum-decline models.

One potential issue with this approach is that improvement (or deterioration) in the calibration of the adjusted hindcasts of the second group could be affected by selection bias. In other words, if we randomly divided the two groups again, we will probably get different calibration curves for both the unadjusted and the adjusted hindcasts. From one iteration, we will likely not be able to distinguish between differences in calibration caused by random selection versus that which was caused by the external adjustment process. To mitigate the effects of selection bias, we repeated the steps presented in the previous paragraph for 1000 iterations using Monte Carlo simulation. From these iterations, we calculated the expected calibration curve (the average proportion correct at each assigned probability value for all of these iterations). Therefore, any improvement or deterioration in calibration caused by the selection bias should cancel out.

**Fig. 3.24** shows the expected calibration curves for the second group of hindcasts before and after external adjustment assuming a PERT distribution for (a) the power-law and (b) the Arps-with-5%-minimum-decline models. External adjustment for the power-law model improved the CS from 0.0129 to 0.0021 (**Table 3.2)**. However, this improvement in calibration resulted in a significant increase in the average width (AW) of the 80% CI from 1,143 MMSCF for the unadjusted to 1,673 MMSCF for the adjusted hindcasts. Note that reducing the CS does not necessarily increase the AW. We have shown in a previous section that the CS value does not distinguish between over and underconfidence. Therefore, external adjustment should reduce both the CS and the AW in the case of underconfidence. Furthermore, CS can also be reduced by reducing the magnitude of DB while keeping CB constant (Fig. 3.11), which keeps the AW roughly

constant (since CB and AW are correlated). For the Arps-with-5%-Minimum-Decline model, external adjustment reduced the CS significantly from 0.0318 to 0.0025. Furthermore, in doing so, the AW was slightly reduced from 772 MMSCF to 747 MMSCF. Fig. 3.24 shows that most of the improvement in calibration came from the reduction of directional bias as suggested by the significant shift downward of the calibration curve.

The choice of fitted distribution type affects the external adjustment process. Table 3.2 shows that, in terms of CS and AW, external adjustment using a PERT distribution did slightly better than using a lognormal distribution. We reiterate here our recommendation to use a PERT distribution for adjustment rather than normal or lognormal, especially if the analyst is not sure of the shape of the estimated parameter. Table 3.2 also shows that while using a normal distribution for the external adjustment process also improves the calibration, the benefits in both CS and the AW are not as significant as when using a lognormal or a PERT distribution. In particular, using a normal distribution with the Arps-with-5%-Minimum-Decline model caused a significant increase in the AW that is not observed when using PERT or lognormal distributions. Ultimately, using a normal distribution fit to externally adjust these specific models is not optimum.

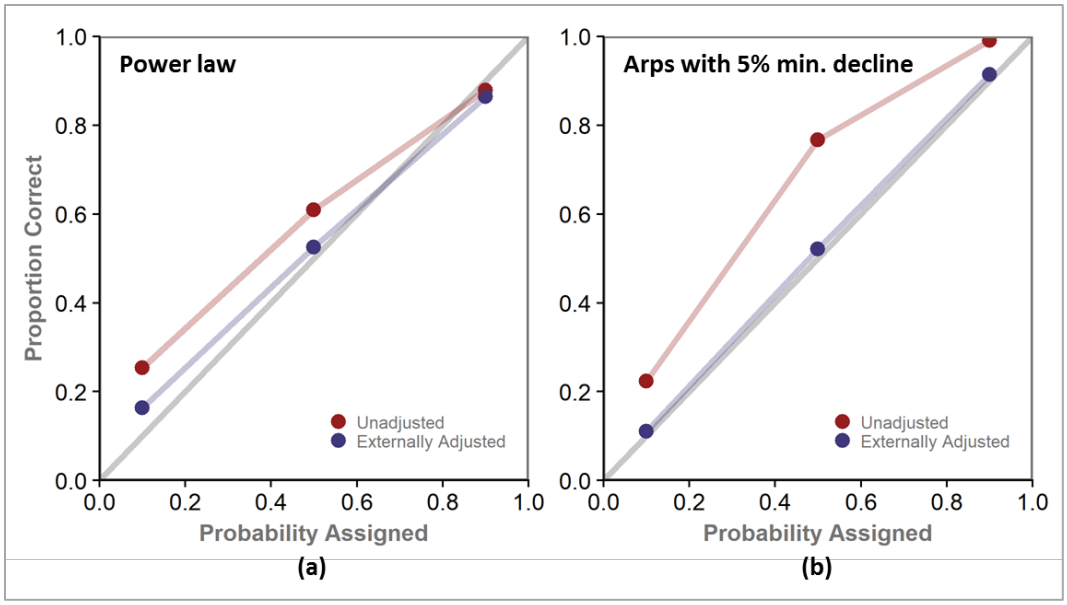


**Fig. 3.24—Expected calibration curves of the (a) the power-law model, and the (b) Arps-with-5%-minimum-decline model externally adjusted by fitting a PERT distribution.**

| | Fitted Distribution | Calibration Score | Average 80% CI Width (MMSCF) |
|---|---|---|---|
| **Power law** | Unadjusted | 0.0129 | 1,143 |
| | normal | 0.0055 | 1,598 |
| | Lognormal | 0.0051 | 2,003 |
| | PERT | 0.0021 | 1,673 |
| **Arps with 5% minimum decline** | Unadjusted | 0.0318 | 772 |
| | normal | 0.0136 | 1,778 |
| | Lognormal | 0.0029 | 772 |
| | PERT | 0.0025 | 747 |

Table 3.2—A comparison of the performance of external adjustment using normal, lognormal, and PERT distributions. The choice of distribution type for fitting the calibration curve and the new estimates affects the method's performance.

*Effects of Using a Lower Number of Historical Probabilistic Assessments*

In the previous experiments, we used nearly half the wells (100) from the data set for calculating the calibration curve and adjusted the remaining half. How would the results be affected by using a much lower number of wells to generate the calibration curve? One of the key assumptions of this external adjustment process is that we have enough probabilistic assessments to measure the calibration curve accurately. We have shown in a previous section that using a number of assessments as low as 10 in the case of moderate confidence and directional biases and as low as 3-5 in the case of extreme confidence and directional biases would result in a calibration curve that adequately describes the magnitude and direction of bias. However, would these low numbers of assessments produce calibration curves that are accurate enough to significantly improve the reliability of new assessments?

To answer this, we repeated the previous experiment three more times using 50, 25, and 10 wells to generate the calibration curve and applied external adjustment to as many wells using a PERT distribution. To allow easier comparison between the experiments, we plotted the net calibration score defined as follows:

$$Net\ CS = CS_{adjusted} - CS_{unadjusted} \quad \text{..... (3.17)}$$

where a negative net calibration score means that the adjusted assessments were more reliable than the unadjusted assessments because CS is negatively oriented (the lower the number, the better).

**Fig. 3.25** shows the net calibration score for the (a) power-law and (b) Arps-with-5%-Minimum-Decline hindcasts, where the middle curve shows the expected net calibration

score, the shaded area signifies an 80% CI and the error bars signifies a 99% CI. Fig. 3.25a shows that there is a chance that external adjustment produces worse adjusted assessments when using a lower number of historical probabilistic assessments (more likely than not for the power-law model when using only 10 historical assessments). However, this chance reduces significantly as more historical assessments are used. Furthermore, increasing the number of probabilistic assessments used to generate the calibration curve will also increase the chance that the adjustment process will improve calibration. Fig. 3.25 also shows that fewer historical probabilistic assessments are needed to improve the calibration score of the Arps-with-5%-minimum-decline model than the power-law model. This is because the Arps model is more biased than the power-law model (Fig. 3.24) and, therefore, the bias can be more easily measured (Fig. 3.14). Therefore, the Arps model stands to benefit more from the adjustment process, because its biases are measured more accurately with fewer historical assessments.

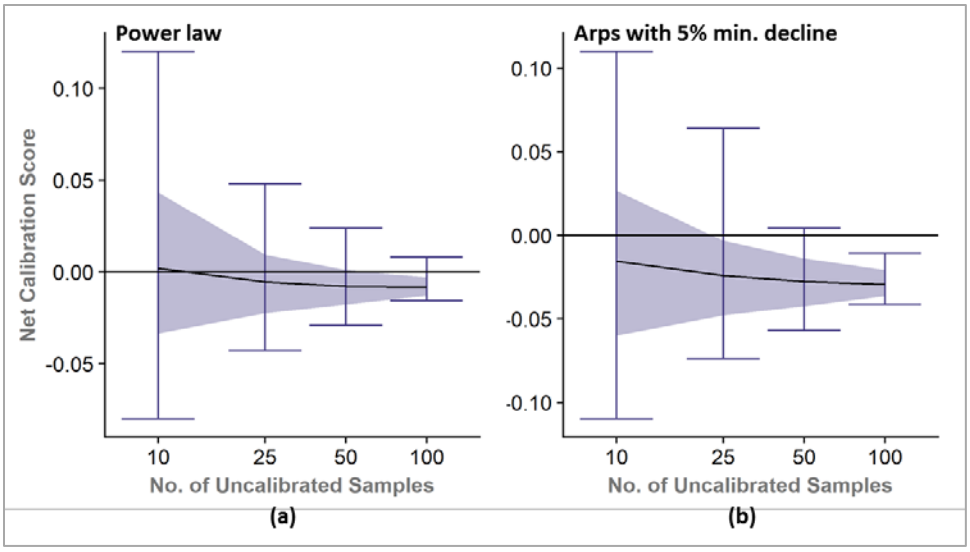


**Fig. 3.25—Net calibration score for (a) the power-law model, and the (b) Arps-with-5%-minimum-decline model externally adjusted by fitting a PERT distribution. The middle curve shows the expected net calibration score, the shaded area signifies an 80% CI and the error bars represent 99% CI.**

*Using Short-Term Assessments to Externally Adjust Long-Term Assessments*

The probabilistic DCA assessments discussed in the previous example are considered long-term. That is, for most wells in this dataset, it will take years before we can measure their calibration and use it as a feedback for new assessments. It is possible that most of the field would have been fully developed by the time we get feedback and the calibration results will add very little value. McVay et al. (2005) suggested generating short-term and long-term assessments, and then adjusting the long-term assessments using calibration information from the short-term assessments. In this subsection, we follow the Fondren et al. (2013) example in evaluating four possible options (explained in the next paragraph) that the operator will be faced with when acquiring new data. We will then introduce a fifth option and compare the performance of each option. We did our analysis on a

normalized-to-time-zero basis. Thus, we assumed that all of the 197 wells start producing at the same time and we measured their 24-month calibration at the same time.

We used the same two sets (power law and Arps with 5% minimum decline) of 197 long-term hindcast assessments introduced in the previous subsections (see for example Well 37 in **Fig. 3.26**). These hindcasts were generated by fitting only the first 6 months of production data. Using these long-term assessments, we generated a number of short-term assessments, which are essentially subsets of the long-term assessments (see the 24-month assessment example of Well 37 in **Fig. 3.27**). At 24 months (after the passage of 18 months from the time the short and long-term assessments were generated), the actual production for these 24 months is available. At this point in time, the operator has at least four options for how to treat the long-term assessments. Option 1 is to keep the initial assessments and do nothing with the newly acquired data. Option 2, which is what is typically done in the industry when updating assessments, is to use the same method to generate new assessments using the 24 months of production data without measuring or relying on calibration information. Option 3 is to use the calibration information to externally adjust the original long-term assessments made using only 6 months of production data. Finally, Option 4 is to generate new assessments by fitting the 24 months of production data and then externally adjust these new assessments using the calibration information from the short-term assessments generated using only 6 months of production data.

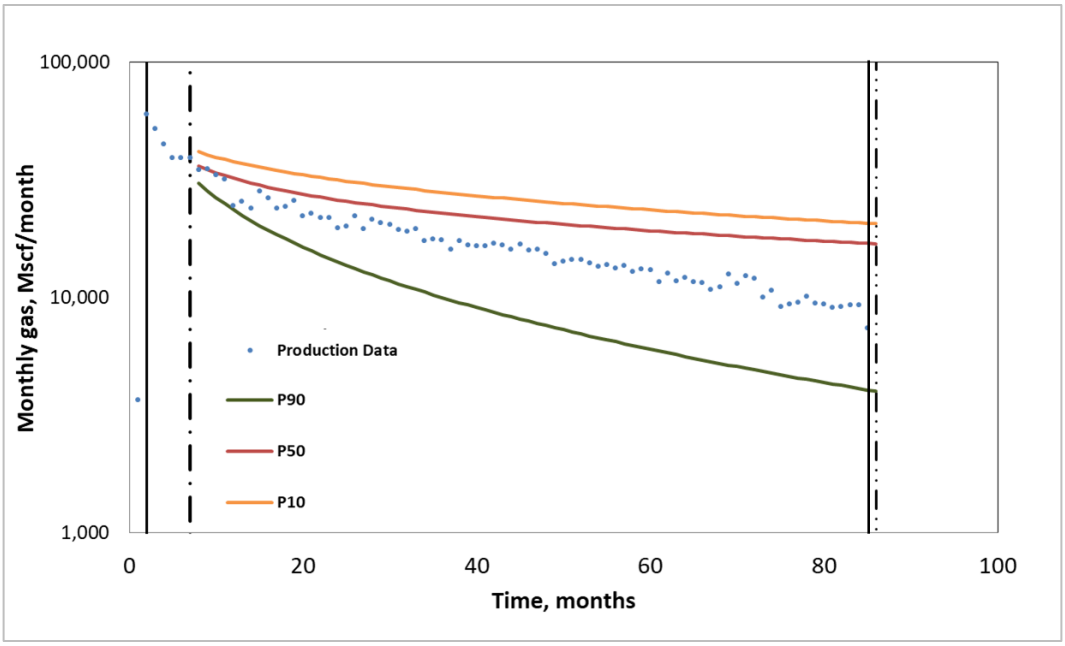


**Fig. 3.26—Long-term probabilistic production hindcast for well 37.**

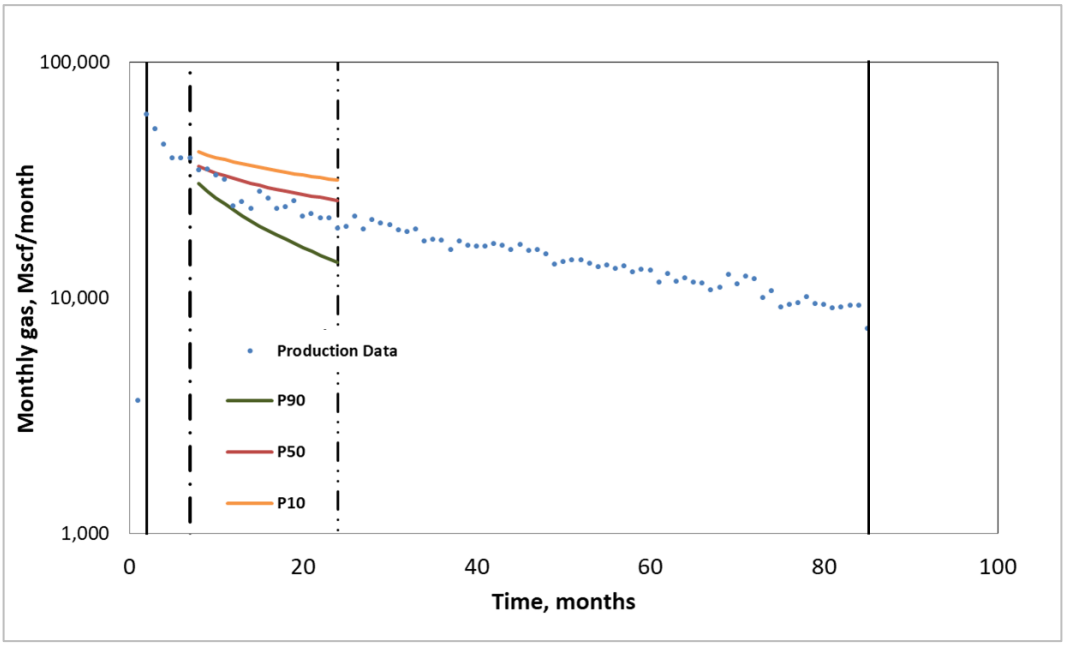


**Fig. 3.27—Short-term probabilistic production hindcast for well 37.**

We ran each of these possible options and then checked the reliability of the long-term assessments. **Table 3.3** shows that not adjusting the assessments (Option 1) produced the worst long-term assessment CS and second-worst AW, for both power law and Arps with 5% minimum decline. Measuring the calibration at 24 months and using it to update the long-term assessments that were generated using 6 months of production data (Option 2) improved the CS significantly, especially in the case of Arps with 5% minimum decline. However, it also resulted in a larger AW, which is not unexpected given that the short-term assessments were overconfident (CB is 0.28 and 0.18 for the Power Law and the Arps, respectively). Updating the forecasts by running the MCMC DCA models using 24 months of data (in contrast to 6 months) without adjustment (Option 3) improved the CS and the AW over Option 1. Updating the long-term forecasts by running the MCMC DCA models using 24 months of data combined with external adjustment (Option 4) improved both the CS and the AW over Option 1; however, the AW was not improved as much as in Option 3 because typically (in the case of overconfidence), there is a tradeoff between the CR/CS and the AW (AW increases as CR/CS decrease). Even though Option 3 has lower AW, Option 4 is more preferable because it has lower CS, which indicates lower biases. We mentioned in the previous section that best decisions are made and portfolio value is maximized when biases are minimized. Operators would benefit most from reducing biases (inferred from CS) first, then reducing uncertainty (inferred from AW) second. Later in this section, we introduce a fifth option that will reduce biases while maintaining a relatively low AW.

| Method | Forecast Update Option at 24 Months | Proportion Correct | | | CR | CS | AW (MMSCF) |
|---|---|---|---|---|---|---|---|
| | | 0.10 | 0.50 | 0.90 | | | |
| **Power law** | Option 1: Do nothing | 0.2767 | 0.6070 | 0.8546 | 0.5779 | 0.0149 | 15,214 |
| | Option 2: External adjustment of original forecasts | 0.1912 | 0.5363 | 0.8411 | 0.6499 | 0.0044 | 20,180 |
| | Option 3: Update MCMC using 24 months of production data but no external adjustment | 0.1874 | 0.5325 | 0.8282 | 0.6408 | 0.0046 | 8,957 |
| | Option 4: Update MCMC using 24 months of production history with external adjustment | 0.1079 | 0.4325 | 0.8408 | 0.7329 | 0.0027 | 12,591 |
| **Arps with 5% min decline** | Option 1: Do nothing | 0.2776 | 0.7383 | 0.9310 | 0.6534 | 0.0298 | 10,286 |
| | Option 2: External adjustment of original forecasts | 0.1163 | 0.5510 | 0.9101 | 0.7938 | 0.0010 | 13,348 |
| | Option 3: Update MCMC using 24 months of production data but no external adjustment | 0.1951 | 0.5969 | 0.8530 | 0.6579 | 0.0069 | 6,065 |
| | Option 4: Update MCMC using 24 months of production history with external adjustment | 0.1036 | 0.3852 | 0.8252 | 0.7216 | 0.0063 | 7,497 |

Table 3.3—Proportion correct, CR, CS, and AW for the MCMC long-term assessments.

Why did the external calibration method improve the CS of the long-term assessments more in Option 4 than Option 3 in the case of power law (from 0.0046 to 0.0027) but not as much in the case of Arps with 5% minimum decline (from 0.0069 to 0.0063)? One of the key assumptions of this external adjustment procedure is that the assessor will have consistent levels of biases over time, similar to what is shown in the hypothetical true and estimated distributions in **Fig. 3.28**. The procedure will not perform as well when the bias levels are changing (**Fig. 3.29a**), or worse, changing from overconfidence to underconfidence or vice versa (**Fig. 3.29b**). We suggest that more frequent look-back, calibration, and external adjustment will improve the results of this process. Suppose that

we are trying to improve the reliability of the estimated distribution in **Fig.** 3.30**a** where, as defined in the Introduction, the true distribution is a hypothetical distribution and represents the perfectly reliable distribution; i.e., not affected by biases. In Fig. 3.30a, the long-term forecast (at $t$ = 54) is underconfident while the corresponding short-term forecast (at $t$ = 12) is overconfident. Adjusting the long-term forecast using the calibration curve of a short-term forecast will result in an even more underconfident long-term forecast (**Fig.** 3.30**b)**. However, it will also cause the newly adjusted forecast to be less overconfident (moving towards underconfidence) at the next time step (see the difference between the unadjusted and adjusted estimates at $t$ = 18 in Fig. 3.30b). So, at the next round (at $t$ = 18), we measure the calibration of the estimate adjusted at $t$ = 12 and use the calibration information to adjust the newly updated model that uses the data available up to $t$ = 18. After calibration measurement, model updating and external adjustment, the long-term forecast will be less underconfident (moving away from underconfidence, **Fig. 3.30c**). We repeat the same process at $t$ = 24 and we get an even better calibrated long-term estimate (**Fig. 3.30d)** After few rounds of calibration measurements and external adjustment, the long-term estimate should be better calibrated.

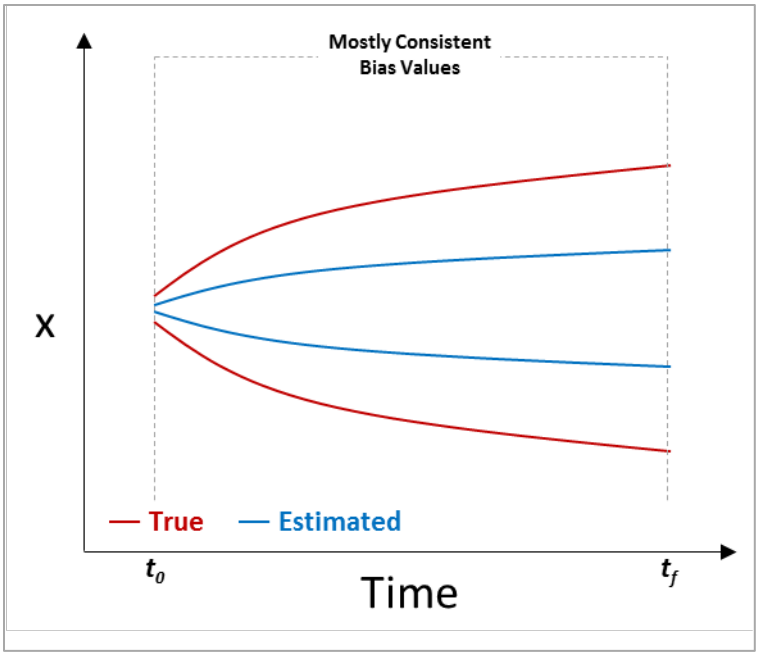


**Fig. 3.28—True and estimated distributions in a time series that exhibit relatively consistent overconfidence bias values over time.**

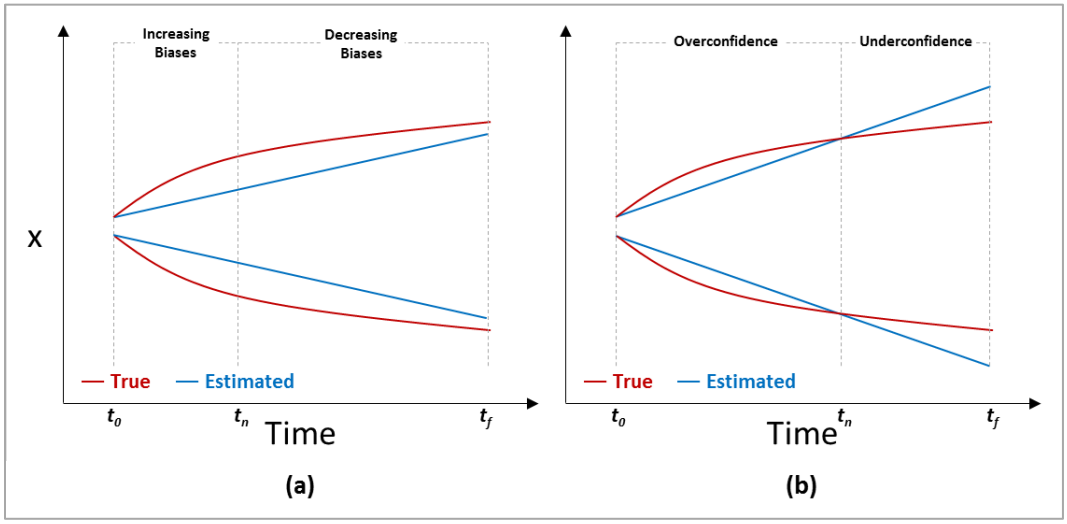


**Fig. 3.29—True and estimated distributions in a time series that exhibit (a) increasing/decreasing bias values over time, and (b) bias values that switches from overconfidence to underconfidence over time.**

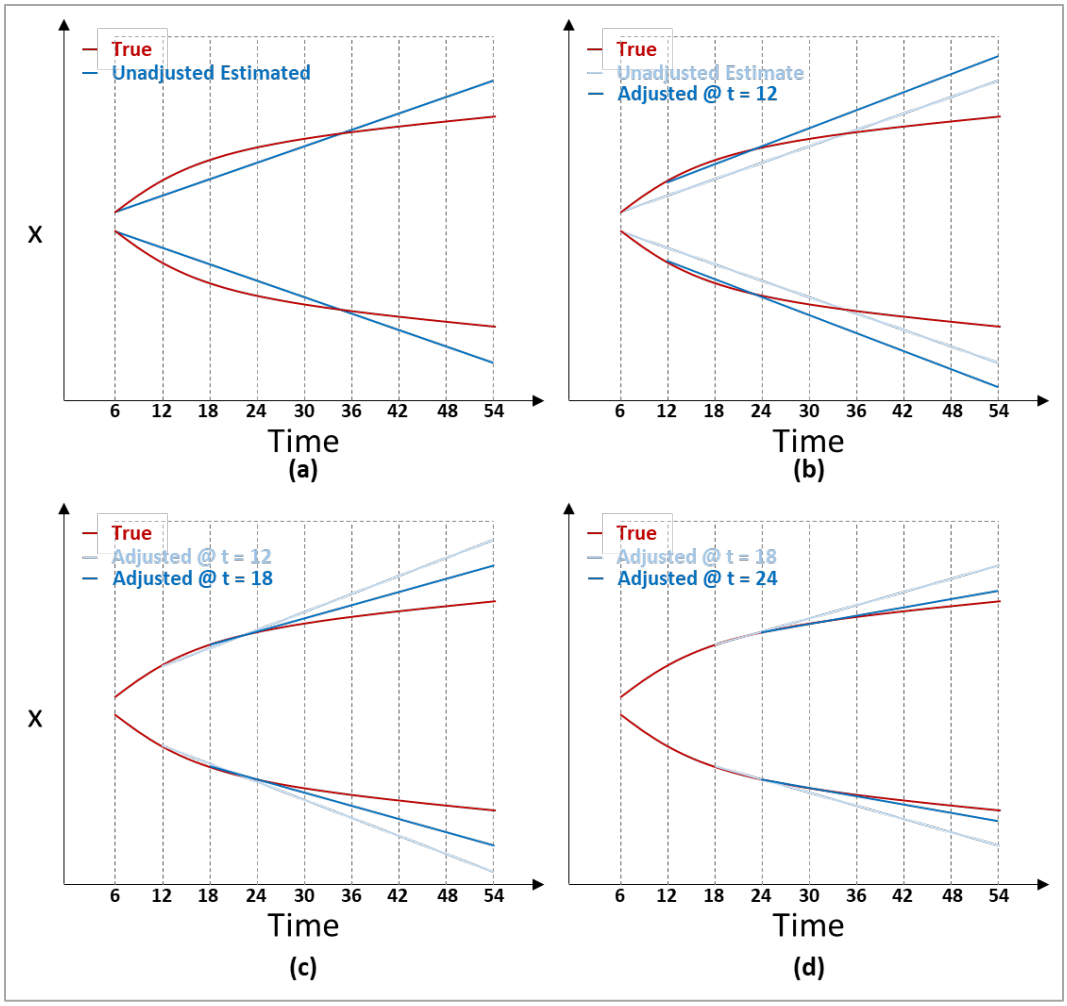


**Fig. 3.30— more frequent look-back, calibration, and external adjustment mitigate the issue of having different biases over a time series.**

We compared frequent look-back, calibration, and adjustment to the one-time external adjustments that we did in the previous example. In **Figs. 3.31 and 3.32,** we compare the CS and AW of the long-term assessments generated by applying the 5 different options with the power law and the Arps models, respectively. In the first option, the assessor keeps the initial assessments made using 6 months of production data and does not make use of calibration results or new production data to update the model (red curves in Figs.

3.31 and 3.32). In the second option, the assessor measures the calibration at given time *t* (12, 18, 24, 30, or 36 months in this figure) and uses the calibration measurements to externally adjust the initial assessments that were made using 6 months of production data (yellow curves). In the third option, all the available production data at time *t* are used to update the MCMC model and generate a new set of probabilistic assessments without measuring calibration or applying external adjustment (green curves). In the fourth option, the assessor updates the probabilistic assessments using all the production data available at time *t* and externally adjusts them using the calibration measurement of the initial assessments (that were generated using 6 months of production data) for the same period (blue curves). Note that the first 4 options here are essentially the same 4 options investigated above (Table 3.3) and involve no more than one update; the difference being that here the one update is performed at different times. The fifth option investigated is to continuously update the assessments with new production data every 6 months and externally adjust the updated assessments using the calibration curves of the adjusted assessments of the last time period (purple curves). Note that in the first four options, the CS and AW measurements at any given time *t* are independent of the CS and AW measurements at other times. That is, suppose we wanted to calculate CS for the fourth option at 30 months. We will need the probabilistic assessments generated using 6 months of production data and all the production data up to 30 months. We do not need to know the CS value at 18 months (for example) to calculate the CS value at 30 months. In contrast, for the fifth option, we use the calibration score of the assessments generated at the last time step to externally adjust the assessments generated at the current time step.

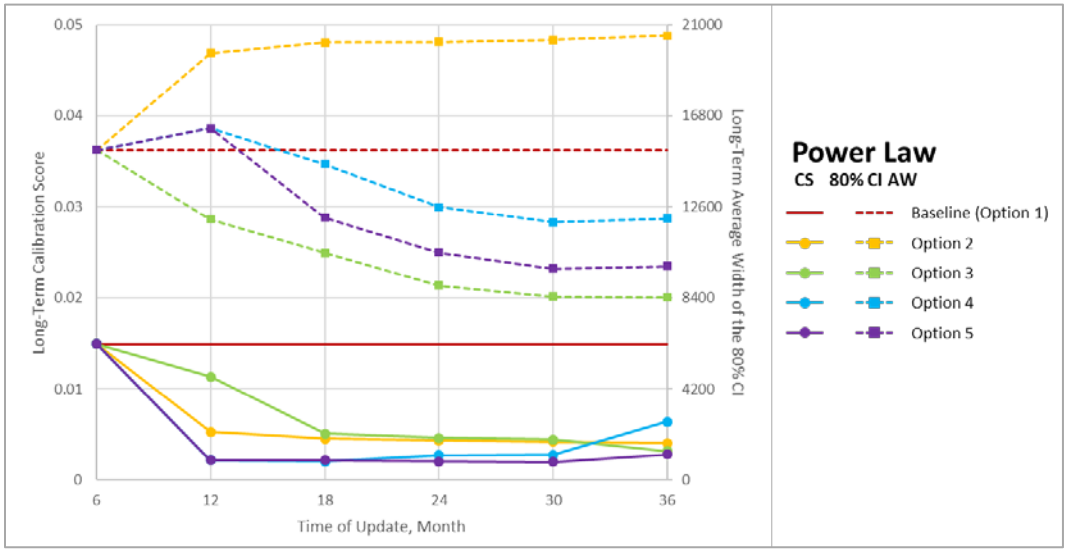


**Fig. 3.31—Calibration score and the average width of the 80% confidence interval of the long-term probabilistic assessments following different look-back, calibration, and external adjustment options for the power-law probabilistic assessment method.**

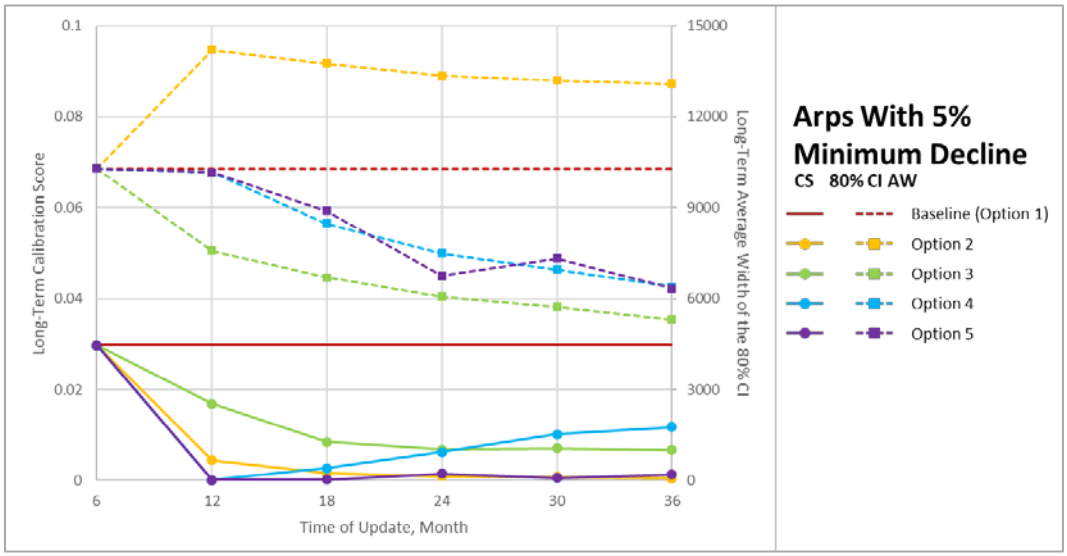


**Fig. 3.32—Calibration score and the average width of the 80% confidence interval of the long-term probabilistic assessments following different look-back, calibration, and external adjustment options for the Arps-with-5%-minimum-decline probabilistic assessment method.**

Figs. 3.31 and 3.32 show that neither updating the model nor using external adjustment results in the worst possible calibration score and a high AW (Option 1, red). External

adjustment without updating the MCMC model with the new production data results in significant reduction in CS (Option 2, yellow). However, this is coupled with a pronounced increase in AW because the historical assessments are mostly overconfident and the option does not benefit from the AW reduction that results from incorporating more production data. Updating the MCMC model with new production data without applying look-back, calibration, and external adjustment (Option 3, green) results in the lowest AW, but it produces more biased estimates than Options 2 and 5. Updating with new production data and adjusting new assessments using calibration information (Option 4, blue) produced low CS initially; however, it started to increase beyond 18 and 12 months for the power-law and the Arps-with-5%-minimum-decline models, respectively. This is caused by the difference in biases of the assessments generated using 6 months of data measured at time of update *t* and the long-term biases of the assessments generated using all the production data available at the time of update *t*. When we applied periodic updating with external adjustment (Option 5, purple), we got the benefits of both using all the available production data (low AW) and externally adjusting assessments (low CS). It almost always resulted in the best (lowest) CS with a relatively low AW (but not the lowest; Option 3 has lower AW but only because it has higher CS). Ultimately, periodically updating the probabilistic model with new production data and external adjustment appears to be ideal; it will result in one of the best possible CS coupled with a relatively low AW.

In this case study, we did our analysis on a normalized-to-time-zero basis and thus we assumed that all of the 197 wells start producing at the same time and we measured their

24-month calibration at the same time. When we do our analysis in such a fashion, we may not be able to utilize all of the available production data at a particular time in the field history, since these wells have varying lengths of production data and we will be limited to the length of production data of the most recent well. It would not be wise to use only 6 months of production data for all wells if the most recent well has only 6 months of production data. One possible solution that should be investigated in future work is to divide the wells into groups that are not mutually exclusive, where the first group contains all the wells that have 6 months of production data or more, the second group contains all the wells that have 12 months of production data or more, and so on. Then external adjustment can be used in combination with aggregation methods such as those explained in Turner et al. (2014) to generate an aggregated probabilistic assessment for each of the wells in the field.

**Conclusions**

The reliability of probabilistic assessments in petroleum engineering can be detected and quantified by conducting look-backs (comparing actual performance to probabilistic forecasts) and constructing and analyzing calibration plots. Confidence and directional biases can be measured from calibration plots. However, the accuracy of these bias measurements is dependent on the number of probabilistic assessment/observation pairs available. In general, the more assessments available, the more accurate the measure of probabilistic reliability. However, even a low number of assessments (as low as 10) is enough to detect the existence and the direction of biases in cases of moderate levels of confidence and directional biases. An even lower number of assessments (2 or 3) is enough

to indicate the existence and direction of biases in cases of extreme levels of confidence and directional biases.

Measurement of confidence and directional biases (CB and DB) from calibration plots offers advantages over more traditional measures of probabilistic forecast reliability, such as coverage rate and calibration score. The coverage rate is insensitive to directional biases although it is sensitive to confidence biases. Coverage-rate values lower than the assumed prediction-interval width typically indicate overconfidence while values that are larger typically indicate underconfidence. Furthermore, while the calibration score is insensitive to the direction of biases (positive vs negative directional biases and overconfidence vs. underconfidence), it is sensitive to the magnitude of these biases. While lower calibration scores indicate lower overall levels of biases present in the probabilistic assessments, they are not helpful in providing feedback to assessors who want to internally adjust their assessment models. On the other hand, CB provides guidance on whether the assessment distributions should be wider or narrower and DB provides guidance on whether the assessments should be shifted positively or negatively.

Measuring the calibration of historical probabilistic assessments and using it to externally adjust new assessments reduces biases and improves calibration. A continuous process, at an appropriate frequency, of look-back, calibration, model update, and external adjustment will result, over the long run, in the best possible calibration while minimizing the average width of probabilistic assessments.

# CHAPTER IV

# CONCLUSIONS AND FUTURE WORK

## Conclusions

A new generalized framework for quantifying the value of reliable uncertainty assessment (or quantifying the cost of biased estimation) that allows full, non-truncated estimated distributions replicates well the results and conclusions from a previously presented simplified framework that used truncated estimated distributions. Moderate overconfidence and optimism can easily produce average portfolio disappointment (estimated value minus realized value) of 30-35% of estimated portfolio EV or more.

Extension of the new generalized framework to underconfidence demonstrates that underconfidence, in combination with directional bias, is similarly detrimental to portfolio performance as overconfidence. Thus, as operators seek to eliminate overconfidence bias, they should be wary of overcorrecting into underconfidence.

Gains from reducing uncertainty are small given moderate levels of confidence and directional biases. At higher levels of confidence and directional biases, reducing uncertainty will result in greater reduction in expected disappointment and increase in expected value attainment. However, these improvements are still less than the improvements that result from reducing biases. The lowest expected disappointment and the highest expected value attainment can be achieved only by eliminating biases.

The reliability of probabilistic assessments in petroleum engineering can be detected and quantified by conducting look-backs (comparing actual performance to probabilistic

forecasts) and constructing and analyzing calibration plots. Confidence and directional biases can be measured from calibration plots. However, the accuracy of these bias measurements is dependent on the number of probabilistic assessment/observation pairs available. In general, the more assessments available, the more accurate the measure of probabilistic reliability. However, even a low number of assessments (as low as 10) is enough to detect the existence and the direction of biases in cases of moderate levels of confidence and directional biases. An even lower number of assessments (2 or 3) is enough to indicate the existence and direction of biases in cases of extreme levels of confidence and directional biases. Armed with quantitative measurements of biases, operators can then make efforts to eliminate these biases in new forecasts through a combination of internal adjustment of uncertainty assessments—via assessment training and/or monitoring—and external adjustment of forecasts using measurements of biases from calibration.

Measurement of confidence and directional biases (CB and DB) from calibration plots offers advantages over more traditional measures of probabilistic forecast reliability, such as coverage rate and calibration score. The coverage rate is insensitive to directional biases although it is sensitive to confidence biases. Coverage-rate values lower than the assumed prediction-interval width typically indicate overconfidence while values that are larger typically indicate underconfidence. Furthermore, while the calibration score is insensitive to the direction of biases (positive vs negative directional biases and overconfidence vs. underconfidence), it is sensitive to the magnitude of these biases. While lower calibration scores indicate lower overall levels of biases present in the probabilistic assessments, they

are not helpful in providing feedback to assessors who want to internally adjust their assessment models. On the other hand, CB provides guidance on whether the assessment distributions should be wider or narrower and DB provides guidance on whether the assessments should be shifted positively or negatively.

Measuring the calibration of historical probabilistic assessments and using it to externally adjust new assessments reduces biases and improves calibration. A continuous process, at an appropriate frequency, of look-back, calibration, model update, and external adjustment will result, over the long run, in the best possible calibration while minimizing the average width of probabilistic assessments.

**Future Work**

In Chapter II, I introduced a method for calculating DB and CB from the least-squares best-fit line of the calibration curve. This method, while simple and fast, will often have some information loss because it is based on the best-fit line of a curve. Consequently, these simplified bias equations will not calculate the exact biases. While the differences between the measured and the actual biases are not significant in typical settings, it may be desirable to develop a bias-measurement method that provides more accurate measurements of biases using the entire calibration curve (all the proportion-correct values available) and not just the best-fit line of the calibration curve.

In the second subsection of the case study in Chapter III (using short-term assessments to externally adjust long-term assessments), the analysis was done on a normalized-to-time-zero basis and, thus, it was assumed that all of the 197 wells started producing at the same time. I did not investigate how to analyze wells when they start producing on

different dates and have varying lengths of historical production data. It would be valuable to determine how to update and externally adjust wells that start producing on different dates and have different lengths of production data given that the probabilistic forecasts for these wells have different levels of biases, as shown by Gonzalez et al. (2012). Furthermore, it would be beneficial to show the value added and the difference in results between performing the analysis using a normalized-to-time-zero basis versus performing it with the wells starting to produce on their respective dates.

# NOMENCLATURE

| | |
|---|---|
| $a$ | Lower bound of truncated distributions; or first intersection between two distributions; or intersection with the y-axis |
| $A_L$ | Area to the left |
| $A_{OVL}$ | The overlapping area between two distributions |
| $A_R$ | Area to the right |
| AW | Average width of the prediction interval |
| $b$ | Upper bound of truncated distributions; or second intersection between two distributions |
| BS | Brier score |
| CapEx | Capital expenditure, dollars spent or committed at the beginning of the project or portfolio |
| CB | Confidence bias |
| $CB_{OC}$ | The overconfidence portion of the confidence bias |
| $CB_{UC}$ | The underconfidence portion of the confidence bias |
| CDF | Cumulative density function |
| CI | Confidence interval |
| CPEOH | Cumulative production at the end of the hindcast period |
| CR | Coverage rate or the empirical coverage of the central prediction interval |
| CS | Calibration score |
| $c_t$ | The proportion correct of the $t$'th cumulative probability or subinterval |
| DB | Directional bias |

| | |
|---|---|
| $DB_{OC}$ | The directional bias value assuming overconfident estimated distribution |
| $DB_{UC}$ | The directional bias value assuming underconfident estimated distribution. |
| DCA | Decline curve analysis |
| E&P | Exploration & Production |
| ED | Expected disappointment, the average of disappointment values over a number of Monte-Carlo iterations. |
| ED%E | Expected disappointment as a percentage of estimated distribution |
| EDE | Expected decision error, the average of decision error values over a number of Monte-Carlo iterations. |
| EDE%E | Expected decision error as a percentage of estimated distribution. |
| EUR | Expected ultimate recovery |
| EV | Expected value |
| EVA%BP | Expected value attainment as a percentage of best possible portfolio value |
| $f_e$ | The PDF of the estimated distribution |
| $f_{e\text{-}reduced}$ | The PDF of the estimated distribution with reduced uncertainty |
| $f_{en}$ | The PDF of an estimated distribution with negative directional bias |
| $f_{ep}$ | The PDF of an estimated distribution with positive directional bias |
| $F_e$ | The CDF of the estimated distribution |
| $f_t$ | The PDF of the true distribution |
| $F$ | A cumulative distribution function |
| $F_{t\text{-}reduced}$ | The PDF of the true distribution with reduced uncertainty |
| $F_t$ | The CDF of the true distribution |

| | |
|---|---|
| $I$ | Indicator function |
| IE | Investment efficiency |
| $m$ | Slope of the best fit line |
| MCMC | Markov Chain Monte Carlo |
| $Mo_e$ | Mode of the estimated distribution |
| $Mo_t$ | Mode of the true distribution |
| $N$ | The total number of propositions |
| NPV | Net present value |
| $n_{Pt}$ | The number of propositions defined at the t'th cumulative probability or subinterval |
| OPB | Optimism-pessimism bias |
| $P$ | Cumulative probability assigned to the proposition |
| $P_t$ | The $t$'th cumulative probability or the average cumulative probability of the $t$'th subinterval |
| PDF | Probability density function |
| PVOCF | Present value of operating cash flow |
| SD | Standard deviation |
| $T$ | The number of defined cumulative probabilities or probability subintervals |
| $x$ | The value of the observed outcome of the quantity assessed |
| $x_P$ | The value which there is a P chance that the observed outcome will be less than or equal to $x_P$ |

# REFERENCES


Alarfaj, M.K. and McVay, D.A. 2018. Improved Framework for Measuring the Magnitude and Impact of Biases in Project Evaluation (Unpublished).

Arps, J.J. 1945. Analysis of Decline Curves. *AIME* **160**: 228-247. https://doi.org/10.2118/945228-G.

Begg, S.H. and Bratvold, R.B. 2008. Systematic Prediction Errors in Oil and Gas Project and Portfolio Selection. Presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 21–24 September. SPE-116525-MS. http://dx.doi.org/10.2118/116525-MS.

Begg, S.H., Welsh, M.B., and Bratvold, R.B. 2014. Uncertainty Vs. Variability: What's the Difference and Why Is It Important? Presented at the SPE Hydrocarbon Economics and Evaluation Symposium, Houston, 19-20 May. SPE-169850-MS. http://dx.doi.org/10.2118/169850-MS.

Bradley, E.L. 2006. Overlapping Coefficient. *In Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc.

Brashear, J.P., Becker, A.B., and Faulder, D.D. 2001. Where Have All the Profits Gone? *J Pet Technol* **53** (6): 20 - 73. SPE-73141-JPT. http://dx.doi.org/10.2118/73141-JPT.

Brier, G.W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* **78** (1): 1-3. https://doi.org/10.1175%2F1520-0493(1950)078%3C0001%3AVOFEIT%3E2.0.CO%3B2.

Capen, E.C. 1976. The Difficulty of Assessing Uncertainty. *J Pet Technol* **28** (8): 843-850. SPE-5579. http://dx.doi.org/10.2118/5579-PA.

Dossary, M. 2016. Automatic Calibration of Uncertainty Estimates Using Anchoring Heuristics. Presented at the SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition, Dammam, Saudi Arabia, 25-28 April. SPE-182797-MS. https://doi.org/10.2118/182797-MS.

Fondren, M.E., McVay, D.A., and Gonzalez, R.A. 2013. Applying Calibration to Improve Uncertainty Assessment. Presented at the, 30 September-2 October. SPE-166422-MS. http://dx.doi.org/10.2118/166422-MS.

Gong, X., Gonzalez, R., McVay, D.A. et al. 2014. Bayesian Probabilistic Decline-Curve Analysis Reliably Quantifies Uncertainty in Shale-Well-Production Forecasts. *SPE J* **19** (06): 1047-1057. SPE-147588-PA. https://doi.org/10.2118/147588-PA.

Gonzalez, R., Gong, X., and McVay, D.A. 2013. Prior Information Enhances Uncertainty Quantification in Shale Gas Decline Curve Forecasts. Presented at the SPE Unconventional Resources Conference-Canada, Calgary, Alberta, Canada, 5-7 November. SPE-167150-MS. https://doi.org/10.2118/167150-MS.

Gonzalez, R.A., Gong, X., and McVay, D.A. 2012. Probabilistic Decline Curve Analysis Reliably Quantifies Uncertainty in Shale Gas Reserves Regardless of Stage of Depletion. Presented at the SPE Eastern Regional Meeting, Lexington, Kentucky, USA, 3-5 October. SPE-161300-MS. https://doi.org/10.2118/161300-MS.

Haynes and Boone. 2018. *Haynes and Boone Oil Patch Bankruptcy Monitor*. http://www.haynesboone.com/~/media/files/energy_bankruptcy_reports/2018/oil_patch_bankruptcy_monitor_03312018.ashx (accessed June 13, 2018).

Hdadou, H. and McVay, D.A. 2014. The Value of Assessing Uncertainty in Oil and Gas Portfolio Optimization. Presented at the SPE Hydrocarbon Economics and Evaluation Symposium, Houston, Texas, 19-20 May. SPE-169836-MS. http://dx.doi.org/10.2118/169836-MS.

Hubbard, D.W. 2014. *How to Measure Anything: Finding the Value of Intangibles in Business*, 3rd edition. Hoboken, New Jersey: Wiley.

Ilk, D., Rushing, J.A., Perego, A.D. et al. 2008. Exponential Vs. Hyperbolic Decline in Tight Gas Sands: Understanding the Origin and Implications for Reserve Estimates Using Arps' Decline Curves. Presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 21-24 September. SPE-116731-MS. https://doi.org/10.2118/116731-MS.

Landman, W.A. and Goddard, L. 2002. Statistical Recalibration of Gcm Forecasts over Southern Africa Using Model Output Statistics. *Journal of Climate* **15** (15): 2038-2055. http://journals.ametsoc.org/doi/abs/10.1175/1520-0442%282002%29015%3C2038%3ASROGFO%3E2.0.CO%3B2.

Lichtenstein, S. and Fischhoff, B. 1977. Do Those Who Know More Also Know More About How Much They Know? *Organizational Behavior and Human Performance* **20** (2): 159-183. http://www.sciencedirect.com/science/article/pii/0030507377900010.

Lichtenstein, S., Fischhoff, B., and Phillips, L.D. 1977. Calibration of Probabilities: The State of the Art. *In Decision Making and Change in Human Affairs*, ed. H. Jungermann and G De Zeeuw, 275-324. Netherlands, Springer.

Mandel, D.R. and Barnes, A. 2014. Accuracy of Forecasts in Strategic Intelligence. *Proceedings of the National Academy of Sciences* **111** (30): 10984-10989. http://www.pnas.org/content/111/30/10984.abstract.

McVay, D.A. and Dossary, M.N. 2014. The Value of Assessing Uncertainty. *SPE Econ & Mgmt* **6** (2): 100 - 110. SPE-160189-PA. http://dx.doi.org/10.2118/160189-PA.

McVay, D.A., Lee, W.J., and Alvarado, M.G. 2005. Calibration Improves Uncertainty Quantification in Production Forecasting. *Petroleum Geoscience* **11** (3): 195-202. http://pg.lyellcollection.org/content/11/3/195.abstract.

Merrow, E.W. 2012. Oil and Gas Industry Megaprojects: Our Recent Track Record. *Oil and Gas Fac* **1** (2): 38 - 42. SPE-153695. http://dx.doi.org/10.2118/153695-PA.

Murphy, A.H. 1973. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology* **12** (4): 595-600. http://journals.ametsoc.org/doi/abs/10.1175/1520-0450%281973%29012%3C0595%3AANVPOT%3E2.0.CO%3B2.

Nandurdikar, N. 2014. Wanted: A New Type of Business Leader to Fix E&P Asset Developments. *J Pet Technol* **66** (10): 15-19. http://dx.doi.org/10.2118/1014-0015-JPT.

Piani, C., Haerter, J.O., and Coppola, E. 2010. Statistical Bias Correction for Daily Precipitation in Regional Climate Models over Europe. *Theoretical and Applied*

*Climatology* **99** (1): 187-192. Piani2010. http://dx.doi.org/10.1007/s00704-009-0134-9.

Rose, P.R. 2004. Delivering on Our E&P Promises. *Leading edge* **23** (2): 165. http://dx.doi.org/10.1190/1.1651465.

Smith, J.E. and Winkler, R.L. 2006. The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis. *Management Science* **52** (3): 311-322. http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1050.0451.

Turner, B.M., Steyvers, M., Merkle, E.C. et al. 2014. Forecast Aggregation Via Recalibration. *Machine Learning* **95** (3): 261-289. http://dx.doi.org/10.1007/s10994-013-5401-4.

Welsh, M.B., Begg, S.H., and Bratvold, R.B. 2007. Modelling the Economic Impact of Common Biases on Oil and Gas Decisions. Presented at the SPE Annual Technical Conference and Exhibition, Anaheim, California, 11-14 November. SPE-110765-MS. http://dx.doi.org/10.2118/110765-MS.

Xu, C. and Bell, L., eds. 2016. Big Losses Reported in 4q 2015 on Writedowns, Low Oil Prices. Vol. **114.4a**, *Oil & Gas Journal*.