The Canadian Journal of Program Evaluation Vol. 21 No. 1 Pages 47–62
ISSN 0834-1516

MAKING COST-BENEFIT ANALYSIS A PRACTICAL TOOL FOR EVALUATION

Kenneth Watson
Rideau Group
Ottawa, Ontario

Abstract: A cost-benefit evaluation requires precise data on program outcomes. However, such data are unavailable when the analysis is prospective, and expensive and time-consuming to collect when the analysis is retrospective. This problem of uncertain data is partly solved by the revised version of the Treasury Board *Benefit-Cost Analysis Guide* (Watson & Mallory, 1997), which allows probabilistic estimates of program results to be used in the analysis. There are not yet many examples of this technique in practice. One is Transport Canada's evaluation of alternative requirements for small commercial vessels to carry emergency signaling equipment. This article describes that evaluation and assesses how well the methodology worked.

Résumé: Une évaluation des coûts-avantages exige des données précises sur les résultats des programmes. Toutefois, ces données ne sont pas disponibles quand l'analyse est de nature prospective, et elles sont dispendieuses et leur collecte exige beaucoup de temps lorsque l'analyse est rétrospective. Ce problème de l'incertitude des données est partiellement résolu par la version révisée du *Guide de l'analyse avantages* (Watson & Mallory, 1997), publiée par le Conseil du Trésor, qui permet d'utiliser des estimations probabilistes des résultats des programmes dans l'analyse. On ne dispose pas encore de beaucoup d'exemples pratique de cette technique. Un exemple est l'évaluation de Transports Canada des exigences modifiées pour les petits vaisseaux commerciaux de transporter de l'équipement de signalisation d'urgence. Cet article décrit cette évaluation et évalue la pertinence de la méthodologie.

Cost-benefit analysis is a powerful evaluation tool. When successful, it shows definitively whether a program is worthwhile, and, if there are alternatives, the analysis shows which of them is best. Nevertheless it has not been used as much as one might expect,

Corresponding author: Kenneth Watson, Rideau Group, 452 Roxborough Avenue, Ottawa, ON K1M 0L2; <watson@rideaugroup.com>

partly because it requires more precise data than the ordinary non-experimental evaluation can usually provide (Hahn, 1991).

However, there are two methodologies that, together, make it possible to deal with the uncertain data problem. First, the Treasury Board *Benefit-Cost Analysis Guide* (Watson & Mallory, 1997) describes a cost-benefit framework that can use, as inputs to the analysis, estimated data values and probabilities (rather than actual measured values). Second, the *Guide* describes ways to obtain these estimates on the basis of partial evidence.

An example of such an evaluation was the Transport Canada (2003) study of regulations that would require small commercial vessels to carry distress-alerting equipment similar to that required on larger commercial vessels. The evaluators were asked to assess whether rapid distress-alerting equipment should be mandatory and, if so, what kind of equipment was likely to be most cost-effective.

The uncertain outcome of the proposed regulation was the number of lives that would be saved. Of course this was not the only uncertain data value, but it was the key one. Most others (such as the appropriate discount rate, or the appropriate dollar value of a "life saved," for instance) could be taken from the research literature rather than estimated anew. Some other values (such as the cost of the distress-alerting equipment) also had to be estimated, but these were less important, either because their values were relatively small or because the uncertainty was relatively small. The key estimate was the number of lives that could be saved. If that could be estimated, and its probability stated, then the cost-benefit analysis could be done.

This particular evaluation was a prospective study, looking forward to the costs and effects of new regulations. However, the problem was essentially the same in principle whether the evaluation was prospective or retrospective, since both commonly have to make do with partial and uncertain data.

HOW WAS THE NUMBER OF 'LIVES SAVED' ESTIMATED?

The Delphi technique was used to estimate probabilities of lives being saved. This technique is one of several types of "contingent value" analysis ("what if" analysis). That is, the outcome estimated is dependent (contingent) on the circumstances. The technique was

developed in the context of forecasting and strategic studies (Adler & Ziglio, 1996; Linstone & Turoff, 1975; RAND Corporation, 2003).

Contingent estimates can be based on information from those directly affected or on expert opinion. Which of these is the best technique depends on the circumstances of the evaluation.

In some cases the evaluator can ask those affected "What would you be willing to pay for the program?" (Gafni, O'Brien, & Diener, 1998; Newhouse, 2002; Viscusi, 1993). This bypasses the measurement of effects and goes straight to values. In this case the evaluators thought that self-reporting of willingness-to-pay had limited potential for a number of reasons. First, it would be difficult for people to judge the level of risk correctly, because fatal accidents are rare occurrences and their frequency would be difficult to judge subjectively. Second, individuals might underestimate the value of requiring all small vessels to carry distress alerting equipment. If all carried it, then ship-to-ship assistance might be more frequent and effective (rather than relying on Canadian Search and Rescue). The more ships that carry the equipment the better off everyone is, but reaching a critical mass might be difficult without mandatory carriage. Third, small fishermen, for example, might not make wise judgments about the value of risk reduction. Some might be unduly influenced by macho attitudes, or tradition, or parsimoniousness, or some other factor that might limit their rational calculation of the equipment's value.

Therefore the evaluators decided to pursue the second method of obtaining contingent estimates — that is, to ask experts to provide them. Specifically the evaluation team convened a group of diverse experts and had them consider each relevant fatal accident during the previous five years and make an estimate of the probability that a death could have been prevented if better distress-alerting equipment had been available.

The evaluators led the group in a Delphi exercise. This technique had been used in long-range forecasting and strategic planning (Linstone & Turoff, 1975; RAND Corporation, 2003), but it was an experiment in this evaluation context. The salient characteristics of the Delphi exercise were:

- Careful marshalling of information beforehand on each accident case-by-case.
- Use of experts with diverse relevant experience.

- Independent estimates by each expert made separately and anonymously to prevent any follow-the-leader effect.
- Repeated rounds of estimates, each time with the group knowing the average results of previous rounds of estimation, and with some discussion preceding each round.

If these requisites are met, it has been demonstrated that group-based judgments are generally more accurate than estimates by a single expert (Surowiecki, 2004).

A key concept in the Delphi technique is the importance of obtaining a "consensus estimate" from participants without problems arising from unhelpful interactions among them. The anonymity of each individual estimate minimizes the suppression of diverse views that might otherwise arise out of authority, prestige, or personality dominance within the group.

Two other factors were important in making the Delphi exercise a success. First, the experts were asked to answer just one, specific, single-dimension question (the likelihood of a life being saved in each case). Second, they were asked to do this on a specific case-by-case basis and given carefully compiled information on which to base their opinion.

To be useable in the Treasury Board cost-benefit model, the experts' estimates were analysed statistically to construct a probability distribution of "expected lives saved," rather than collapsing the estimates into a single figure and treating that single figure as if it were certain.

THE VALUE OF LIVES SAVED

Determining the "value of life" raises many conceptual and analytical issues. Some that were considered included values for avoidance of risk to oneself vs. responsibility to protect the lives of others (Volokh, 1997); loss of life in different ways that evoke different attitudes (Viscusi, 1993); willingness to pay for insurance as an indication the value of avoidance of financial risk to dependents and survivors (Newhouse, 2002); and the ability (or lack thereof) of people to rationally assess risk and to imagine and judge what they would actually pay to avoid a certain risk if they had the option (Harris, 1987; Morrow & Bryant, 1995).

In this study the value of life was not estimated directly. The evaluators looked for a consensus figure in the research literature and from accepted practice at Transport Canada and the Treasury Board of Canada. The main reason for this was practicality. An evaluation budget will seldom stretch to cover new estimation of all of the uncertain parameters.

Accepting a consensus value for the parameter might also be justified by a need for consistency across evaluations. It might be argued that the "value of life" should not vary from one federal government study to another, and, therefore, it is best estimated once by a central agency and then used consistently in evaluations in all departments and agencies.

There is some force to these arguments but, on balance, if budget had been available, it would have been better to estimate the "value of life" anew, in relation to those actually affected. There were various populations that the new regulations would cover — small fishers (owners and crew) and small commercial boat operators (operators and passengers). Their attitude to risk, their willingness to pay to ameliorate it, and their responsibilities to themselves and to the public may have varied (Cutler, 2004).

Deciding to accept a consensus figure for the "value of life" did not end the uncertainties because, in fact, the evaluators found no consensus. There was little agreement in the literature to draw upon. The Treasury Board of Canada, based on the research literature and legal precedents, had indicated that "life values" from $2,500,000 to $3,000,000 appeared reasonable (Watson & Mallory, 1997). A previous evaluation of marine safety regulations by Transport Canada (distress-alerting equipment carriage requirements for a different class of vessels) had used a value of $1,500,000 for each fatality avoided. This was stated to be a minimum economic value (Transport Canada, 2001). On the other hand, the Regulatory Assessment Methodology commissioned by the Council of Marine Carriers in 2001 had stated that this is the low end of the plausible range:

> An economic value of life of $US1.5 million is well below the low end of the range of estimates based on individuals observed willingness to pay for reductions in the risk of accidental death. Most scholars' estimates fall into the range $US3.5 million to $US6 million.

That is, approximately C$5.4 million to C$9.2 million. The U.S. Department of Transportation, since 1996, has used a "value of life" figure of US$2.7 million adjusted annually by the U.S. Gross Domestic Product Implicit Price Inflator.

In the absence of consensus among methodologists in regard to the value of a statistical life, the range selected for use in this evaluation, somewhat arbitrarily, was from the minimum Transport Canada value of C$1.5 million to the U.S. Department of Transportation value of C$3.7 million. Probabilities within that range were assumed to be distributed normally (bell curve). Again this assumption of a normal probability curve was arbitrary. The lack of consensus in the literature could as easily have led the evaluators to assume a flat probability distribution curve (no compelling evidence in favour of one value or another within the plausible range).

ESTABLISHING A BASELINE

Estimates of the number of small commercial vessels operating in the sea areas to be regulated varied from 27,000 to 35,000. The study used an average 31,000. Over the seven years from 1995 to 2001, small commercial vessels experienced 38 fatalities in 20 accidents in Canadian waters. In the same accidents, another 28 people were at risk but did not die. Most of the incidents involved fishing vessels, and, in most of these cases, there was only one person on board. Only three fatalities involved other kinds of work boats. Only one incident that caused loss of life involved a passenger vessel, but a relatively large number of people had been put at risk by this incident.

Table 1
Fatalities in Small Commercial Vessel Incidents, 1995-2001, by Type of Vessel

Type of Vessel	Number of Accidents Involving a Fatality	Number of Persons at Risk	Number of Fatalities	Percent of Total Fatalities
Fishing Vessels	15	35	29	76.3%
Work boats	3	10	6	15.8%
Passenger Vessels	1	20	2	5.3%
Other	1	1	1	2.6%
Total	20	66	38	100.0%

The evaluators used BestFit software to estimate the expected number of incidents and resulting fatalities per annum (in the absence of further regulation). This was the baseline. The BestFit software tested various continuous distributions against the time series of annual accident data and compared the fits of alternative types of distributions.

The best fitting general statistical distribution was found to be a Poisson distribution. This is a common type of probability distribution for the number of individual events that occur in a given unit of time, such as the number of deaths by maritime accident per year. The statistical form of the probability distribution for annual fatalities without new equipment carriage requirements was observed to be:

$$f(x) = e^{-\lambda}\lambda^{i}/x!, \text{ where, in this case, the mean } \lambda = 8.41$$

This risk distribution has no value below zero (no fatalities in a particular year), clusters around the mean, and has a long positive tail. That is, there is a small possibility that fatalities in a particular year might, for example, be as high as the 41 people who either died or were at severe risk in the worst year observed in the seven year sample covered by the study.

Estimating the Incremental Impact of the Proposed Regulations

The key question was how much the risk of fatalities would be lessened by distress-alerting equipment. The evaluators did not ask this as a general question to the experts. Instead they asked the experts to consider each of the accidents where a fatality had occurred and to estimate, on the evidence, how much distress-alerting equipment might have reduced the risk of a fatality in that accident. This is an important distinction between good Delphi work and poor Delphi work, in the opinion of the author.

The evaluators spent about three person-weeks compiling information on the fatal accidents that had occurred in small commercial vessels during 1995-2001. In 10 of the incidents, nothing was known of the accident until the vessel was overdue. Five incidents were in isolated locations and 13 were not. Five occurred in inland waters and 13 on home-trade voyages. In 7 cases there was a visual alert. In only 2 of 20 incidents was an emergency alert broadcast by radio.

Table 2
Type of Distress Signal in Incidents where a Fatality Occurred

Type of Distress Signal	Number of Incidents	Number of Fatalities
Overdue	10	23
Visual	7	12
Radio	2	2
Other	1	1
Total	20	38

Estimates of the Impact on the Fatality Rate

Step 1: Do we really need to be precise?

At one stage of the evaluation, it was thought that hypothetical calculations might be sufficient in themselves. That is, if the required reductions in fatalities were so small as to be self-evidently achievable by the proposed equipment carriage regulations, then it might not be necessary to push the analysis further. If, for example, a mere 5% fewer fatalities would have been sufficient to justify the regulations, then stakeholders (including those who would have to pay for the new equipment) might have been convinced of the worth of the regulation without further analysis. However, it did not turn out to be so simple. The evaluators found that a 33% reduction in fatalities was needed to justify requiring all small commercial vessels to carry the most cost-effective package of equipment; and a reduction of 21% was needed if only fishing vessels were required to carry the equipment. This lower figure reflects the fact that fishing vessels accounted for most of the fatalities.

Table 3
An Example of the Hypothetical Reduction in Fatalities Sufficient to Justify the Regulations

Assumed % Reduction in Fatalities	Net Present Value - All Small Commercial Vessels	Net Present Value - Small Fishing Vessels Only
40%	$17.05 million	$47 million
33%	0	$29.9 million
30%	-$7.4 million	$22.51 million
21%	-$29.4 million	0
20%	-$31.87 million	-$1.95 million
15%	-$44.1 million	-$14.18 million
10%	-$56.33 million	-$26.41 million
05%	-$68.56 million	-$38.64 million

The reductions in fatality rates that were required to justify the regulations were sufficiently large that further work was needed to ascertain whether or not it was credible that they would be achieved.

Step 2: What reduction in fatalities was to be expected?

To move beyond the hypothetical "what if" analysis of the impact of the proposed regulations, the evaluators needed estimates of what actual reductions in fatalities were to be expected. It was decided to convene a group of experts (the "Delphi Group") to address the issue of expected impact. The Delphi Group considered only those accidents where a scan of the file had indicated a possibility of life being saved.

Accident data for small commercial vessels during 1995 to 2001 was available from the Canadian Transportation Safety Board and from Canadian Search and Rescue. Their databases noted the number of incidents that involved fatalities and the characteristics of each accident (how authorities became aware of the accident, what the conditions were at the time, and how long a response team took to arrive at the scene). The information on the number of people at risk and the number of fatalities was reliable. The information on conditions at the time of the accident was less complete and sometimes not easy to interpret.

The Search and Rescue database contained information on 15,741 marine incidents from 1995 to 2001. The evaluators found that 153 of these involved small commercial vessel accidents in conditions where it was possible that more rapid distress alerting might have saved lives. The relevant accidents proved to be only 2% of all search-and-rescue incidents. Of the 153 accidents, 20 resulted in deaths — 38 deaths in total. The key question for the evaluators was how many deaths might be prevented, over a similar period of time, if all such vessels carried rapid distress-alerting equipment.

The experts were not asked to make a crude single judgement on overall change in fatality rates. Rather they were asked to bring their expertise to bear on each relevant case. What was the probability that carriage of rapid distress-alerting equipment would have prevented a fatality in each specific instance? The expected overall change in fatality rates was then calculated from these case-specific estimates of probabilities.

The exercise was conducted in a face-to-face group setting. A group of seven experts in marine safety and search and rescue met for about five hours to consider the 20 accidents that had resulted in fatalities and to consider in each case the probability that lives might have been saved by rapid distress and location alerting.

The group considered each case in several stages, facilitated by the consultant evaluator. First, a researcher presented the details of the case to the group, describing circumstances and outcomes. The group then discussed the characteristics of the accident and, especially, the possible role of distress-and-location alerting. Second, each expert, anonymously, made an estimate of the probability that a fatality might have been averted in that particular accident case if rapid distress-alerting equipment had been carried by the vessel. The worksheets used by the experts listed five categories of probabilities: 100% (the lives lost in the incident would have been saved for sure), high (70–99%), medium (40–69%), low (1–39%), and zero (no chance that carriage of rapid distress-alerting equipment would have led to saving lives).

It is interesting that in no case did any expert think that it was certain that a life would have been saved by better distress alerting. In contrast, the experts estimated 59 times in total (in assigning 214 probability scores) that there was zero chance that better distress alerting could have saved a life in that instance.

The facilitator calculated the average "probability values" after each iteration, based on the individual "probability scores" submitted by the experts. Without revealing the score given by any particular expert, the evaluator told the group what the pattern of response was (range and average, and, sometimes, other descriptive statistics). The group then discussed the case further, and each person made another estimate of probabilities, either revising his or her initial estimate or letting it stand (again, anonymously). After discussion and before each round of scoring, the group was asked whether anyone wished to change his or her score. Generally, after two rounds of scoring probabilities in a particular accident, no one wanted to change his or her score further, and in no case did scoring go past three rounds.

At the end of the workshop, the facilitator collated the final probability scores for each of the 20 accident cases and calculated average probabilities and ranges based on those scores using standard techniques to encode probability distributions (Spetzler & Holstein, 1975; SRI, 1978).

The net result was that mandatory carriage of Package 1 of equipment was judged likely to result in a 37% decrease in fatalities (12.76 lives saved over six years). Alternatively, mandatory carriage of Package 2 was judged to be likely to result in a 44% decrease (14.98 lives saved).

The calculation of a "bottom line"

This probability data was entered into a standard cost-benefit/risk-analysis model as described in the Treasury Board *Benefit-Cost Analysis Guide*.

Both equipment packages had similar positive outcomes on average. There was a significant possibility that Option 1 would have the higher pay-off, but the option was more risky. That is, it had lower minimum values as well as higher maximum values.

Table 4
Net Present Values (NPV) of Two Equipment Options

	Minimum NPV	Mean NPV	Maximum NPV	95% probable that NPV will be:
Option 1	-$51.6	$64.2	$317.6	More than $154.1
Option 2	-$30.0	$68.7	$284.7	More than $131.3

Sensitivity Analysis

The outcome of the cost-benefit analysis was influenced most by the level of fatality reduction achieved by the various carriage requirements and secondly by the "value of life" assumed. At low average values for these variables, there was a risk (30% for Option 1 and 20% for Option 2) that the economic outcome of regulation would be negative. At higher (but still plausible) estimates of the number of lives saved, there was little risk that the proposed regulations would have a negative outcome.

The expected change in the value of regulation over time would have been useful to know but was not calculated. Equipment prices were falling from year to year, and owners were gradually adopting the equipment without regulation. If late adopters were high-risk takers (accident prone), then a voluntary approach to equipment carriage might not have worked as well as a mandatory approach.

Figure 1
Net Present Values, Option 1 Equipment, All Small Commercial Vessels

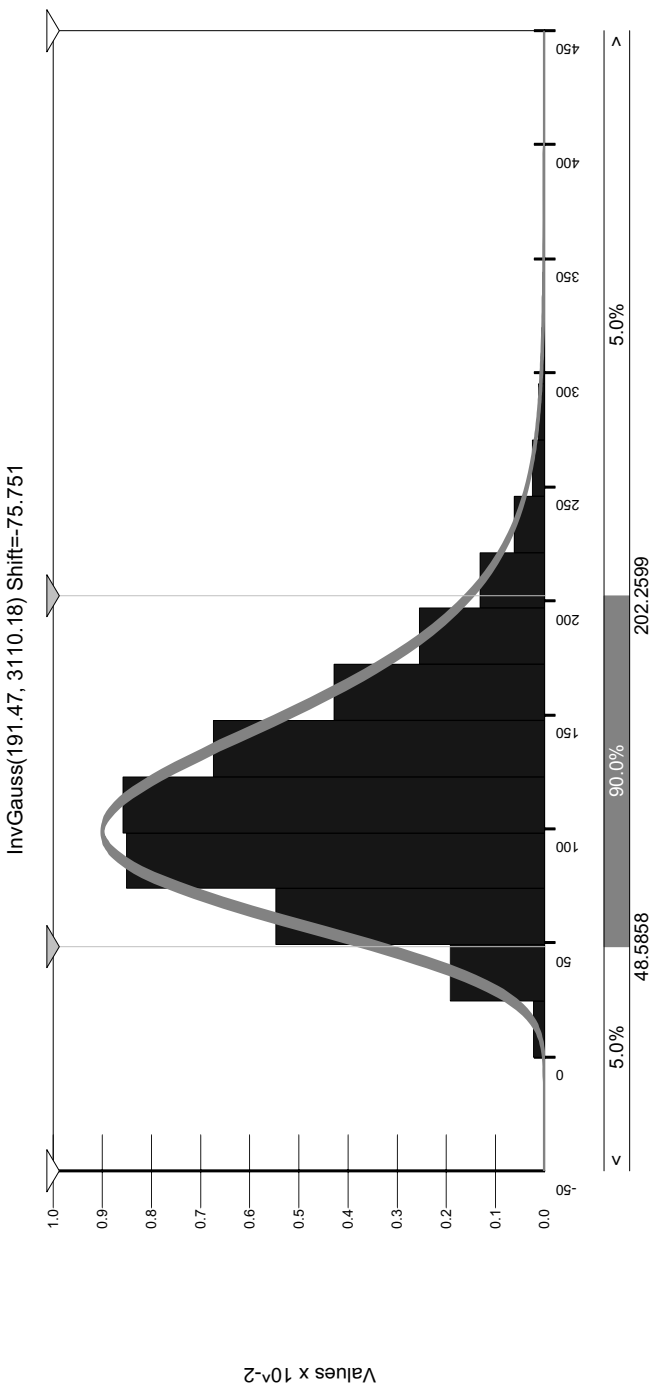
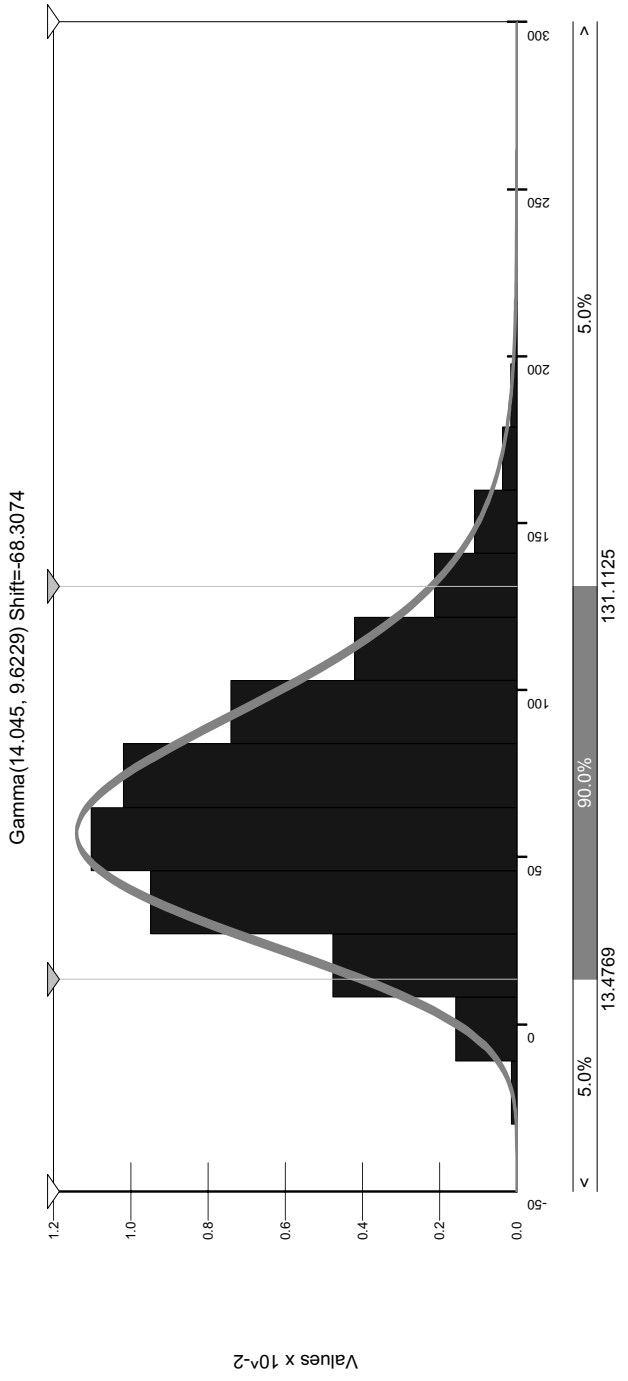


Figure 2
Net Present Values, Option 2 Equipment, All Small Commercial Vessels



CONCLUSION

The main interest of this study is that, on a small budget and despite data uncertainties, the evaluation team succeeded in producing a defensible cost-benefit analysis of a major regulatory program that applied the framework set out in the Treasury Board *Benefit-Cost Analysis Guide*.

The result was that decision makers were given good advice on the basis of the evidence and also made aware of the level of risk (uncertainty) that remained. The decision was not taken out of their hands by the evaluators. In our opinion this is the ideal outcome of an evaluation.

The Delphi exercise appeared to work well. Its success was dependent on good evidence being available on a case-by-case basis to inform expert opinion, on the key question being relatively simple, and on the diversity and independence of the experts and their estimates.

The main weakness of the evaluation was that not every uncertain parameter could be estimated *de novo*. There simply was not enough time or money to do that. There were several parameters, including the value of a "statistical life" and the appropriate discount rate that were taken from the literature despite the lack of consensus. This meant that a wide probability distribution for the value of each parameter had to be used, and this added to the residual uncertainty at the end of the cost-benefit analysis. It would improve the accuracy and consistency of such evaluations if the central agencies of the Government of Canada, assisted by departments with a particular interest in each topic (for instance, Health Canada in regard to the "value of life") were to publish consensus figures for parameters that are common to many cost-benefit evaluations.

REFERENCES

Adler, M., & Ziglio, E. (Eds.). (1996). *Gazing into the oracle: The Delphi Method and its application to social policy and public health*. London: Kingsley.

Cutler, D. (2004). Pricing the priceless. In *Your Money or Your Life* (chap. 2). New York: Oxford University Press.

Gafni, A., O'Brien, B., & Diener, A. (1998). Health care contingent and valuation studies. *Health Economics, 7*, 313–326.

Harris, J. (1987, September). QALY-ifying the value of human life. *Journal of Medical Ethics, 13*(3), 117–123.

Hahn, R.W. (1991, Winter). The costs and benefits of regulation: Review and synthesis. *Yale Journal on Regulation,* 72–80. New Haven: Yale.

Linstone, H., & Turoff, M. (1975). *The Delphi Method: Techniques and applications*. New York: Addison-Wesley.

Morrow, R.H., & Bryant, J.H. (1995). Health policy approaches to measuring and valuing human life: Conceptual and ethical issues. *American Journal of Public Health, 85*(10), 1356–1360.

Newhouse, J.P. (2002). *Pricing the priceless*. Cambridge, MA: MIT Press.

RAND Corporation. (2003). *Delphi and long-range forecasting: A bibliography* (Rev. ed.). Palo Alto, CA: RAND Corporation.

Spetzler, C.S., & Von Holstein, C.-A.S. (1975). Probability encoding in decision analysis. *Management Science, 22*(3), 340–358.

SRI. (1978). *Manual for encoding probability distributions*. Menlo Park, CA: Stanford Research Institute.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economics, societies and nations*. New York: Doubleday.

Transport Canada. (2001). *Evaluation of rapid distress alerting equipment carriage requirements for non-SOLAS vessels*. Ottawa: Author.

Transport Canada. (2003). *Evaluation of rapid distress alerting equipment requirements for small commercial vessels*. Ottawa: Author.

Viscusi, W.K. (1993, December). The value of risks to life and health. *Journal of Economic Literature, 31*, 1912–1946.

Volokh, A. (1997). *n* guilty men. *University of Pennsylvania Law Review, 146*, 173–216.

Watson, K., & Mallory, C. (1997). *Benefit-cost analysis guide*. Ottawa: Treasury Board of Canada.

Dr. Kenneth Watson is a consulting specialist in evaluation, program strategy, and organization design. He studied economics and government at the University of British Columbia and Harvard University and has held visiting appointments at Oxford and the Australian National University. He received the Canadian Evaluation Society award for contribution to evaluation in 2002.