# *From* Guesswork *to* Informed Judgment

## Using Calibration Training for Better Decision-Making

BY SHAYNE KAVANAGH, DOUG HUBBARD, PHILIP MARTIN, AND ROBERT WEANT

©2023 AAD GOUDAPPEL C/O THEISPOT.COM

The job of a local government finance officer is to help elected officials and other decision-makers make better decisions. This involves helping elected officials make sense of uncertain information and situations—which requires us to express uncertainty clearly, especially for assumptions that might underlie financial plans and budgets.

To help, GFOA worked with Hubbard Decision Research (HDR) and 40 GFOA member volunteers to go through what is known as "calibration training." The objective of calibration training is to prepare people to express their estimates of uncertainty in quantitative terms, using only their own judgment and without relying on outside data. Quantified estimates of uncertainty are desirable, and special training is necessary to make them.

## WHY DO WE NEED TRAINING TO QUANTIFY UNCERTAINTY?

Let's start with why quantified estimates of uncertainty are desirable. Uncertainty is a key element of risk, which could be defined as the downside of uncertainty. Finance officers help decision-makers manage the financial risk of their decisions. But as Peter Bernstein put it in *Against the Gods: The Remarkable Story of Risk*,[1] "Without numbers, there are no odds and no probabilities; without odds and probabilities, the only way to deal with risk is to appeal to the gods and the fates. Without numbers, risk is wholly a matter of gut."

Bernstein is advocating for quantifying uncertainty (and, thereby, risk) by using odds and probabilities. He cautions against our gut instincts because people are not wired to think about risk correctly. One reason for that is what psychologists call the "overconfidence bias." This means that we are overconfident in our predictions and tend to underestimate uncertainty. We can illustrate this with data from a calibration training experiment GFOA conducted (see Exhibit 2).

Exhibit 2 provides a comparison of how well the group of GFOA volunteers thought they did on making a prediction versus how well they actually did, before they went through calibration training. For instance, if we look at the horizontal axis and find the 0.8 mark, this is the point where someone believes their forecast has an 80 percent chance of being correct. We can then look upward until we intersect the 0.8 mark on the vertical axis, which is on the green line. This is the point of perfect calibration: if you say your forecasts have an 80 percent chance of being correct, and your forecasts are, in fact, correct 80 percent of the time. Any point on the green line is perfect calibration.

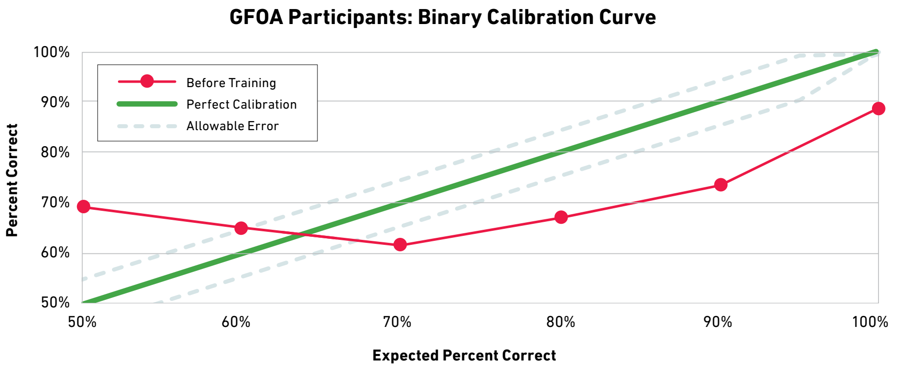
We also see a red line on Exhibit 2; this is data from participants before they took the calibration training. If we revisit 0.8 on the horizontal axis, we see the red line meets the vertical axis at 0.68. In other words, the uncalibrated person believed their forecasts had an 80 percent chance of being correct but are actually correct only 68 percent of the time; this is overconfidence bias. Anytime the red line is below the green line, it represents overconfidence. The dotted gray lines represent allowable error in our measurement of the participants in the training—we can't expect our tests to be a perfect measure of accuracy.

With this in mind, we can see that the participants are not overconfident when assessing their chances at 50 or 60 percent. This is because the participants are highly uncertain and are, in essence, flipping a coin when estimating if they will be correct or not. In fact, because the red line is above the green line, our participants were underconfident: they thought they were flipping a coin; but,

EXHIBIT 1 | UNQUANTIFIED VERSUS QUANTIFIED ESTIMATES

| **Unquantified estimate** | **Quantified estimate** |
|---|---|
| Next year's total revenue **will be somewhere around** $20 million. | There is a **90 percent chance** that next year's revenue **will be between** $19 million and $21 million. |
| I think the new sales tax law **will likely** pass by the start of our next fiscal year. | I give an **80 percent chance** that the new sales tax law will pass by the start of our next fiscal year. |

EXHIBIT 2 | RESULTS OF "PRETEST" OF CALIBRATION TRAINING PARTICIPANTS



©2023 AAD GOUDAPPEL C/O THEISPOT.COM

in fact, they had better information about the correct answer than they thought they did. But as soon as the participants believed their estimate has a reasonable chance of being right, overconfidence crept in.

Mitigating or eliminating overconfidence bias is why calibration training is necessary. Quantifying your estimates of uncertainty is not simply a matter of replacing words such as "likely" or "probable" with numbers like "80 percent" or "4 in 5 chance." We need to take steps to make the numbers as accurate as possible. Ideally, we would use data to come up with our estimates of uncertainty. For example, if your annual hotel tax forecast has been within +/– 5 percent of actual revenue for each of the last ten years, then it is a good bet that next year's forecast will be within that range as well (assuming your forecast method remains substantively similar). But the data for estimating uncertainty is often unavailable or incomplete, and there isn't the time and/or resources to get more data. In this case, the estimator will be forced to rely on their judgment, at least in part.

Calibration training moves the participants closer to the green line. Exhibit 3 shows results from training with an added blue line. We can see that, at all points, the blue line is closer to the green line and is even within the gray lines at a few points, which indicates perfect calibration.

## HOW DOES CALIBRATION TRAINING WORK?

Calibration training helps people recognize overconfidence bias and mitigate it. It does this by providing immediate feedback on the quality of a high volume of predictions made by the participant. In other words, the participants make a lot of predictions and then find out how well they did right away. Rapid feedback is essential to learning.

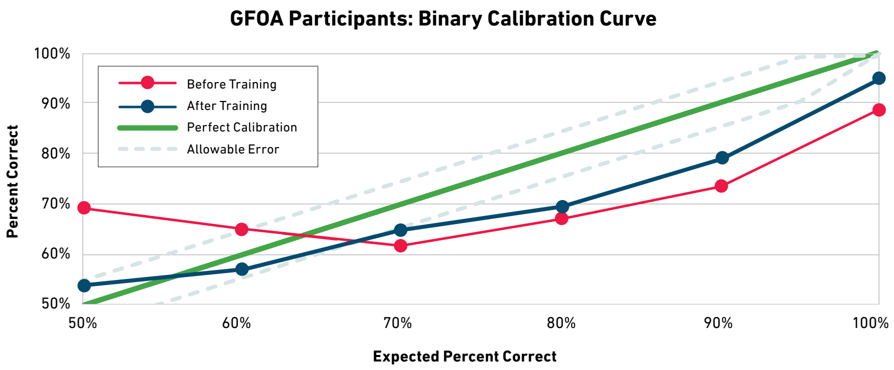
The predictions that participants make come in the form of difficult "trivia questions," where it is highly unlikely that the participants know the right answer (without looking it up).

These questions were asked in two formats. One format was true/false, where participants picked either true or false as their answer and then assigned a level of confidence to their answer. For example, a question might be: "True or false? A gallon of oil weighs less than a gallon of water." Though you may not know the exact weight of either, you may recall that oil floats on water; therefore, oil probably weighs less. For that reason, you predict the statement is true and give your answer a high level of confidence (90 or even 100 percent).

Other questions might be harder. For example, "Mars is always farther away from Earth than Venus." Imagine you have no idea, so you guess true....but you rate your confidence at 50 percent, which indicates you basically flipped a coin to get your answer. Most questions are somewhere in between, where the participant has some idea of the correct answer but is not highly certain. Participants usually overstate their confidence at the start of the training and learn to be more circumspect about their predictions as the training progresses. By answering many questions and getting feedback, participants learn to more accurately assess the chance that they are correct. This prepares them to develop real-life predictions, like whether some important event will come to pass (for example, whether a major retailer in the community will close in the next year).

After one half-day of calibration training, participants get much closer to 90 percent correct for their range questions. This prepares them to develop ranges for situations they might encounter in their work life, such as a range of potential yields for a revenue source.

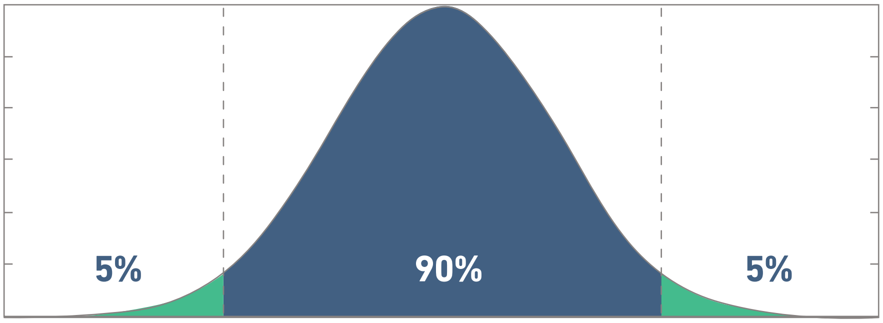
EXHIBIT 3 | RESULTS OF "POST-TEST" OF CALIBRATION TRAINING PARTICIPANTS



In the other format of the trivia questions, participants are asked to define a range where they are 90 percent sure the correct answer lies within the range. For example, "What is the wingspan, in feet, of a Boeing 737 aircraft (the type of plane used by Southwest Airlines)?" Unless you are an aeronautical engineer, you probably don't know the exact answer and would need to make a projection. You are not making a wild guess because you have information to go on. You may have been a passenger on such a plane or have seen it in pictures, so you have an idea of how big it is. However, it will still be a highly uncertain forecast for you. Picking a single number, like "75 feet" is of limited use because it doesn't express your uncertainty and the degree of risk that you might be wrong, so the calibration training asks participants to express their forecasts as a 90 percent confidence range. Put another way, they pick a high and a low value where they are 90 percent confident that the value will be in between. For example, they might say, "I'm 90 percent confident that the wingspan is between 65 and 100 feet."

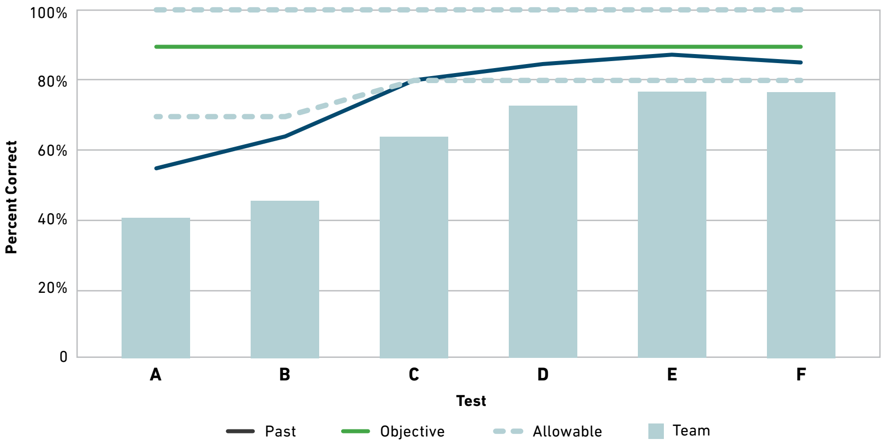
As it turns out, the actual wingspan is 113 feet. Making the range too narrow is a common problem for participants; this is overconfidence bias in action. When participants first try forecasting 90 percent confidence ranges for a large number of trivia questions, the correct answer falls outside of their ranges for around half of their estimates. If they are truly "90 percent confident," then the correct answer should fall outside of

EXHIBIT 4 | 90 PERCENT CONFIDENCE INTERVAL SHOWN AS A BELL CURVE



Rather than estimating the blue portion, many people find it easier to estimate each green portion separately.

EXHIBIT 5 | PERCENT CORRECT OVER SUCCESSIVE PRACTICE ROUNDS



their ranges in only 10 percent of their forecasts (as in, 90 percent of the forecasts should be correct). After one half-day of calibration training, participants get much closer to 90 percent correct for their range questions. This prepares them to develop ranges for situations they might encounter in their work life, such as a range of potential yields for a revenue source.

Participants do not learn purely through trial and error. They also learn techniques to help them make better predictions. For example, they learn to estimate each side of their ranges separately. When people think of a "90 percent confidence range," they tend to generate both ends of the range simultaneously. Considering each side separately, however, causes people to slow down and give more consideration to their estimate (see Exhibit 4). Looking at one side of the range at a time reframes the questions from "90 percent of the time the correct answer will be in this range" to: 1) "Only five percent of the time the correct answer will be higher than my upper value"; and 2) "Only five percent of the time will the correct answer be lower than my low value." (The five percent on either side of the range adds up to 10 percent—a 100 percent minus 10 percent is 90 percent for your 90 percent confidence interval.) Five percent is only 1 out of 20! When participants consider that the correct answer can only go beyond a given side of their interval 1 out of 20 times, it causes them to adjust that side of their interval.

In Exhibit 5, we can see how participants improved as they completed multiple rounds of feedback and incorporated the ideas from the training in their estimates. The graph focuses on the range questions, and the "percent correct" means the proportion of times the value fell within the 90 percent confidence interval that the participant defined. By the sixth round, the participants were quite close to the gray line, which is the allowable error.

Finally, some readers may wonder why the "objective" in Exhibit 5 (the green line) is only 90 percent correct instead of 100 percent correct. This is because scoring 100 percent on the test is easy—just make all your ranges infinitely wide. For example, you might set your range for the wingspan of a 737 aircraft between 1 and 1 million feet. You are guaranteed to be "correct." However, such wide ranges do not have much, if any, practical use in real-life estimation problems. The training aims to create balance, with ranges narrow enough to focus in on the 90 percent most likely possible values but still wide enough to not fall victim to overconfidence bias.

## APPLICATIONS OF CALIBRATION TRAINING

In this section, we'll discuss three possible applications of calibration training for the work of the finance officer. We invite you to also consider other applications.

**First, it equips finance officers to communicate uncertainty in their forecasts and estimates.** Research shows that people prefer advisors who quantify their uncertainty but are still confident.[2] To illustrate, a revenue forecast might be presented in these ways:

- "I'm 75 percent certain that revenues will increase by at least one percent next year."
- "I'm 90 percent certain that revenues will be between $50 million and $55 million next year."

Our examples omit hedging language like "maybe," "I'm not sure, but...," and so on. The statements come across as confident, even though expressed as a probabilistic likelihood.

We can also use the second bullet to illustrate overconfidence bias. Overconfidence causes people to make their ranges too narrow, often by around 50 percent.[3] Revisit the second example and imagine that the person who made that estimate has calibrated. In contrast,

### Pro Tip For Communicating Uncertainties

Probabilities like 90 percent are an abstract concept and can be difficult to grasp for people who aren't accustomed to thinking in probabilities. Research suggests that ratios are easier for people to grasp.[6] For example, rather than saying, "I'm 90 percent certain revenues will be...," one could say, **"There is a 9 in 10 chance revenues will be...."**

an overconfident estimator might say, "I'm 90 percent certain that revenues will be between $51 million and $54 million." This is a range of only $3 million, compared to the $5 million range given by the calibrated estimator. In other words, the overconfident estimator has left less room for error and increased the chance of a bad decision. Calibration training helps you reach intervals that are of the right size, given your degree of certainty for the question at hand.

For many applications in public finance, the government needs to settle on a single number. For example, you can't put a range of possible revenues in your budget—but you can use the range to assess the risk you're taking on by adopting any given number for your budget. The GFOA book *Informed Decision-Making through Forecasting: A Practitioner's Guide*,[4] provides several case examples of how local governments have applied ranges when presenting forecasts. To summarize one application, the range of possible future revenues could be presented to the governing board, with the midpoint of the range considered the single most likely outcome. If the board were to adopt an expenditure budget equal to this midpoint, they give themselves 1:1 odds (or a 1 in 2 chance) of a deficit. This is because there is a 50 percent chance of revenues being less than (or more than) the midpoint of the estimate.

Many elected boards wouldn't like those odds and would prefer to give themselves a better chance of avoiding a deficit at the end of the year. In one of our case studies, the finance officer facilitated a conversation about the odds the elected board would prefer. The board landed on a spending plan that provided 2-1 odds (or a 2 in 3 chance) of producing a surplus, which was more aligned with the board's goal of building up reserves after a recent natural disaster. So, they landed on a single number for the budget but were also fully aware of the degree of risk that number represented. After a few years, when the reserves had been replenished, the board opted for lower odds of budgetary surplus (though still better than 1 in 2) by picking a higher number for the expenditure budget.

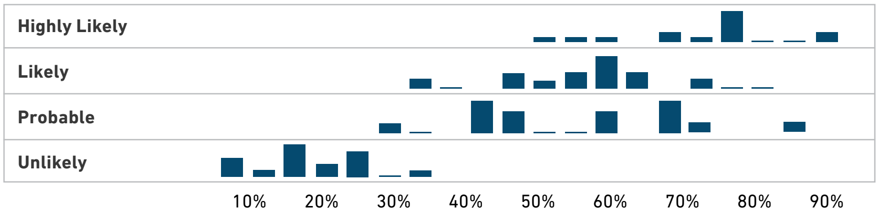
You can download a spreadsheet to help you with the process described in this paragraph as part of an online article published by GFOA titled "Silicon Valley Bank and Stress Tests: What Can Local Governments Learn?"[5] After looking at the spreadsheet, you will probably be able to think of other applications for inserting subjective estimates of probabilities into larger quantitative models.

**The second practical application of calibration training is that it helps the finance officer avoid the "illusion of communication" when discussing risks.** The illusion of communication is when people use vague terms to communicate uncertainty, such as "likely," "probably," and more. The problem is that people often have very different numbers in mind when asked to quantify these terms.

To illustrate, one study of military intelligence examined a set of standard terms that intelligence officers were supposed to use to describe their certainty of military threats.[7] The order from most to least certain was: 1) "highly likely," 2) "likely," 3) "probable," and 4) "unlikely." When the intelligence officers were asked to quantify their uncertainty, the results exposed these standardized terms as worthless. For example, it turned out that some intelligence officers thought "highly likely" meant anything with more than a 50 percent chance of occurring, while others thought it meant a chance of 90 percent or more. "Likely" and "probable" meant essentially the same thing, given the quantities the officers assigned. And the range variation in the quantitative definitions the officers came up with resulted in some officers using "probable" (the term intended to describe the second lowest chance) to describe the same percent chance that other officers described as "highly likely" (the term for the highest chance). You can see this illustrated in Exhibit 6.

**EXHIBIT 6 | FREQUENCY WITH WHICH MILITARY INTELLIGENCE OFFICERS ASSIGNED QUANTITATIVE VALUES TO SUBJECTIVE DESCRIPTIONS OF RISKS**

The blue bars represent how often a given quantitative measure (such as, 20%) was believed to describe what a given subjective term (such as "unlikely" meant). The higher the bar, the more frequent the belief.



©2023 AAD GOUDAPPEL C/O THEISPOT.COM

**EXHIBIT 7 | A CONVENTIONAL RISK MATRIX**

| RISK PROBABILITY | RISK SEVERITY | | | | |
|---|---|---|---|---|---|
| | Catastrophic **A** | Hazardous **B** | Major **C** | Minor **D** | Negligible **E** |
| Frequent **5** | 5A | 5B | 5C | 5D | 5E |
| Occasional **4** | 4A | 4B | 4C | 4D | 4E |
| Remote **3** | 3A | 3B | 3C | 3D | 3E |
| Improbable **2** | 2A | 2B | 2C | 2D | 2E |
| Extremely Improbable **1** | 1A | 1B | 1C | 1D | 1E |

The illusion of communication is not limited to military intelligence. Many local governments have probably fallen into the same trap by using the popular "heat map" style of risk matrix, as shown in Exhibit 7. It is easy to imagine the same problem with the terms on either axis of Exhibit 7. This is not just a theoretical problem. Research into how these kinds of risk matrices affect real-life decisions has concluded that "[qualitative] risk matrices should not be used for decisions of any consequence."[8]

Calibration training prepares the finance officer to replace vague terms with meaningful, quantified, and calibrated estimates of uncertainty. For example, you could imagine a version of Exhibit 7 where the two axes are replaced by ranges of percent likelihood that the risk will occur and a potential dollar loss on the other. There are also other presentation methods that quantify risks into probabilities as a fundamental design feature. One example is called a "loss exceedance curve." You can see examples applied to cyber security risk in the GFOA report Cyber Risk Savvy.[9]

**The third practical application of calibration training is that it hones our intuitions about chance and prepares us to think probabilistically.** According to The Great Mental Models project:[10] "Successfully thinking in shades of probability means roughly identifying what matters, coming up with a sense of the odds, doing a check on our assumptions, and then making a decision. We can act with a higher level of certainty in complex, unpredictable situations. We can never know the future with exact precision. Probabilistic thinking is an extremely useful tool to evaluate how the world will most likely look so that we can effectively strategize."

The benefits of probabilistic thinking are measurable. For example, one study showed that exposing participants to probabilistic training was associated with as much as a 50 percent increase in the accuracy of their predictions.[11] Having a general idea how a revenue might perform is one thing. However, being able to examine the question of revenue yield from multiple vantage points and think about how new information might cause you to adjust your probabilistic expectations can be much more powerful.

## CONCLUSION

Public finance officers deal with uncertainty and risk across many domains of public finance, including forecasts, rainy day funds, insurance, and more. The ability to think probabilistically enhances a finance officer's ability to make savvier decisions about risk and uncertainty. Calibration training hones probabilistic thinking skills. Even better, it provides a basis for communicating quantified probability estimates to others. These skills are essential for helping the finance officer guide their local government in testing assumptions in an uncertain world. This allows the finance officer to be a bona fide decision leader—someone who doesn't only make good decisions themselves but can also help others make wise decisions.[12]

***Shayne Kavanagh*** *is senior manager of research in GFOA's Research and Consulting Center.* ***Doug Hubbard*** *is chief executive officer of Hubbard Decision Research.* ***Philip Martin*** *is chief operating officer of Hubbard Decision Research.* ***Robert Weant*** *is senior analyst at Hubbard Decision Research.*

The Decision Architect
How budget officers can reduce the impact of bias and noise for better decision making
BY JASON RIIS AND JARED PETERSON

GFOA's "Budget Officer as Decision Architect" (Jason Riis and Jared Peterson, GFOA, February 2023) offers a complete discussion of the finance officer's role in helping their governments make better decisions. The ideas in this article support the skills needed to be a decision architect.

**READ THE ARTICLE:**

**gfoa.org/materials/gfr0223-decision-architect**

[1] Peter L. Bernstein, *Against the Gods: The Remarkable Story of Risk...* (Wiley: 1998).

[2] Celia Gaertig and Joseph P. Simmons, "Do people inherently dislike uncertain advice?" forthcoming, *Psychological Science*, 2017 (available at SSRN: ssrn.com/abstract=3041566).

[3] Jack B. Soll, and Joshua Klayman, "Overconfidence in interval estimates," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 2004.

[4] Shayne Kavanagh and Daniel W. Williams, *Informed Decision-Making through Forecasting: A Practitioner's Guide*, GFOA, January 2017.

[5] Shayne Savage, Sam Savage, and Matthew Raphaelson, "Silicon Valley Bank and Stress Tests: What Can Local Governments Learn," GFOA (gfoa.org/silicon-valley-bank-and-stress-tests-what-can-local-governments-learn).

[6] Chip Heath and Karla Starr, *Making Numbers Count: The Art and Science of Communicating Numbers*, (Avid Reader Press/Simon & Schuster: 2022).

[7] Richards J. Heuer, *The Psychology of Intelligence Analysis*, Center for the Study of Intelligence, CIA, 1999.

[8] Philip Thomas, Reidar Bratvold, and J. Eric Bickel, "The Risk of Using Risk Matrices," SPE Economics and Management, 6(02), 2013.

[9] Shayne Kavanagh, Rob Roque, and Teri Takai, "Cyber Risk Savvy," GFOA, January 2022.

[10] *The Value of Probabilistic Thinking: Spies, Crime, and Lightning Strikes*, Mental Modes, fs blog, Farnam Street Media.

[11] In *Smarter Faster Better: The Transformative Power of Real Productivity* (Random House Publishing: 2016), author Charles Duhigg discusses the results obtained by the "Good Judgment Project" in their forecast experiments.

[12] Decision leader is a term from: Don A. Moore and Max H. Bazerman, *Decision leadership: empowering others to make better choices* (Yale University Press: 2022.