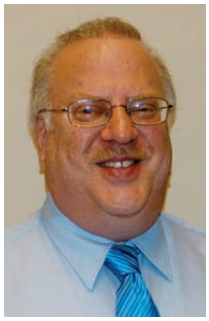
ANALYSIS PLACEBOS



THE DIFFERENCE BETWEEN PERCEIVED AND REAL BENEFITS OF RISK ANALYSIS AND DECISION MODELS.

BY DOUGLAS HUBBARD (LEFT)
AND DOUGLAS A. SAMUELSON (RIGHT)

In the first few years of the 21st century, events like terrorist attacks, Hurricane Katrina and the financial crisis have focused much attention on risk management. Did risk analysis itself fail? Were the analyses right, but not heeded? Was the whole cascade of events beyond anyone's ability to predict? While the definitive answer may be years away, we can say, based on substantial evidence, that many methods used in risk assessment do not work, even if they seem useful at the time they are applied. Of course, risk analysis is really just a subset of decision analysis and the same observations can apply.

How do we know what works? That depends on what we mean by "works." Suppose an organization is considering a new way to assess some critical set of decisions. Perhaps this decision process involves selecting from among several alternatives for major capital investments,

improved security measures, improving safety at a major chemical plant, or deciding whether to proceed with a risky surgical procedure. Now suppose we asked the users of this new method to select among the following objectives for this new decision analysis process:

1. The decisions will be better (i.e. observable outcomes over time will actually be improved)

OR

2. Users will *feel* better about the decision (i.e. there is high acceptance of the analysis process and the recommendation it produced)

Would the users of the decision process really think the second objective would be satisfactory? Would they really think that improved outcomes are not the primary criterion in choosing a decision analysis method – even if the decision process is known to be a "soft" method?

Just a few decades ago, a freezing mountain climber might have thought it was a good idea to drink brandy since it created a sensation of warmth (remember the image of the Saint Bernard carrying a flask of brandy on his collar to a freezing hiker?). We now know that the alcohol causes capillaries in the skin to expand and that the sensation of warmth is actually heat leaving the body faster. The mountain climber felt better off but was actually worsening his hypothermia. If you are freezing and only care about *feeling* better, drink the brandy. If you care about your actual chance of survival, don't drink it.

Unfortunately, many popular decision analysis methods seem to confuse *feeling* better with *doing* better, or don't bother to include any measurements that would highlight this distinction. Of all the things that might be measured in organizations, all too often the actual effectiveness of decision analysis methods is among the least measured. Whether a method is soft and informal or rigorously quantitative, the question of whether it actually improves decisions in the long run is rarely even questioned much less quantified. This has led to what are probably a long list of unproven methods (even though they are touted as "proven") being used for major, critical decisions that affect the financial well-being of organizations as well as health and safety of the public.

Recently, decision scientist Robert Clemen of Duke University made a call to action to measure the effectiveness of decision analysis methods (Clemen 2008). He would call a method "strongly effective" if it measurably increased desired outcomes such as higher returns on portfolios, reduced industrial accidents,

increased sales, improved surgery patient results, average monetary return on investment in movie projects, and so on.

Clemen differentiates strongly effective methods from the merely "weakly effective" methods. These methods only show that the results of the analysis were preferred by the users or that satisfaction with/acceptance of the process was high. Weakly effective methods may not, for example, actually improve the ability to select movie projects that make more money than they cost. They can only show that the users of the method would be satisfied with the outcome of the analysis.

Obviously, for such observations to support the claim that a method is strongly effective, they must be repeated for a significant number of decisions tracked over a long period of time. These results must then be compared to individuals, teams or organizations using alternative methods over another sufficient sample size of decisions. Like a clinical drug trial, a test group and control group would be tested side by side and outcomes would be observed over time. But there is a dearth of any evidence of strong effectiveness. As Clemen noted: "Virtually no research has been done that compares DA with other decision-making techniques in terms of strong effectiveness. Relatively little work has been done to show weak effectiveness."

In other words, if a decision analysis method is evaluated with any "performance metric" at all, it is, at best, a measure of whether the users felt good about the process and the results. But, like that feeling of warmth from brandy, this is the very kind of performance metric we should be suspicious of. For example, a study published in *Organizational Behavior and Human Decision Processes* showed that gathering more information makes you feel better but, at some point, begins to reduce decision quality while confidence continues to increase (Tsia et. al. 2008). An earlier study in the same journal showed how interaction with others also increases decision confidence but, again, at some point decisions are not improved while confidence continues to increase (Heath et al, 1995). It appears to be a little too easy for decision-makers to increase their confidence in decisions (i.e. a form of weak effectiveness) without improving decisions.

This distinction between strong and weak effectiveness might help resolve

some long controversies in DA. Popular methods, in particular, should be scrutinized in this light. One popular method supported by several software products is Analytic Hierarchy Process (AHP). AHP has many passionate proponents and this alone may be evidence that AHP is at least weakly effective. Almost all of the published material to date about AHP has been criticism regarding theoretical flaws (Dyer 1990; Holder 1990, 1991; Schenckerman 1997; Perez 1995; Perez et al 1996) or, more frequently, merely case studies about the application of AHP to a particular problem that never actually measure the benefits of the method in controlled environments of a large number of samples. The case for strong effectiveness is not made by the proponents, and the case against it is not made merely by pointing out the possibility of theoretical errors. (Perhaps in real-world problems the remaining theoretical issues of AHP are too infrequent to matter.) Only a

controlled experiment could prove the strong effectiveness of AHP either way. So far, only a 2007 study published in the *European Journal of Operations Research* comes fairly close to accomplishing this (Williams et al 2007).

Like Clemen, the authors of the EJOR study believed that decision support systems like AHP software tools are rarely if ever tested to determine if actual results are an improvement. In a task of selecting applicants for college admission, teams using different methods were asked to select which students should be accepted. The authors found that AHP showed no measurable improvement in selecting applicants. (The researchers already knew what the optimal answers would be, so the results were evaluated according to which methods picked the known best answers). In some cases, subjects using AHP actually performed worse than subjects not using AHP. At the same time, they found that satisfaction with the result was higher with AHP users, even though the decision quality was clearly not improved. Perhaps a broader longitudinal study of problems closer to real-world business decisions would be a fairer test of AHP. But there is as yet no such study, and previously mentioned research suggests the possibility that the perception of benefits from a decision method could be as much an illusion as the sensation of brandy staving off hypothermia.

Regarding observed outcomes as the measure of success, one comment has to be made about something frequently heard from decision analysts ¬ that there is a difference between good outcomes and good analysis. But this is only true at the level of an individual decision. If you were to bet even money on the roll of a single die and the choices were betting on a roll of "6" and "not 6," it would be rational to choose "not 6," even if it subsequently turns out that you lost the bet. Over the long run, betting on the "not 6" result would be a strongly effective strategy. It is possible for individual outcomes to go against the rational choice but over the long run good outcomes *are* how we identify good decision processes.

Yet the subjective perception of value is apparently what suffices for most "evidence" of the effectiveness of a method. Feeling better about a decision without actually making better decisions might possibly be a goal, and we won't question whether it should suffice as a goal or not. Perhaps the mountain climber knows his time is

limited and just wants to feel better in his last few moments alive. But we propose that most decision-makers use a method because they actually believe that their decisions will improve and not merely because they "helped build consensus" or "improved communication about the proposed ideas."

Fortunately, plenty of methods have been proven to be strongly effective. While other methods are waiting to be validated empirically (for strong effectiveness, not just weak), keep the following research in mind:

- Decomposition of extremely uncertain quantities improves estimates (MacGregor et al 1994). It has been shown that decomposing an estimate into its component parts and estimating those, does improve the overall estimate but more so for extremely uncertain quantities. For example, the estimate of project cost can be improved by decomposing the cost into multiple components, estimating each of them and then adding them up. This study, however, only proves the benefits of "explicit" models. That is, where the mathematical relationships of the variables are known.
- "Calibrated Probability Training" improves the ability of experts to subjectively assess probabilities (Lichtenstein et al 1982; Murphy, Winkler, 1977; Hubbard 2007). Several studies show that the ability of experts to assess subjective odds or intervals can be improved through training. This is critical since most Monte Carlo models include at least some subjective estimates and about a third are mostly subjective estimates. But since the very few models take advantage of calibration training, the experts providing estimates will tend to be overconfident and risk will be underestimated.
- Certain types of linear models improve on the intuition of experts. Paul Meehl showed that regression models based on historical data consistently outperform expert intuition in a wide variety of topic areas – even when the experts argue that each situation is so complex and unique that historical models can't possibly apply (Meehl 1986). The "Lens Method" developed from Egon Brunswick's research in the 1950s also shows consistent improvements in the forecasts and

estimates of experts using their unaided intuition. The Lens Method is also based on a regression model but, instead of using historical results, the Lens uses the subjective judgments of experts on a series of hypothetical scenarios (Karelaia et al 2008, Hubbard 2007).

- The use of Monte Carlo modeling methods appears to outperform soft risk analysis methods when applied to exploratory oil companies or NASA space missions (Hubbard 2009). In this case, the analysts at NASA were using decomposition (fundamental to any Monte Carlo model), calibration and historical models. In over a hundred missions, the Monte Carlo simulations and historical models consistently outperformed the project scientists and engineers (who used a "softer" ordinal scoring method) in the forecasts of costs, schedule and mission failure risks.

What is striking here is that none of the validated methods above have much in common with the more popular ordinal scoring methods used widely throughout government and business. In fact, weighted ordinal scores add errors that may easily outstrip any actual benefits (Cox, 2008, Budescu Broomell 2009; Hubbard 2009). Again, some decision analysis methods can't show strong effectiveness, but some can.

"Placebo" is Latin for "I shall please" and it is not so much of a stretch to apply this word to many types of decision-making methods. The research shows that the placebo effect might be the only effect for some methods. If there were "truth in labeling laws" for DA as for prescription medications, some DA methods would have to come with a disclaimer like this:

"This method is a placebo. While there is evidence that this method can cause a sense of elation and increased confidence about the decision, there is no scientific evidence that decisions will actually be improved over the long run. Side effects include a complete waste of time and money and, in some cases, decisions may be worse than what unaided intuition would have yielded."

Douglas W. Hubbard *(dwhubbard@hubbardresearch.com) is president of Hubbard Decision Research, in Glen Ellyn, Ill. Since it was published in July 2007, his book, "How to Measure Anything: Finding the Value of Intangibles in Business" has been the best selling business math book on Amazon. His new book, "The Failure of Risk Management: Why It's Broken and How to Fix It," was released in April 2009.*

Douglas A. Samuelson *(samuelsondoug@yahoo.com) is president and chief scientist of InfoLogix, Inc., an R&D and consulting company in Annandale, Va. He is a frequent contributor to Analytics and OR/MS Today.*

REFERENCES

For complete references, see: http://www.analyticsmagazine.com/fall09ref.html