SOCIETY OF ACTUARIES

Article from:

Forecasting & Futurism

July 2010 – Issue 2

Best Methods and Practices in Judgmental Forecasting

By Alan Mills

"We should be quite careful in trusting the intuitions of experts."
Daniel Kahneman[1]

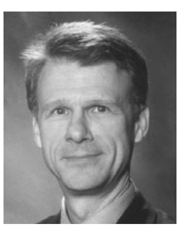
All forecasting is, at least in part, judgmental forecasting. No matter how formal, computerized or technically sophisticated your forecasting method, it is rooted in human judgment. Judgment is required to select the method, to choose its parameters, to filter its data and to interpret its results.

And some forecasting, even in actuarial work, is mainly judgmental. Actuaries sometimes forecast results using mainly their experience, "gut feel," opinions, and intuition; and chief actuaries often use their judgment to summarily override all model results. In the business world, such behavior is common: A 2003 survey of 240 U.S. corporations found that only 11 percent use formal nonjudgmental forecasting methods at all, and of these 60 percent routinely adjusted their forecasts based on judgment.[2] Economists are no different: judgment is "the primary factor that the economist uses in converting mere statistical and theoretical techniques into a usable forecast."[3]

Given the ubiquity of judgmental forecasting, it is important to understand the potential biases inherent in human judgment—the traps that can snare us—and to learn methods and best practices to avoid them. The aim of this article is to help you with these goals.

BIAS AND ERROR IN HUMAN JUDGMENT

Our understanding of bias and error in human judgment comes mainly from the work of Daniel Kahneman and Amos Twersky *(see sidebar, Kahneman and Twersky)*. Their experiments in human judgment and decision-making, together with the experiments of scientists who followed them, uncovered a startling result: much of human judgment is based not on rational cognitive processes, but



Alan Mills

Alan Mills, FSA, ND, is a family practice physician. He can be reached at *alan.mills@earthlink.net.*

Kahnemann and Twersky

Psychologists Daniel Kahneman and Amos Twersky shared one of the most productive collaborations in the history of social science. Starting in 1969, for more than 25 years they conducted groundbreaking experimental research into human judgment and decision-making. Their research had such a profound impact that in 2002 Kahneman became the first psychologist to win a Nobel Prize in economics (an honor that, had he lived, Tversky would have shared).

As an example of one of their experiments: In two trials, participants immersed a hand in cold water until instructed to remove it. The first trial lasted 60 seconds at 57 degrees Fahrenheit (very painful), and the second trial lasted a total of 90 seconds with 60 seconds again at 57 degrees followed by 30 seconds at 59 degrees (a little less painful). When asked which of the two trials they would choose to repeat, the remarkable finding is that 65-80 percent of subjects elected to repeat the second trial, even though it was longer than the first trial and produced more pain.

This, and a host of similar experiments, led Kahneman and Twersky to conclude that we store memories of our experiences according to what they called a "peak/end rule heuristic:" our memory of events is primarily an amalgam of the peak point of the experience and its end point. Nothing else matters.

For an excellent introduction to their work, see the YouTube videos Kahneman (2008a) and Kahneman (2008b).

FOOTNOTES

1 Kahneman (2008a)
2 Sanders & Manrodt (2003)
3 McAuley, 1986, p. 384

rather on *heuristics*—unconscious "rules of thumb" that humans have developed over millennia to deal with our environment. And, although our ingrained and unconscious heuristics may have served us well in dealing with the dangers and opportunities faced by hunters and gatherers, in today's complex world, they can produce serious errors.

Following are research-based examples of heuristics, biases (i.e., heuristics that are systematically skewed from rationality), and common cognitive difficulties that often cause us trouble when we make judgmental forecasts:[4]

Peak/end rule heuristic: In an experience, we tend to remember only the most extreme point, and the final point (the peak and the end). Thus, our memories—even the memories of experts are highly inaccurate. Since our intuitions are largely based on memory, our intuitions also can be highly inaccurate.

Framing bias: We judge an issue according to how it is presented. Many times Kahneman and Twersky demonstrated that opinions about an issue can be reversed if the issue is simply presented in a way that is logically equivalent, but expressed differently.

Anchoring bias: In our judgments, we often rely too heavily on one piece of information. For example, in one of their first studies, Kahneman and Twersky asked people to write down the first three numbers of their telephone number, and add 400 to it. They asked the subjects to consider this number as a year AD, and then asked them to guess when Attila the Hun was defeated in Europe. Invariably, the resulting guess was very close to the result of the addition.

Representativeness bias: We tend to make judgments based on small samples that are not statistically representative.

Availability bias: We make judgments based on data that is easily available, rather than finding appropriate data.

Confirmation bias: We focus on aspects of the past that conform to our views, and generalize from these to the future. We are blind to what would refute our views, and only look for corroboration. This is the central problem of induction: we generalize when we should not.

Conjunction bias: In an experiment, researchers found that when a terrorist attack had occurred recently, people about to board a plane are willing to pay more for insurance that covers terrorism than for insurance that covers any cause of death, including terrorism. Thus, recent significant events cloud our ability to reason.

Narrative bias: We automatically fabricate stories, weaving narrative explanation into a sequence of historical facts, and thereby deceive ourselves that we understand historical causes and effects and can apply this understanding to the future. This bias gives us a false sense of forecasting confidence, a sense that the world is less random and complex than it really is—a complacency that leads to forecast error.

Proximate cause bias: In our search for cause and effect relationships, we tend to consider only the most proximate causes.

Expert bias: We overvalue expert opinion.

Difficulty judging probabilities: Kahneman and Tversky found that people, even those who are statistically sophisticated, are not good at judging probabilities:

- When people are asked to estimate the probability that a randomly selected group of men has an average height over six feet, they give about the same probability whether the group consists of 10 men or 1,000.
- When presented with new information, people tend to ignore other probabilities. For example, suppose a reliable test for a rare medical condition is positive for

FOOTNOTES

[4] Most of these can be found in Kahneman, Slovic, & Tversky (1982) and Kahneman & Tversky (2000).

CONTINUED ON **PAGE 8**

you. The doctor tells you that one person in 10,000 has the condition, that for a person with the condition the test returns an accurate positive result 99 percent of the time, and that the test gives an accurate negative result 99 percent of the time for people who do not have the condition. Should you worry? Most people worry, because they focus on the new information, the positive test result, rather than the problem as a whole. They conclude that the chances are overwhelming that they have the disease, when in fact the chance is only about 1 in 100.

To understand this result, consider a group of 10,001 people who are tested for the disease. On average, only one of the people in this group actually has the disease, and 10,000 do not. For that one person, 99 percent of the time, the test will return a positive result. However, for the remaining 10,000 people, the test will return a positive result for 100 people (because the test returns an accurate negative result for only 0.99 x 10,000 = 9,900 people). Thus, if you are someone who received a positive result, you may be the one person who has the disease, or you may be one of the 100 for whom the test returned an incorrect positive result. Your chance of having the disease is therefore 1/101, or about 1 percent. You need not worry too much.

Overconfidence: When asked for the probability that their prediction of some event will come true, people—especially experts—systematically report a probability that is far too high. Similarly, they are overconfident that particular disasters will not happen. As a dramatic example, using their judgment, NASA managers assessed the probability of failure for the space shuttle Challenger as 0.00001, even though their engineers assessed the probability as 0.01.[6] In studies performed by numerous researchers, experts provide judgmental prediction intervals for their forecasts that are far too narrow; they are overconfident in their forecasts by a wide margin, and especially so when in addition to prediction intervals they are asked to provide point predictions.[7]

The list of biases and cognitive difficulties goes on and on. But you get the point: no matter how much mathematics and statistics we study, or how much experience we have, our judgment is largely governed by unconscious heuristics, and is prone to substantial error. We are human and we err. (*see sidebar, Downright Humiliating for Experts*).

METHODS

To guard against our inherent biases that lead to judgmental forecasting error, there is much that we can do. First, we can follow methods that, according to research, are more accurate. The following chart on page 9 shows 10 common judgmental forecasting methods.

In the chart, the methods are arranged according to the type of forecasters (whether individuals, groups, individuals or group, or automated) and according to their degree of struc-

Downright Humiliating for Experts

"… one of the great classics in the history of psychology, work by Paul Miel, was a seminal study looking at all the previous studies that had compared the performance of experts (like clinical psychologists) predicting various criteria, to very simple linear combinations of variables. So, you have a clinical psychologist looking at a lot of information, with a subset of that information used in a statistical model, and you compare how well the intuitions of people do compared to the statistical model. The statistical model is based on only part of the information, and is applied in a restricted way as a simple linear combination of variables. The stunning result was that in all the studies that Miel looked at, the linear equations beat the experts hands down ... a simple model does better.

Now, 50 years later, at last count there are 180 studies along the same lines, and basically the conclusion is the same: When you compare people to very simple combinations of variables, the combination wins just about every time. It's hard to find any exception. The results of this research are downright humiliating for people who try to forecast complicated events."

Daniel Kahneman[5]

FOOTNOTES

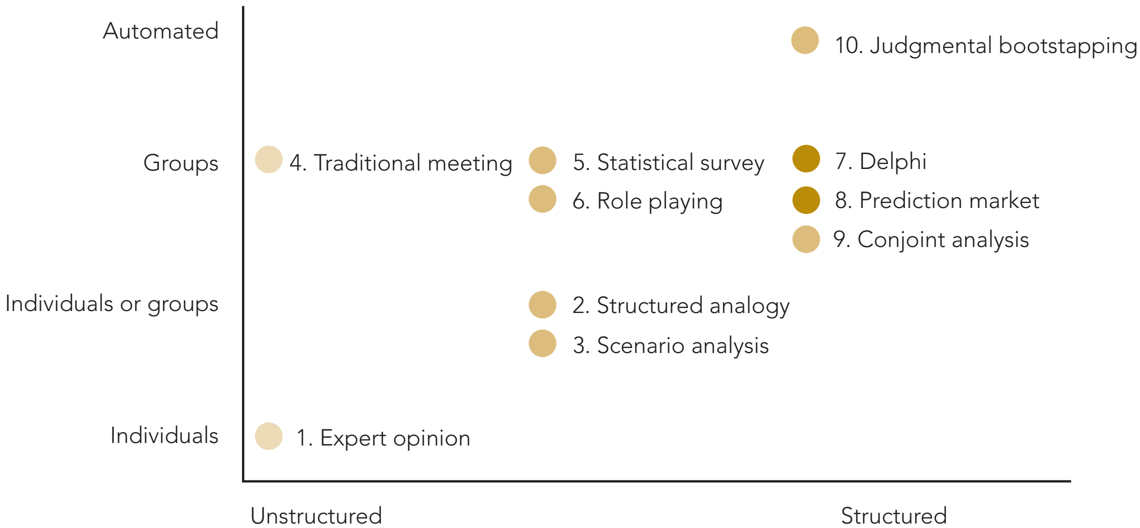
5 Kahneman (2008a).

6 From the Feynman appendix (Appendix F) to the Rogers commission report on the space shuttle Challenger accident *(science.ksc.nasa.gov/shuttle/missions/51-l/docs/rogers-commission/table-of-contents.html).*

7 Lawrence, Goodwin, O'Connor, & Onkal (2006), pages 505-506.

ture. Structured methods consist of systematic and detailed steps that can be described and replicated. The color depth of the dots in the chart represents the methods' relative accuracy; the deeper color is more accurate. As the chart shows, unstructured methods are generally less accurate than structured, and the individual method is less accurate than group methods.

Following are brief descriptions of these methods, in the order indicated on the chart.



Individual method

1. Expert opinion

Perhaps the most common judgmental forecasting method is to ask the opinion of an expert. Although common, this method is perhaps the most error-prone. It is generally unstructured, and fraught with all the biases and cognitive difficulties of an individual human.

Individual or group methods

2. Structured analogy

This method compares a recent series of events to a similar series that occurred earlier in another context. Forecasted outcomes are then based on past actual outcomes in the other context. The two series generally share important characteristics such as time scale, severity, reversibility, impacted sector, aggravating factors, etc. The purpose of the method is to constrain judgmental forecasting to a situation that actually happened in the past.[8]

3. Scenario analysis

The scenario analysis method is a process of forecasting future events by framing alternative possible outcomes in terms of story-like narrative scripts that often include the impact of events such as new technology, population shifts or changing consumer preferences. Usually, the method includes development of a most likely scenario, along with at least one optimistic and one pessimistic scenario. The primary purpose of a scenario is to constrain judgmental forecasting to a narrative structure, with the aim of producing more realistic forecasts. Scenario analysis is used by many organizations for long-term forecasting, and has been found to be quite accurate when used properly.[9]

Group methods

4. Traditional meeting

The most common method to obtain a judgmental forecast from a group of people is the traditional meeting, with unstructured discussion around a table. But this method has a number of drawbacks: The outcome is often unduly influenced by expert bias, difficulties of communication, psychological factors such as yielding to the opinions of

FOOTNOTES

[8] Armstrong (2001), Chapter 7: Analogies, pages 193-213.
[9] Mills & Bishop (2000)

CONTINUED ON **PAGE 10**

BEST METHODS AND PRACTICES … | FROM PAGE 9

authority figures or aggressive personalities, and the excessive influence of majority opinion.

5. Statistical survey

The statistical survey is a method of obtaining opinions about a specific topic from a sample of a population that is considered to be representative of the whole population. Statistical surveys are a widely used judgmental forecasting method.[10] For example, the Library Research Service recently surveyed libraries about the impact of e-readers on the future of libraries, and specifically about whether libraries will cease to exist. Used properly, a survey can produce a useful judgmental forecast.[11]

6. Role playing

In this method, people play roles to enact a situation in a realistic manner. It is particularly useful in forecasting the outcome of competition or conflict. The method has been shown to be substantially more accurate than expert opinion.[12]

RAND and Delphi

Olaf Helmer and Norman Dalkey of the RAND Corporation developed the Delphi method in 1953. It was based on a prior method, also developed by RAND, to combine the opinions of horse-racing handicappers to improve the chances of winning horse races.

Helmer and Dalkey thought a group of experts is the best way to develop a judgmental forecast (as Dalkey put it, "n heads are better than one"), but they knew from experience that a group of experts around one table can lead to argument and little progress. They designed Delphi to maximize the information that can be obtained from such a group.

In 1964, RAND published the first Delphi study, titled "Report on a long-range forecast." The study included a panel of 82 experts, including Issac Asimov, Arthur Clarke, Bertrand de Jouvenel, and Dennis Gabor. Its purpose was to forecast scientific and technological advances through the year 2000 and beyond. Most of the forecasts turned out to be amazingly accurate.[13]

7. Delphi

The Delphi method is a structured group consensus method for obtaining judgmental forecasts from experts *(see sidebar RAND and Delphi)*. It polls the experts anonymously, over successive rounds, with summary feedback in between. Over the course of a Delphi study, participating experts see where they stand in the group and may adjust their views accordingly. The feedback in successive rounds includes reasons for the more extreme views. Sometimes those reasons convince others that they are unwittingly making an erroneous assumption or ignoring an important piece of evidence. The result is a deeper exploration of the reasons behind expert opinions, without the biases, psychological impediments, and conflict that often appear in unstructured group discussions.

The Delphi method generally produces a rapid narrowing of opinions, and has been demonstrated to provide more accurate forecasts than unstructured group discussions. Interestingly, a face-to-face group discussion following the application of the Delphi method generally degrades forecast accuracy.[14]

8. Prediction market

A prediction market is another structured group consensus method, one that develops judgmental forecasts based on the mechanism of a speculative market. In this method, participants buy or sell shares of 'claims' regarding a particular forecast (e.g., the next president, an Oscar winner, or the increase in health care expenditures next quarter). If a claim turns out to be true, then one share is worth a stated amount (e.g., a claim is worth $1 if a particular candidate becomes president). A participant places a bet on the outcome by buying or selling shares at a market-determined price. For

FOOTNOTES

10 Fowler (2009)
11 Surowiecki (2004) and (2007).
12 Armstrong (2001), Chapter 2, pages 13-30.
13 Mills & Bishop (2000)
14 Mills & Bishop (2000) and Adler & Ziglio (1996).

example, a participant might buy 100 'yes' shares that a certain candidate will become president at 60 cents per share. The participant would then win $100 if the candidate won. The current market price can thus be interpreted as the participants' forecast of the probability of the event or the expected value of the parameter. Prediction markets are strikingly prescient.[15]

9. Conjoint analysis

This method quantifies respondent judgments and opinions, by having the respondents trade conflicting event or object attributes against one another. Analysis of these trade-offs reveals the relative importance of the attributes. The method is often used to forecast consumer response to new products. For example, pharmaceutical companies use this method to understand physician opinions about drugs, in order to forecast product sales and market share.[16]

AUTOMATED METHOD

10. Judgmental bootstrapping

Judgmental bootstrapping is an automated expert system that models an expert's reasoning process. To develop the model, an expert's forecasts are regressed against the information the expert used to make the forecasts. Because such models apply the reasoning processes of experts in a consistent way, studies from psychology, education, personnel, marketing and finance have shown that bootstrapping forecasts are more accurate than forecasts made by experts with unaided judgment.[17]

BEST PRACTICES

In addition to using methods that researchers recommend, we can employ well-researched best practices in applying these methods. Following is a review of such practices. They are organized according to the methods to which they apply.

Practices that apply to all methods

Provide feedback

One of the key findings of researchers is that records should be kept about judgmental forecasts, in order to provide the forecasters with feedback. Feedback is valuable because it enables the forecaster to learn. There are three main types of feedback:

- **Outcome:** The most common type, providing the forecaster with the latest observation in a series.
- **Performance:** Describes the accuracy and biases of the forecaster's forecasts.
- **Cognitive process:** Describes the strategy the forecaster used to arrive at the forecast. For example, such feedback might include a graphical display of the weights the forecaster attached to different data.

Researchers have found that outcome feedback is the least effective form.[18] For actuarial judgmental forecasts, a combination of all the feedback forms would likely be most useful.

Provide checklists

Provide the judgmental forecaster with a checklist of information categories relevant to the forecasting task. Checklists remind forecasters about factors relevant to their forecasts, and prevent them from being influenced by extraneous information.[19]

Present data clearly, in both graphs and tables

Present information to the forecaster clearly; in particular, avoid presentations that require forecasters to recognize complex patterns or to mentally aggregate many numbers. Because people vary in their ability to extract information from graphs and tables, present data in both formats.[20]

Frame questions in different ways

To avoid the framing bias, pose questions in various ways, from various perspectives.[21]

FOOTNOTES

15 Servan-Schreiber, Wolfers, Pennock, & Galebach (2004) and Surowiecki (2004).

16 Armstrong (2001), Chapter 5: Conjoint analysis, pages 145-167.

17 Armstrong (2001), Chapter 6: Judgmental bootstrapping, pages 169-192.

18 Lawrence, et al. (2006), page 507; and Armstrong (2001), page 63.

19 Armstrong (2001), page 61.

20 Lawrence, et al. (2006), page 497-498; and Armstrong (2001), page 64 and 93.

21 Armstrong (2001), page 697.

CONTINUED ON PAGE 12

Use mechanical methods
To help forecasters process complex information, especially statistical information, use mechanical methods rather than relying on judgment or mental processes.

Combine forecasts
Researchers have found that combining judgmental forecasts with either statistical forecasts or with other judgmental forecasts improves forecast accuracy.[22]

Practices that apply to groups

Use heterogeneous groups
A forecast developed by a group, especially a heterogeneous group, is generally more accurate than one by an individual, even if the individual is an expert. Generally, the various structured consensus methods will produce more accurate results than an individual expert.[23]

Employ an adequate number of forecasters
In surveys and prediction market models, make sure that the sample size is adequate to represent the entire population. In expert consensus methods, use between five and 20 experts.[24]

Pretest questions
Prior to data collection, questions should be tested on a sample of potential forecasters to ensure that they are understood and that they relate to the objectives of the problem.

Practices that apply to groups of experts

Use heterogeneous experts
If you use an expert consensus method, such as the Delphi method, make sure the experts are heterogeneous, that they vary in their information sources and in the way they approach the problem.

Request justification in writing
Experts should provide the reasons for their forecast, in writing. The Delphi method incorporates this best practice.

Provide numerical scales with several categories
To avoid anchoring bias, use as many categories of potential forecast responses as reasonable.[27]

Require multiple forecasts
Ask experts to make forecasts, and then repeat the process some days later. The Delphi method incorporates this practice.[28]

Require confidence intervals
Require experts to use confidence intervals, rather than point predictions.

Early in his career, Daniel Kahneman coined the term "illusion of validity," to capture the truth that we—especially if we are "experts"—often harbor an illusion that we are good at judgmental forecasting, when in fact we are not. In his research, he and Amos Twersky showed us that the accuracy of our judgment is severely compromised by inherent and unconscious cognitive weaknesses. Based on their work, others have developed methods and best practices to circumvent these weaknesses, and improve our judgmental forecasts. In your work and the work of your colleagues, do you see any opportunities to use these methods and best practices to improve your judgmental forecasts?

FOOTNOTES

22 Lawrence, et al. (2006), page 508.
23 Surowiecki (2004), (2007), and Armstrong (2001) page 698.
24 Armstrong (2001), pages 698-699.
25 Armstrong (2001), page 698.
26 Armstrong (2001), page 697.
27 Armstrong (2001), page 698.
28 Armstrong (2001), page 699.

REFERENCES

Adler, M., & Ziglio, E. (1996). *Gazing Into The Oracle: The Delphi Method and Its Application to Social Policy and Public Health.* London: Jessica Kingsley Publishers, xi, 252 p.

Armstrong, J. S. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners.* Boston, Mass.: Kluwer Academic, xii, 849 p.

Fowler, F. J. (2009). *Survey Research Methods* (4th ed.). Thousand Oaks: Sage Publications, x, 201 p.

Kahneman, D. (2008a). "Explorations of the Mind — Intuition: The Marvels and the Flaws," *Hitchcock Lectures*. Retrieved from *www.youtube.com*. 56 minutes.

Kahneman, D. (2008b). "Explorations of the Mind — Well-being, *Hitchcock Lectures.* Retrieved from w*ww.youtube.com*. 59 minutes.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty:* Heuristics and Biases. Cambridge; New York: Cambridge University Press, 555 pages.

Kahneman, D., & Tversky, A. (2000). *Choices, Values, and Frames.* New York; Cambridge, UK: Russell sage Foundation; Cambridge University Press, 840 pages.

Lawrence, M., Goodwin, P., O'Connor, M., & Onkal, D. (2006). "Judgemental Forecasting: A Review of Progress Over the Last 25 Years. *International Journal of Forecasting,* 22, 493-518.

Mills, A., & Bishop, P. (2000). *"Applied Futurism: An Introduction for Actuaries.* Society of Actuaries, 128 pages.

Page, S. E. (2007). The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton: Princeton University Press, xxiv, p. 424 .

Sanders, N. R., & Manrodt, K. B. (2003). "The Efficacy of Using Judgmental Versus Quantitative Forecasting Methods in Practice. *Omega* (31), 511-522.

Servan-Schreiber, E., Wolfers, J., Pennock, D. M., & Galebach, B. (2004). "Prediction Markets: Does Money Matter? *Electronic Markets,* 14 (3 - September 2004), 243-251.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (1st ed.). New York: Doubleday, xxi, 296 p. ▼