Hubbard Decision Research

CALIBRATED PROBABILITY ASSESSMENTS

An Introduction

Measure What Matters — Make Better Decisions

Introduction

Calibrated probability assessments are subjective assessments of probabilities or confidence intervals that have come from individuals who have been trained specifically to minimize certain biases in probability assessment and whose skill in such assessments has been measured.

Although subjective probabilities are relied upon for estimates of key variables in major business decisions, research shows that the subject matter experts who provide these estimates are consistently overconfident.

Fortunately, it can be shown that training significantly improves the ability of experts to quantify their own uncertainty about these estimated values.

BACKGROUND

Stochastic models of business decisions are used to assess the uncertainty and risk around that investment. Often, these models take the form of a Monte Carlo simulation where thousands of randomly generated scenarios may, for example, be used to generate the likelihood of returns on a given investment.

Subjective probability assessments by identified experts are widely used in stochastic modeling methods to assess uncertainties and risks about major business decisions.

A Key Risk in Major Business Decisions

1. Many of the most sensitive variables in models of business decisions rely on subjective estimates of experts.

2. Experts systemically underestimate their uncertainty about these variables – resulting in consistently understated risk.

3. Training experts to provide "calibrated probability assessments" is known to be effective – but it is rarely used.

Surveys show that subjective probability estimates are present in **over 80%** of stochastic models used in business decisions and the variables in nearly half of these models are mostly subjective estimates (Hubbard 2009).

There are two general types of estimates experts may be asked to provide for stochastic models:

1. An expert may be asked to provide a subjective confidence interval for some continuous quantity. For example, they may say that the cost of a project has a 90% chance of being between $2 million and $2.9 million.

2. An expert may be asked to provide a probability that some discrete event will occur. For example, an expert might say that there is a 75% chance that a competitor will go out of business.

But, unless certain controls are employed, most people will not provide probabilities that realistically represent their uncertainty about an estimate (Kahneman et. al, 1972, 1973, 1982; Lichtenstein et. al. 1982, Hubbard 2009).

For example, a project manager may state that there is a 90% chance of a project finishing on time. After a large number of the project forecasts where this project manager claimed 90% confidence were observed, it would probably turn out to be correct at a much less than 90% of the time. Of all the times an expert in any field says they are 90% confident that a cost estimate will fall within a particular range, something less than 90% of actual outcomes will turn out to be within their stated ranges.

Consistent over-estimation of the chance that a claim is (or will turn out to be) correct is called overconfidence. In some cases - although much more rarely - individuals may be under-confident. Experts in most every field that have been measured are overwhelmingly overconfident (Russo 1989). Since subjective estimates are common in most stochastic models of business decisions, the models will consistently underestimate risk – and this comes with a real and calculable cost.

"CALIBRATING" THE EXPERTS

If an individual demonstrates that they are neither overconfident nor under-confident for a large number of subjective estimates, they are said to be calibrated regarding subjective probability assessments. Such an individual will be right 80% of the time they said they were 80% confident, 90% of the time they are 90% confident, and so on. Certain methods do improve subjective estimates and can make estimators nearly perfectly calibrated.

The methods involve training with iterative feedback, training specific strategies, incentive systems, and aggregation of the estimates of groups of individuals. Several of these methods used in combination have shown to be effective in calibrating the majority of experts (Hubbard 2009). Using a battery of calibration tests based on trivia questions and training with equivalent bets and other strategies, a majority of typical managers reach calibration in less than one day.

Findings about calibration training include:

- Training may simply involve a series several trivia tests where estimators are asked to provide probabilities that some statement is true or to provide a range with a stated confidence (e.g. 90%). After each test, subjects are shown their results and, at least initially, will discover that they are overconfident or (less likely) under-confident. Subjects then attempt to improve their performance on subsequent tests by reducing their confidence (if they were initially overconfident) or increasing their confidence as needed (Hubbard 2009; Lichtenstein et. al. 1982). Feedback training alone may take a very long period of time or may have limited effectiveness.

- Other training involves teaching specific strategies to overcome confidence biases. It can be shown that habitually thinking of pros and cons for each estimate (Lichtenstein 1982) can slightly improve calibration. Also, subjects may learn to treat each estimate as a kind of bet. The "equivalent bet" or "equivalent urn" method involves identifying a bet with a stated chance of winning and comparing it to a bet on the estimate. If a subject's 90% confidence interval is realistic, they should be indifferent between a bet that their interval contains the correct value and spinning a dial that has a given 90% chance of winning.

- Anchoring is a significant effect on calibration and causes ranges to be narrower than they otherwise would be (Kahneman 1982). Anchoring is a phenomenon where prior estimates or even arbitrarily chosen numbers influence subsequent estimates. Since anchoring is a more significant issue for estimating ranges than binary probabilities, one strategy for avoiding anchoring includes estimating ranges by starting with apparently absurdly high and low values and asking "What is the probability the quantity is more (less) than X?". After answering several such questions for various values of X, the range can be inferred.

- Incentives based on "proper" scoring methods are also effective in calibrating estimators. A proper score is one which is impossible to game by the estimators. The only way for them to maximize their expected score with a series of predictions is to give the honest best estimate on each estimate. A person who flipped a coin and said they were 50% confident in each true/false prediction or a person who gives absurdly wide 90% confidence intervals 90% of the time (and deliberately misses 10%) would not score as well as a person given the best estimate or each. A Brier score is an example of a proper score and it has been shown to be an effective way to calibrate weather forecasters (Brier 1950).

- Another calibration method is simply averaging groups of individual estimates. This tends to be an improvement, but, since most individuals are overconfident, averaging the estimates of a group will not compensate for this. Averaging several estimates together may simply make the estimates more consistent (Clemen, Winkler 1986).

- Another method for aggregating the subjective estimates of a group of estimators is the use of prediction markets. In prediction markets participants buy and sell coupons on future events. A common type of coupon would pay a fixed amount if the event comes true. If the event does not come true, then the coupon is worth nothing. The coupon is bought and sold in a market where participants attempt to price the coupon in a way that reflects its expected future value. The calibration effect from using prediction markets provides fairly well-calibrated estimates and this effect is only slightly diminished when participants don't use real money (Servan-Schreiber 2004; Hubbard 2009).

THE EFFECT OF CALIBRATION TRAINING

The literature cited previously shows that training which combines several of these strategies has a significant effect on the ability of experts to realistically assess their own uncertainty using probabilistic statements. Hubbard Decision Research finds that about 3 hours of training is sufficient to train most individuals. Calibration training of similar duration has been tested by a variety of researchers.

Figures 1 and 2 below show the combined findings of multiple studies in the effects of calibration training. Figure 1 shows the results of training on the assessment of probabilities of discrete events (e.g., the chance a project will finish on time). Figure 2 shows the results of calibration training for interval estimates of continuous quantities (e.g., the first-year revenue of a new product).

Figure 1: The Combined Results of 11 Studies in Probability "Calibration" Training

Most experts significantly overstate their confidence in forecasts. Calibration training corrects this.

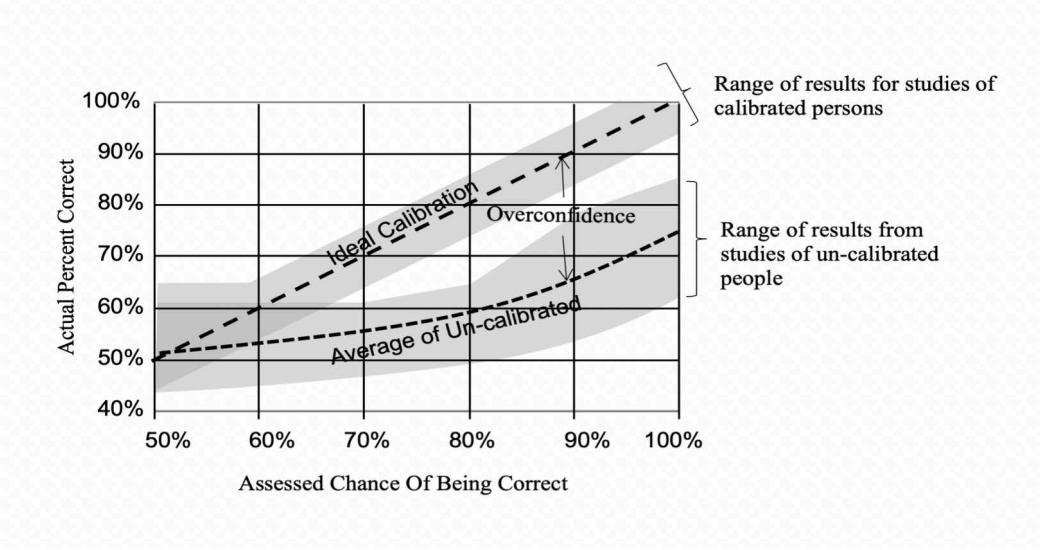

Figure 2: The Combined Results of Calibration Studies for Estimating Confidence Intervals

Most experts provide intervals that are far too narrow for the stated confidence but this is also corrected by calibration training.

Group	Subject	% Correct (target 90%)
Harvard MBAs	General Trivia	40%
Chemical Co. Employees	General Industry	50%
Chemical Co. Employees	Company-Specific	48%
Computer Co. Managers	General Business	17%
Computer Co. Managers	Company-Specific	36%
HDR Seminar (before training)	General Trivia & IT	35%-50%
HDR Seminar (after training)	General Trivia & IT	~90%

The overwhelming findings are:

1. Almost all experts are consistently overconfident in their estimates prior to training and

2. Most experts are able to reach nearly ideal calibration (within statistically allowable error) after training.

Given the size and risk of decisions that rely on subjective estimates, the consistent overconfidence of experts who provide those estimates, and the relative ease of calibration training, calibration training will be one of the most critical improvements in the decision processes of most organizations.

References:

1. Hubbard, D. The Failure of Risk Management: Why It's Broken and How to Fix It, Wiley, 2009, pp 237-8
2. Kahneman, D., Slovic, P., and Tversky, A., Judgement under Uncertainty: Heuristics and Biases, Cambridge University Press, Cambridge, 1982.
3. Kahneman, D. and Tversky A, "Subjective Probability: A Judgement of Representativeness," Cognitive Psychology 4, 430-454 (1972).
4. Kahneman, D. and Tversky A, "On the Psychology of Prediction," Psychological Review 80, 237-251 (1973).
5. Lichtenstein S., Fischhoff B., and Phillips L.D., "Calibration of Probabilities: The State of the Art to 1980," in Judgement under Uncertainty: Heuristics and Biases, eds. D. Kahneman, P. Slovic, and A. Tversky, Cambridge University Press, Cambridge, 1982, pp. 306-334, .
6. Servan-Schreiber, E. Wolfers, J, Pennock D.M., and Galebach, B, "Prediction Markets: Does Money Matter?" Electronic Markets 14, No. 3, 243 - 251 (2004).
7. Brier, G.W. "Verification of Forecasts Expressed in Terms of Probability," Monthly Weather Review 75, 1950, 1–3.
8. Clemen, R. and Winkler, R., "Combining Economic Forecasts," Journal of Business & Economic Statistics 4(1), January 1986, 39–46.

Hubbard Decision Research

About Doug Hubbard

Douglas Hubbard is the inventor of the Applied Information Economics (AIE) method and founder of Hubbard Decision Research (HDR). He is the author of *How to Measure Anything: Finding the Value of Intangibles in Business, The Failure of Risk Management: Why It's Broken and How to Fix It, Pulse: The New Science of Harnessing Internet Buzz to Track Threats and Opportunities* and his latest book, *How to Measure Anything in Cybersecurity Risk* (Wiley, 2016). He has sold over 100,000 copies of his books in eight different languages. Two of his books are required reading for the Society of Actuaries exam prep. In addition to his books, Mr. Hubbard has been published in several periodicals including *Nature, The IBM Journal of Research and Development, OR/MS Today, Analytics, CIO, Information Week, and Architecture Boston.*

About Hubbard Decision Research

Hubbard Decision Research (HDR) is a risk management consulting firm that applies quantitative analysis methods to the most difficult measurements and challenging decisions across many industries and professions. Using Applied Information Economics, HDR has developed quantitative analysis solutions to information technology investments, military logistics, entertainment media, major policy decisions, and business operations, for clients ranging from small businesses to Fortune 500 companies. More information can be found at hubbardresearch.com.

Live Calibration Training:

Attend our live, three-hour Calibration Training webinar and learn how to significantly improve your ability to accurately assess probability.

Learn More

CONTACT US

For a consultation on the methods used by HDR to use scientific quantitative methods to identify, assess, measure, and mitigate risk, contact the team.