RFF REPORT

MARCH 2004

Valuing Health Outcomes:

Policy Choices and Technical Issues

Alan J. Krupnick

RESOURCES
FOR THE FUTURE

MISSION

Resources for the Future improves environmental and natural resource policymaking worldwide through objective social science research of the highest caliber.

As the premier independent institute dedicated exclusively to analyzing environmental, energy, and natural resource topics, Resources for the Future (RFF) gathers under one roof a unique community of scholars conducting impartial research to enable policymakers to make sound choices.

Through a half-century of scholarship, RFF has built a reputation for reasoned analysis of important problems and for developing innovative solutions to environmental challenges. RFF pioneered the research methods that allow for critical analysis of environmental and natural resource policies, enabling researchers to evaluate their true social costs and benefits.

RFF Reports address major issues of public policy in a manner designed to enrich public debate and meet the needs of policymakers for concise, impartial, and useful information and insights.

Core areas of knowledge at Resources for the Future include:

Energy, Electricity, and Climate Change
Environment and Development
Environmental Management
Food and Agriculture
Fundamental Research
Natural and Biological Resources
Public Health and the Environment
Technology and the Environment
Urban Complexities

RESOURCES FOR THE FUTURE
1616 P Street, Northwest · Washington, D.C. 20036-1400
Telephone: (202) 328-5000 · Fax: (202) 939-3460 · www.rff.org

Valuing Health Outcomes

Policy Choices and Technical Issues

Alan J. Krupnick

Board of Directors

Contents

TABLES AND FIGURES

IN MEMORIAM

The author wishes to dedicate this report to his friend and colleague Dr. Bernie O'Brien. Bernie was a professor of clinical epidemiology and biostatistics at McMaster University in Hamilton, Ontario. He died unexpectedly at the age of 44 in February 2004. He was one of the very few experts in the field of health valuation who was equally credible and insightful in addressing both the willingness-to-pay and the quality-adjusted life year approaches. Bernie's internationally recognized contributions to the field of health economics will be deeply missed.

Acknowledgments

The author would like first to provide a special thanks to the agencies and the representatives on the Interagency Steering Committee who provided generous support for this paper and comments on the OMB draft guidelines, the RFF conference, and the workshop. These agencies and representatives include: EPA (Kelly Maguire and Chris Dockins, NCEE), FDA (Clark Nardinelli, CFSAN), AHRQ (Bill Lawrence), NIH (Jim Shuttinga, OSP), USDA (Elise Golan, ERS), and CDC (Scott Grosse, Phaedra Corso, and Laura Blanciforti). Other members of the Steering Committee who were all central to shaping all the aspects of this project were Joseph Lipscomb, NCI/NIH, Debbie Aiken, OSHA, Jim Schaub, ORACBA/USDA, Peter Belenky, OTP/DOT, and Warren Prunella, formerly of CPSC. Thanks also go to the expert participants at the "Valuing Health Outcomes" conference and at the follow-up workshop. Those who did double duty include: Maureen Cropper, University of Maryland and World Bank; Bernie O'Brien, McMaster University; Dan Brock, NIH; David Feeny, University of Alberta, Institute of Health Economics; F. Reed Johnson and Josephine Mauskopf, Research Triangle Institute; James Hammitt, Harvard University, Center for Risk Analysis; Kerry Smith, North Carolina State University; and Ted Miller, Pacific Institute for Research & Evaluation. Needless to say, the views expressed in this report do not necessarily reflect the views of any of the above participants, nor the agencies and institutions they represent. Here at RFF, I wish to thank Mike Taylor, who as co-chair of the Steering Committee and a former "decisionmaker" at FSIS kept the group firmly grounded in the practical aspects of policymaking and offered wise counsel throughout the process. A special thanks also goes to Meghan McGuinness, who has been my sidekick throughout this project, the main conduit between the Steering Committee and Mike and me, and the research assistant performing both substantive and editorial work on all written products. David Feeny deserves special mention as he provided detailed comments on this manuscript in its final iteration. Finally, I would like to thank the RFF staff members, including Karin Warner and Michael Hensler-McGinnis, who made the events and the website happen.

Executive Summary

Regulatory and nonregulatory activities by governments at the federal, state, and local levels often affect the economic and physical health of their constituents. The effectiveness, efficiency, and distributional implications of such activities are often compared through cost–benefit analysis and cost-effectiveness analysis. These analytical tools can also be used to help design policies. Their use requires measures that track and aggregate expected or realized health outcomes.

The government's choice of analytical tools varies. Some statutes (such as the Clean Air Act) are silent on the subject of cost–benefit analysis (which the Supreme Court has recently affirmed means that it cannot be used to make certain types of decisions), while others (such as the Safe Drinking Water Act) mandate its use. Nevertheless, all federal agencies are required to evaluate the costs and benefits of any regulation expected to cost $100 million or more.

The choice of health outcome measure is as varied—and as controversial. The Environmental Protection Agency (EPA) uses monetary measures of the value of a statistical life (VSL), but the EPA's specific value differs from that used by the Department of Transportation. In contrast, the National Institutes of Health and the Food and Drug Administration tend to use health indices because these measures help rationalize and prioritize health interventions under their purview.

The Office of Management and Budget (OMB) is charged with overseeing the use of these tools and in September 2003 issued new regulatory guidance in Circular A-4. This document boosts the standing of cost-effectiveness analysis in Regulatory Impact Analyses (RIAs) carried out by agencies and promises more widespread use of health indices and health effect measures in describing outcomes.

This report sorts out the assumptions underlying the alternative analytical tools and health valuation measures and informs practitioners about the implications of their choices on the effectiveness, efficiency, and equity of the policies they evaluate. The research grew out of a February 2003 conference and an April 2003 workshop, cosponsored by Resources for the Future and a host of federal agencies, at which these tools and methods were discussed and compared.

Analytical Tools

Cost–benefit analysis (CBA)

CBA is a normative accounting technique for capturing the advantages and disadvantages of a policy in monetary terms: subtracting costs from benefits yields the net benefits to society. These net benefits are measured according to the efficient allocation of society's scarce resources and leave out by design other important but less quantifiable criteria, such as the distribution of these benefits across income classes. Policies that deliver negative net benefits are judged inferior to those that offer positive net benefits, and policies with the larger net benefits are considered superior to those with smaller net benefits, all other things being equal. While CBA focuses on aggregate measures, it can be used to track the distributional effects of a policy, such as by income group, sex, or race.

CBA offers transparency because the results can be clearly linked to the assumptions, theory, methods, and procedures used. It fosters revelation of uncertainty because its template permits the practitioner to see whether important information is missing. And it allows comparability because it attempts to capture in a single index all the features of a policy decision that affect the well-being of society, that is, in terms of the efficiency of resource allocation. In the context of this report, note that CBAs also routinely add valuation of nonhealth benefits to the analysis. This is important because what is being regulated, such as pollution, often affects more than human health.

In CBA, the monetary values of both benefits and costs are expressions of an aggregation of individual well-being, or utility. But a fundamental tenet of welfare-based valuation approaches is that one cannot make interpersonal utility comparisons and aggregate individual utility measurements. The solution embraced by CBA—the compensation criterion—evaluates policies on the basis of whether they could, in principle, lead to greater welfare: Can the winners from the policy fully compensate the losers and be at least as well off as before the policy? That such compensation hardly ever takes place, coupled with the difficulty of even identifying all winners and losers, leads to questions about using the compensation criterion to aggregate individuals' welfare into social welfare.

Applied to health policy, CBA attempts to capture preferences for alternative health states and put them in monetary terms so that they can be compared with other monetary estimates of the policy's effects. Some people, however, consider it unethical to place a value on human life. This view reflects a persistent misunderstanding about the valuation process. This process measures preferences of individuals, not researchers. These preferences are also for small changes in reducing death risks, some things people do every day—from deciding whether to have medical tests to deciding whether to cross the street against a red light to save a few seconds of waiting. However, because of technical difficulties and expense, some health endpoints are left unvalued. With benefits (and possibly costs) only partially estimated, the resulting net benefit estimates will likewise be incomplete.

Cost-effectiveness analysis (CEA)

CEA is a form of CBA in which the benefits are not monetized; therefore, net benefits cannot be calculated. Instead, one calculates regulatory costs per unit of an effectiveness measure, such as lives saved. Although CEA does not help determine whether a policy increases social welfare, like CBA, CEA can help identify the policy that achieves the specified goal with the smallest loss in social well-being, and it can help rank alternative policies—in this case, according to their cost-effectiveness. By avoiding monetization of health benefits, CEA may be less controversial than CBA, and, through the use of a single effectiveness measure, CEA can be simpler to conduct and communicate.

Those advantages come at a cost, however. The results can be misleading for social welfare, since the smallest welfare loss might not be associated with the smallest dollar cost. Also, CEA can unambiguously compare only policies that have a single outcome (such as lives saved), outcomes that move proportionally, or outcomes described by a health index. As a result, it addresses uncertainty less comprehensively and offers less comparability than CBA.

Cost-utility analysis (CUA)

Cost-utility analysis, as defined in this report, is the same as CEA except the effectiveness measure claims to represent (or be based on) utility, or individual welfare. It is used most often to consider the appropriateness of alternative medical interventions as well as to analyze health policy. Relative to CBA, the advantages and disadvantages of CUA are the same as for CEA except that CUA includes an effectiveness measure reflecting utility, albeit not in monetary terms.

Health Valuation Measures

Health indices, such as those using a quality-adjusted life year as the unit of account are based on multiplying the duration of a health state by a score reflecting the quality of the health state. Monetary measures include willingness to pay, which measures what individuals would be willing to give up to obtain health improvements, and cost-of-illness, which measures medical costs and forgone wages associated with health effects. A conversion factor to translate a health index outcome into a monetary measure, such as $/quality-adjusted life year, also appears in the literature and in practice.

Quality-adjusted life year (QALY)

The term “QALY index” was used to stand for all health indices. QALY indices are in use throughout the world, primarily to examine the effectiveness of medical interventions. Whether they are ready for and appropriate for use in any given policy setting are open questions.

The QALY approach uses the quality of a life year as the basic unit of account and aggregation. With dead represented by a score of zero and perfect health by a score of one, living five years longer would add five life years, subject to any adjustment for a less-than-perfect quality of life during those years. In general, numeric values are assigned to various health states to permit morbidity effects (such as severity and types of illness) to be combined with mortality effects (or likelihood of death) to develop an aggregated measure of health outcomes. For example, a year of extreme pain may be valued at 0.5. A basic assumption is that the QALY values are additive, so that a treatment eliminating extreme pain for one year for two individuals (2 x 0.5) is equivalent to a treatment that adds one healthy year of life. Life years are treated equally for all individuals, implying that a single healthy year is weighted the same regardless of age or income.

The health states are based on specific symptoms or general functionality, such as chronic limitations in one’s motor functions. A crucial decision for eliciting scores is whose opinions will be sought—experts, health care professionals, affected groups, or the general population.

Willingness to pay (WTP)

The WTP approach is based on the trade-offs that individuals make (or think they would make) between health and wealth. Such trade-offs in daily life are easily recognized and as easily quantified: we may take a riskier job if higher pay will compensate us for the greater risk. WTP health valuation studies attempt to make such preferences explicit by either uncovering the trade-offs people actually make (revealed preference), as in the job choice example, or presenting them with realistic but hypothetical choices and eliciting their preferences (stated preference). WTP studies are used around the world (and particularly in the United States) in cost–benefit analyses of government regulations.

Cost-of-illness (COI)

COI estimates typically include direct medical expenditures, forgone wages, and lost household services associated with illness and premature death. Also known as the human-capital approach, COI does not purport to be a measure of individual or social welfare, since it excludes such intangibles as pain and suffering. Its advantage is its simplicity. COI estimates generally supplement WTP estimates in cost–benefit analyses of government regulations where WTP estimates are missing.

$/QALY

Another approach used to analyze the benefits of policies (or more often medical treatments) is to monetize the QALY estimate of effectiveness. QALYs are converted to dollars generally using a single $/QALY factor and then the resulting monetary estimate of benefits can be used in a CBA. These numbers are generally taken from studies that place ceilings on the cost-effectiveness of various medical interventions. Others have attempted to use value of a statistical life (VSL) estimates to derive a value of a QALY. It was suggested at the conference that these approaches as they have been used to date are not theoretically sound, in part because individuals cannot be expected to have a constant rate of substitution between QALYs and wealth.

Policy-Level Choices

In choosing their metrics, practitioners and the policymakers receiving their work need to be aware of the assumptions, both explicit and implicit, that they are accepting. Some assumptions are not amenable to technical solutions; rather, they lie more in the domain of ethicists and philosophers. Chapter Three takes up these policy-level choices.

Efficiency

Efficiency—meaning better allocation of resources—has two dimensions in the context of regulatory analysis. The first is the normative dimension, that is, does the regulation generate positive public benefits and does this particular regulatory design generate the largest public benefits? The second is a relative dimension, that is, does this regulation rank highest in terms of incurring the lowest social cost per chosen measure of effectiveness? Under the first perspective, CBA using WTP measures of value are favored. Under the second perspective, the choice of type of analysis is unclear because both CEA and CBA rank alternative regulatory designs.

Equity

The tools themselves—CBA and CEA—have no equity implications. Equity concerns with respect to the valuation measures, at a minimum, may involve how age, health status, and income factor into the analysis. Health indices, unlike WTP measures, implicitly value extensions to younger people's lives and healthy people's lives more than life extensions for those who are older or infirm, primarily because of lower life expectancies in elderly populations and lower health status in infirm populations. Adjustments could be made to eliminate such biases, as has been suggested by OMB in Circular A-4 with respect to the bias against the disabled and ill. For their part, WTP measures are constrained by income, which may be regarded by some as unethical, although in practice, values averaged over all income classes are usually used in CBAs and therefore do not discriminate against any income groups.

Normative guidance or relative rankings

How much normative guidance does the decisionmaker want? CBA offers a possibility of rejecting a course of action on efficiency grounds, if it generates net social losses. CEA can provide this service only when net costs are negative. With CBA, policies can be ranked from the largest net benefit to the smallest. With CEA, they can be ranked from the smallest cost-effectiveness ratio to the largest.

Interpersonal utility

In adding up the gains (or losses) from a policy, aggregation over individuals is required. CUA assumes that individual utilities can be added by simply summing QALYs. Thus, for any given intervention, some will gain, others will lose, and the choice of intervention will be ordered for cost-effectiveness based on the net effect on QALYs and cost. Similarly, WTP measures assume individual values can be summed.

Individual versus social perspective

Welfare economics places individual preferences at the center of the "story," with government intervention to correct market failures. In this market story, consumers are sovereign, their preferences create demand for goods, and government should not interfere with this demand if it wishes to maximize social welfare—defined as the sum of individuals' welfare.

Whether individual preferences should be at the core of government activity is an open question. Individuals have preferences for what their own health states are; they also have preferences for what happens to others (in their household, in the community, and so on). Either or both of these objects of individual preferences may be important for decisionmakers to take into account. They may also want social preferences to play a role in their decisions, where such preferences are different from the sum of the individual preferences. This issue affects WTP and QALYs alike. WTP and QALY using the person trade-off approach—where respondents choose between helping individuals in one health state over those in another—can be designed to capture preferences for community health improvements.

Underlying the choice between an individual and social perspective may be, at least in part, beliefs about the reliability of individual preferences. Certainly, there is ample evidence of gaps between individual risk perceptions and scientific estimates of risk and other factors relevant to valuing health. Some of these differences may be cognitive difficulties, say in understanding probabilities, which would argue for reducing the weight given to individual preferences. Then again, arguing in the other direction is the well-known phenomenon that individuals imbue risk preferences with many qualitative attributes, such as degree of voluntariness and dread, which lie outside of the standard probabilistic treatment of risk. Both WTP and QALY approaches are conceptually indistinguishable on this perception issue, although this issue has received far more attention by economists in the WTP literature.

Health versus utility

What measure is appropriate to maximize in distinguishing among regulations—health or social welfare? If an aggregate measure of health changes associated with a regulatory design is the preferred measure of effectiveness, then QALYs are favored. Needless to say, describing changes in social welfare, not just health, would seem more important in a policy context, and this view would favor WTP measures.

Avoiding controversy

Using monetary measures of health effects such as WTP, particularly where premature mortality is at issue, can be more controversial than using either physical effects or QALYs (although conferees agreed that this controversy was unwarranted and is largely the result of a misunderstanding of what is being valued and how it is being valued). However, the use of physical criteria does not permit aggregation with other health or nonhealth effects. Using QALYs does eliminate the need to express benefits in monetary terms, but this merely postpones the problem, since eventually a decision must be made about whether to spend a given amount of money to save a specified number of QALYs.

Completeness

Developing WTP measures is more labor-intensive than developing QALYs because the latter approach provides weights for many different health states or domains in a given survey, while WTP studies yield values for at most a few health endpoints at a time. Attempts to develop benefit-transfer approaches to extend the range of health effects valued by WTP generally have not been successful, but choice experiment (conjoint analysis) techniques (where attributes of a health state are valued) may prove useful for this purpose. "Scores" for health states are generally more available than WTP values, although they tend to be for endpoints that are more detailed than those typically specified in epidemiological studies used in RIAs.

Credibility

A large part of Chapter Four is devoted to analyzing the credibility of the various valuation measures. The bottom line is that, under a welfare economics paradigm, WTP measures are theoretically more credible than QALYs. QALYs provide a valid utility measure only under very restrictive conditions. Within the WTP literature, much attention has been devoted to validating the credibility of WTP measures. Within the QALY literature, however, treatment of the issue of credibility (at least according to critics) has been far less extensive, though this is a flaw in the literature rather than the measure itself. Within the monetary measures, those based on willingness to pay using stated preference techniques are beginning to be viewed as more credible for use in evaluation of policy interventions related to health than those based on revealed preference techniques, either because of the paucity of the latter or because the market behavior being studied is too far removed from the policy context. Typically, WTP measures derived from revealed preference (RP) techniques are based on wage-risk tradeoffs, which differ in context from most environmental and health risks in terms of the populations and the nature of the risk at hand. Stated preference (SP) methods can be more easily tailored to the policy context, thus potentially providing more credible estimates for policy evaluation. A panel of QALY experts and others to be convened by the Institute of Medicine later in 2004 will be attempting to sort out the advantages and disadvantages of various health indices for use in the policy context.

Consistency

Consistency can be assessed along two dimensions: consistency of the approaches used in studies and consistency in values across studies. WTP measures tend to be reasonably consistent regarding approaches—the two approaches in use (SP and RP) are derived from the same economic theory, and both approaches have been standardized (and improved upon) within the literature. The QALY literature also recognizes a number of approaches to calculating health status scores, but these approaches do not descend from the same theoretical origins. These include standard gamble (SG), which asks individuals to give the probability of death that would make them indifferent to a certain health state; time trade-off (TTO), which asks how much time—that is, how many years of life—people would trade to forgo certain symptoms; and ratings scale (RS), in which respondents are asked to simply rate various conditions on a numeric scale. While SG is based on expected utility theory, TTO and RS are not.

Transparency

Perhaps the least transparent of the WTP measures is the VSL. VSL measures are frequently misinterpreted as representing a market value for human life, rather than their true meaning, which is a value of a *statistical* life, derived by aggregating individuals' willingness to pay for small changes in risk. The metric, then, is representative of choices and trade-offs individuals make (or say they would make) in the face of risk. WTP measures for acute effects are more transparent in that they are simply the average willingness to pay to avoid a case of an illness.

On one level, QALYs are more transparent than willingness-to-pay measures, as they are simply the product of two components: a health state score and its duration. What is less transparent about QALYs, however, is the information underlying the health state scores. Such scores can be taken from a number of different indices, which in turn are developed using a number of different approaches for deriving preference weights (RS, TTO, SG) that have implications for health state scores (just as using RP or SP methods for WTP has implications for VSL values).

Time preferences

Another issue is how to account for the incidence of costs and benefits over time. Discounting the future is appropriate because it is perceived that getting something today is worth more than getting it later. Both costs and benefits should be expressed in terms of present discounted value, which involves applying discount rates to future costs and benefits. QALY gains should also be discounted, depending on their time of realization.

Consensus among economists once was that the same rate should be used to discount both costs and benefits. However, some health economists have now concluded that there is no compelling basis for that approach. In the OMB guidelines, it is said that costs can be discounted using the real market rate of interest and benefits using the rate of time preference, generally from 1% to 5%.

Technical Choices

This report examines from a technical perspective the advantages, disadvantages, and underlying assumptions associated with QALYs, WTP, COI, and $/QALY measures. Comparisons between WTP and QALY measures are summarized in Table E-1, according to a list of attributes identified as desirable during the Valuing Health Outcomes conference. The criteria for judging these measures include different types of validity (criterion, context, convergent, construct, and content validity); comprehensiveness; ease of application; costs of developing estimates; how well uncertainty is addressed; whether averting behavior is captured; whether qualitative risk attributes are included; and whether these measures bias choices toward certain groups. This detailed treatment is provided because agencies tend to be familiar with and use one measure to the exclusion of the other.

Those looking for a clear winner in the WTP versus QALYs comparison will be disappointed. According to experts at the conference, WTP studies can provide reasonable and credible social welfare-based estimates of value for some health endpoints, but not for others. The labor market studies provide a particularly robust set of studies on the VSL, with a growing body of contingent valuation method (CVM) studies on this health endpoint, which is generally considered to be the most important.

The literature on health effects valued with QALY indices is extensive and covers a wide variety of health endpoints, although these do not necessarily match endpoints appearing in RIAs. These indices enjoy wide acceptance by the medical community for discriminating the efficacy of alternative medical interventions. Judging on the comprehensiveness of estimates, QALYs do better than WTP, generally because any given survey to develop weights covers many types of health states that can be repackaged as a particular health effect is redefined. In general, WTP methods only apply to one health effect at a time, but newer studies taking the choice experiment approach promise to develop "prices" for a variety of health attributes. This approach, also called conjoint analysis, asks individuals to choose among different attributes such as health states. However, another important comparison is between the credibility of the underlying weights on health states in calculating QALYs and the credibility of the WTP estimates for individual health conditions. On this dimension, the literature on refining WTP approaches to improve their credibility is more advanced.

Concerning the QALY measures in the RIA context, some of the health indices do not pass this test and there is ambiguity and controversy about the others. Specifically, using indices based on the standard gamble to develop weights are favored because they incorporate the notion of trade-offs and some notions of risk. On the other hand, indices based on the person trade-off approach to weighting also have some appeal, as this weighting approach is the only one that may address individual preferences for effects to the community, in contrast to the standard gamble, which is concerned with individual preferences for one's own health. Not surprisingly, most studies use the simpler, less defensible indices (as measures of utility) because they are easier to understand and use when scoring aspects of disease.

TABLE E-1

Comparisons of Health Valuation Measures for Technical Attributes

Attributes	Quality-Adjusted Life Year (QALY)	Willingness to Pay (WTP)
Criterion Validity		
Tested against conditions for preferences to represent utility	Key assumptions violated by individuals, but may perform better in the aggregate.	Performs well.
Comparison to actual choices	Standard gamble (SG) scores predict treatment choices	Concern over hypothetical bias for stated preference (SP) studies; difficult to make head-to-head comparisons of SP with actual choices
Context Validity	SG does fairly well in invoking trade-offs, but not in context of reduced health risks; person trade-off (PTO) reflects community-level choices; health domains/states defined on medical interventions may not match health outcomes relevant for policy interventions.	Performs well; however, most health valuation studies are for individual preferences rather than community preferences.
Convergent Validity	Differences in preference weights by approach; SG is the only utility-consistent approach and depends on cardinal utility assumption, but is insensitive to changes in health status.	Differences in revealed preference (RP) and stated preference (SP); SP has potential to better match choice context.
Construct Validity	Focus is more on testing validity of indices than validity of weights. Weights are sensitive to duration of effect, violating independence assumption. Difficult for people to make SG tradeoffs. Duration estimates often unreliable or ad hoc. Yet, QALY indices can predict medical consumption.	Performs well, except proportionality to scope/scale for contingent valuation method (CVM).
Content Validity	Critics charge little attention given to "weights" surveys except in the construction of health state descriptions. Proponents say there is extensive work on this topic.	Major thrust of SP literature
Comprehensiveness	More comprehensive than WTP, but for health only. Combines mortality and morbidity.	Less comprehensive than QALYs, but covers more than health; doesn't combine mortality and morbidity.
Ease of Application	Easy	Easy
Cost	Cheap to apply, but getting weights is expensive (though only a one-time effort).	Cheap to apply, but getting unit values is more expensive per endpoint than QALYs. Presumption is that measures have to be estimated for each health effect–duration combination and by context, but research approaches are changing.
Address Uncertainty in Weights (QALYs)/Prices (WTP)	Relatively little attention here, only in sensitivity analysis. Uncertainty in duration of health states not addressed.	Yes
Recognizes Avoidance Behavior	No	Yes
Inclusion of Qualitative Elements of Risk	Embedded in preferences to unknown degree.	Embedded in preferences to unknown degree; beginning to be an object of research.

WTP measures may be better than QALYs at capturing preferences regarding acute health effects and can, at least in theory, capture qualitative attributes of risk (voluntariness, dread, and so on) that are not quantified in standard surveys to derive weights used in QALY indices. WTP measures can also be applied in consistent fashion to nonhealth effects, such as the effects of emissions on both ecosystems and childhood asthma. For QALYs to incorporate nonhealth effects in a CEA, the effects would have to be monetized using WTP techniques and then subtracted from costs—a confusing hybridization.

Many of the shortcomings of the QALY literature could be remedied by using best practices or even reforming some practices. Two examples of remedies are better for describing uncertainty of the scores and performing studies to better test the credibility of the weights. The WTP literature expends great effort on these features, routinely reporting uncertainty of the WTP estimates and performing a variety of content validity tests.

Improvements in WTP methods as well as research that could integrate the two approaches also are addressed in this report.

Concluding Thoughts

Regulatory decisionmaking by agencies will always be complicated, as decisionmakers evaluate incomplete and uncertain information while seeking to meet their legislative requirements, respond to stakeholders, and take their own readings about what is best for society. The new, more complex RIAs that will be produced under Circular A-4 could lead to better decisionmaking if they produce multifaceted, more complete results.

More information is not always better, however, and unfortunately WTP and QALY measures cannot be unambiguously ranked in their usefulness for policy. Such rankings would depend on the policymaker's philosophical choices and on the relative weight given to the various technical criteria. Similarly, despite CBA's advantages in measuring and ranking regulatory outcomes, the gaps in valuation and other practical problems posed by CBA that are not present in CEA would discourage sole reliance on the former method. Researching to improve all the tools and measures, evaluating each new RIA for how it approaches these issues, and examining how decisions are influenced by RIAs before and after implementation of the new guidelines would make improvements in decisionmaking more likely.

■ ■ ■

CHAPTER ONE

Introduction

Regulatory and nonregulatory activities by governments at the federal, state, and even local levels affect the economic and physical health of their constituents. Analytical *tools* and concepts are available and are often used to evaluate and compare the effectiveness, efficiency, and distributional implications of these activities—namely cost–benefit analysis and cost-effectiveness analysis. These tools can be used either *ex ante*, that is, to help design the best policy or *ex post*, that is, to evaluate a previously implemented policy. Once the decision is made to use these tools, a further decision is needed about the *measures* to use for tracking and aggregating health outcomes. These measures range from a simple count of the number of deaths or cases prevented, or other single indicators of the scope of the improvement, to the uses of indices to aggregate over the variety of health effects that might accompany any particular course of action taken by the government. There are basically two types of indices: monetary measures and measures that use life years as their metric, adjusted for the quality of life over time. Some of these indices may not be consistent with both cost–benefit and cost-effectiveness analyses, so the choice of analytical tool partly determines the choice of health measure.

The choices of tools and health outcome measures are highly controversial. Some statutes (such as the Clean Air Act) forbid use of cost–benefit analysis to make certain types of decisions (for example, on the stringency of the National Ambient Air Quality Standards), while others (such as the Safe Drinking Water Act) mandate its use. Irrespective of statutory requirements, federal agencies are currently required under Executive Order 12866—and executive orders going back to administrations from Clinton to Reagan—to evaluate the costs and benefits of major regulations, defined as those expected to have a cost of at least $100 million. For agencies whose activities are not regulatory or whose regulations don't exceed the $100 million threshold, the choice of tools is a matter of agency culture. While the U.S. Environmental Protection Agency (EPA) typically performs cost–benefit analyses and may or may not supplement them with cost-effectiveness analyses, agencies such as the Occupational Safety and Health Administration (OSHA) do not perform cost–benefit analyses.

Controversies are probably even more intense and agency practice even more varied over the choice of the types and the specific values of the health outcome measures. Indeed, a particular agency may not even follow a consistent set of practices. For example, EPA tends to use a monetary measure of the value of a statistical life (VSL) based on a particular reading of the environmental economics literature, but this measure differs from that used by the Department of Transportation (DOT), which reads this literature differently and uses it in a different regulatory context. The National Institutes of Health (NIH) and the Food and Drug Administration

(FDA), agencies whose culture is influenced by the health services community that they serve and regulate, tend to use the health indices, because these measures are used in a medical-and-health-insurance setting to help rationalize and prioritize health interventions for insurance support and for development of standard medical practices.

The Office of Management and Budget (OMB) is charged with bringing some order to this process, primarily through implementation of the executive order noted above and its guidelines on the conduct of cost–benefit analysis. Recently, these efforts have been significantly augmented with OMB's new regulatory guidance document (OMB 2003).[1] It calls for expanded use of cost-effectiveness analysis in agency Regulatory Impact Analyses (RIAs) and promises a more frequent role for the health indices and physical health effect measures in describing health outcomes.

The purpose of this paper is to sort out the assumptions underlying the use of alternative tools and health outcome measures in the analysis of government activities and, in so doing, provide guidance to government practitioners and policymakers about the implications choices of these tools and measures have for the effectiveness, efficiency, and equity of the activities they evaluate. Thus, the intended audience is the practitioner of analyses for governments and the experts and others who have a need to understand such analyses.

Readers of this report should be aware of how the topics covered fit into the overall role of government in protecting the health of its citizens. The analytical approaches discussed here have, in the best of circumstances, a fairly minor supporting role in a play that involves the legislative, executive, and judicial branches, as well as the public. The play begins with legislation providing the broad (and occasionally quite specific) framework for health protection. Then agencies use their discretion in implementing this framework, taking into account all the factors that appear important in the scientific and political processes in which they are embedded. Analyses of the types discussed here—cost–benefit analysis (CBA) and cost-effective analysis (CEA) involving health outcomes—can inform and support these legislative and agency activities but are certainly not substitutes for them.

This paper is a product of two major meetings and much review by government practitioners and policymakers, as well as experts, both within and outside government. The first meeting—a conference titled "Valuing Health Outcomes"—was held February 13–14, 2003, at the Resources for the Future Conference Center in Washington, D.C. It involved some 25 speakers and 175 participants who contributed a great deal of information describing the conceptual and empirical foundations of health valuation measures and their use in various analytical tools. The second meeting was held April 29 and involved a small subset of this group in detailed discussions leading to the creation of this paper. While this paper relies heavily, indeed almost exclusively, on the discussions and writings of the participants at these meetings, and has been reviewed by many of them, this paper is nevertheless the responsibility of the author. This paper is a product of RFF, not of the Interagency Steering Committee.

The paper is organized as follows: Chapter Two defines the key tools and valuation measures that are the subject of the paper. Chapter Three identifies the "policy choices" implicitly made with the use of these tools and measures, while Chapter Four addresses the technical issues underlying the use of particular health valuation measures. This distinction between policy choices and technical issues is fundamental to the report. Policy choices are not amenable to analytical arguments and further research per se, but require choices made at the highest levels of an agency about the fundamental values that will undergird its analytical efforts. An example of policy choices is the metric to be maximized—whether health outcomes, social welfare, or some other measure. Technical issues include the assumptions, advantages, and disadvantages associated with particular valuation measures, including, most importantly, the conceptual and empirical validity of the measure to describe what it purports to describe. Such issues, by their nature, are capable of being informed by evidence and logical arguments. Chapter Five presents and comments on the recent OMB guidance (Circular A-4, OMB 2003). Chapter Six lists research topics flowing from the report. Finally, Chapter Seven summarizes the differences in the tools and measures from the decisionmaker's point of view.

CHAPTER TWO

Key Concepts

This chapter begins with a definition and explanation of analytical tools. We consider three: CBA, CEA, and cost-utility analysis (CUA). For health outcome measures, we focus on quality-adjusted life years (QALYs), monetized QALYs, cost-of-illness (COI), and willingness to pay (WTP). These terms are further defined in Appendix I.

Cost–Benefit Analysis

CBA is essentially a normative accounting technique for capturing the advantages and disadvantages of a course of action in monetary terms. Subtracting costs from benefits yields the net benefits to society or net improvements in social welfare of a course of action. Policies that reduce welfare or well-being are a priori (and other things being equal) inferior to those that improve well-being, and policies that improve well-being a great deal are superior to those that improve it only marginally. Conceptually, then, CBA could be used to cardinally rank policies on the basis of their improvements or reductions in well-being. CBA focuses on the aggregate measures of well-being, taking the existing distribution of income as given. CBA is capable of tracking the effects of policy by income group, race, and other factors, but the distribution of these effects does not figure in the ranking process.

The advantages of CBA are in its transparency, its ability to reveal areas of uncertainty, and its comparability. CBA offers transparency because the results of a well-executed cost–benefit analysis can be clearly linked to the assumptions, theory, methods, and procedures used in it. This transparency can add to the accountability of public decisions by indicating where the decisions are at variance with the analysis. It fosters revelation of uncertainty because the template character of CBA permits the decisionmaker to determine the adequacy of the information collected and see whether important information is missing. This knowledge can provide the decisionmaker with valuable insight into the level of uncertainty regarding important attributes of the policy. CBA fosters comparability because it attempts to capture in a single index all the features of a policy decision that affect the well-being of society. The single-metric approach permits the comparison of policies that affect different attributes of well-being differently, that is, it permits the decisionmaker to compare "apples" and "oranges" on the basis of a single attribute (in this case, the index of social welfare) common to both.

As noted, in CBA, the monetary values of both benefits and costs are expressions of an aggregation of individual well-being. Moving from individual to societal valuations is a controversial step. A fundamental tenet of welfare-based valuation approaches is that one cannot make interpersonal utility comparisons. An implication of this tenet is that it is impossible to provide an unambiguous means of aggregating individual utility measurements. More than 50 years ago, the economist Kenneth Arrow proved an "Impossibility Theorem" stating that no simple representation of total social welfare—additive or otherwise—simultaneously satisfies a number of intuitively desirable properties (Arrow 1951). Any approach that attempts to measure welfare as the summation of individual utilities must confront this issue.

The solution embraced by CBA—the compensation criterion—evaluates policies on the basis of whether they could, *in principle*, lead to greater welfare. This is said to occur if the winners from the policy could provide full compensation to the losers and be at least as well off as before the policy. However, there is no requirement that compensation actually occur. The fact that such compensation hardly ever takes place, coupled with the difficulty of even identifying specific winners and losers from large policy changes, leads to questions about the credibility of the compensation criterion as a satisfying fix to the aggregation issue.

Some people feel that it is unethical to place monetary value on health, but this view reflects a misunderstanding in the valuation process.

Concerning applications of CBA to policies specifically affecting health, the analysis attempts to capture preferences for alternative health states and put them in monetary terms to be commensurable with other monetary estimates of the policy's effects. One additional feature of CBA that *in practice* distinguishes it from some of the tools to be discussed below (CEA/CUA) is that CBAs routinely add valuation of nonhealth benefits (or costs) to the analysis. This feature is becoming increasingly important as, for example, recognition grows of the interconnectedness of pollution and the environment. In principle, CEA/CUA could incorporate such effects as well (by, for example, subtracting monetized nonhealth benefits from costs in a cost-effectiveness analysis), but this is not standard practice.

Aside from conceptual issues with the aggregation of preferences—an issue shared or ignored by the other tools discussed below—CBA has other disadvantages. Probably its main disadvantage is that it seeks monetization of all effects of a policy. Some people feel that it is unethical to place monetary value on health or mortality risk changes because it "places a value on human life." As shown below, this view reflects a misunderstanding about the valuation process, but this misunderstanding is widespread and durable. A related problem with CBA is that the technical difficulties and expense in obtaining such values mean that some health endpoints will be left unvalued. With benefits (and possibly costs) only partially estimated, the resulting net benefit estimates will likewise be incomplete. Of course, the partial estimation of benefits, for instance, may not matter for policy if even with a partial estimate benefits exceed costs.

Cost-Effectiveness Analysis

CEA is a particular form of CBA, where the benefits are not monetized, and therefore, net benefits cannot be calculated. Instead, one calculates costs per unit of an effectiveness measure (such as lives saved). Therefore, while CEA cannot help in determining whether a policy increases social welfare, it can help in the choice of policy that achieves the specified goal with the smallest loss in social well-being and can help rank alternative policies according to their cost-effectiveness.

The advantages of CEA over CBA then are that it can avoid monetization of health benefits (although in a "net cost-effectiveness analysis" any monetized benefits of a policy are subtracted from costs) and, through the use of a single effectiveness measure, can be simpler to conduct and communicate.

These advantages come at a cost, however. The results of a CEA can be misleading for social welfare, as the smallest welfare loss might not be associated with the smallest dollar cost. Said another way, CEA does *not* imply choosing the policy with the smallest dollar price tag (although many people believe that it does). Also, CEA can only unambiguously compare policies that have a single outcome (such as lives saved), that have outcomes that move proportionally, or that have outcomes described by a health index. Thus, CEA provides less revelation of uncertainty and provides less comparability than CBA. Because CEA is a simpler analysis than CBA, it may be more transparent in this sense but less so in that the health effects may not be clearly linked to assumptions and an underlying conceptual model. Important effects may not be weighted by public preferences.

Cost-Utility Analysis

Cost-utility analysis is the same as CEA except the effectiveness measure has a claim to represent utility, or well-being. As noted below, not all health indices make this claim. It is used most often to consider the appropriateness of alternative medical interventions as well as in health policy analysis (see Gold et al. 1996, Haddix et al. 2003, and Drummond et al. 1997). In these applications, it is sometimes used in comparison to a benchmark cost-effectiveness value. Thus, if the utility-based effectiveness measure is from the Health Utilities Index and the cost-effectiveness of medical interventions A and B are $5,000 per quality-adjusted life year (QALY) and $50,000 per QALY, respectively, and the benchmark is $25,000 per QALY, intervention A would pass this test but B would be judged too expensive.

The advantages and disadvantages of CUA (relative to CBA) are the same as those of CEA except that CUA can lay claim to including an effectiveness measure reflecting utility, albeit not in monetary terms.

Effectiveness Measures

There is no conceptual restriction on the effectiveness measures that can be used in cost-effectiveness analysis and, in fact, those *in use* range widely, including, in the air pollution area alone: reductions in emissions, concentrations, exposures, health effects, lives lost,[2] life years lost,[3] and a health index. The choice of effectiveness measure could be based on the dominant effect of concern in a particular regulatory setting, since only one effectiveness measure can be used in the denominator of the cost per unit effectiveness calculation, or a major effect that cannot be monetized could be used in the denominator. In the latter case, the monetized effects can be subtracted from the cost term in the numerator and then divided by the nonmonetized effect, creating a "net cost-effectiveness" measure. Finally, a health index could be used when more than one health outcome needs to be added together for the effectiveness measure. As noted above, if the health index is derived from weights on health states consistent with utility, then the CEA would be called a CUA.

Because the conceptual basis for defining and measuring "costs" is the reduction in welfare associated with some activity, cost-effectiveness measures will generally involve mixing the welfare measure in the numerator—costs—with a nonwelfare measure, such as lives saved, in the denominator. This measure or other measures of effectiveness may be related to changes in welfare or satisfaction, but not without placing very restrictive conditions on the preferences of individuals (see Chapter Four).

Health indices

The health indices in the literature generally share the common characteristic that they are based on the product of the duration of a health state and a score reflecting the quality of the health state. Scores (weights for the index) can be derived from several approaches. The surveys use one of four elicitation approaches: rating scales (RS), such as the visual analog scale (VAS); time-trade-off (TTO); person-trade-off (PTO); and standard gamble (SG) to elicit population average preferences for health states and derive weights for these states (Kaplan et al. 1993). In this paper, we use the term quality-adjusted life year (QALY) to describe all of these indices.

The QALY approach uses the quality of a life year as the basic unit of account and aggregation. With dead represented by a score of zero and perfect health by a score of one, living five years longer would add five life years, subject to any adjustment for a less than perfect quality of life during those years. In general, numeric values are assigned to various health states so that morbidity effects can be combined with mortality effects to develop an aggregated measure of health outcomes.[4] For example, a year of extreme pain may be valued at 0.5. A basic assumption is that QALYs are additive,[5] so that a treatment that eliminates extreme pain for one year for two individuals (2 x 0.5) is equivalent to a treatment that adds one healthy year of life. Life years are generally treated equally for all individuals, so that a single healthy year is weighted the same regardless of age or income.

Underlying the health states is a set of domains of health. These domains can be based on general functionality or specific symptoms. An example of a set of domains from Gold et al. is in Table 2.1.

The class of indices described by the term "quality-adjusted life years" as used in this paper encompasses both preference-based and nonpreference-based summary measures of health. Among these are the Health Utilities Index (HUI, see Torrance et al. 1995, 1996; Furlong et al. 2001; Feeny et al. 2002), EuroQol or EQ-5D (EuroQol Group 1990), the Functional Capacity Index (MacKenzie et al. 1996), the disability-adjusted life years (DALYs) index, the Years of Healthy Life Scale (Erickson et al. 1995) and others.[6] The variations in these approaches have to do with the methods used to elicit the weights to be assigned to various health states (see Table 4.3, Gold et al. 1996) and the specifications of the domains underlying the health states.[7] A key decision for eliciting weights is whose will be sought—experts, health care professionals, affected groups, or the general population.[8]

The survey methods for arriving at weights for health states are not perfect substitutes. The rating scale (RS) is seemingly the simplest method for estimating weights.[9] Individuals are given a description of a health state and asked to rate it on a numeric scale. One type of rating scale is a visual analog scale (VAS), also called a "feeling thermometer." Respondents are shown a picture of a line with one endpoint representing dead and the other perfect health. They indicate their evaluation of where the described health state ranks by placing a mark on the line. Rating

TABLE 2.1

Domains Used in QALY Indices

Concepts and Domains	*Indicators*
Health Perceptions	Self-rating of health, health concern, health worry
SOCIAL FUNCTION	
Social relations	Interactions with others, participation in the community
Usual social role	Acute or chronic limitations in usual social role (major activities) of child, student, or worker
Intimacy/sexual function	Perceived feelings of closeness, sexual activity and/or problems
Communication/speech	Acute or chronic limitations in communication/speech
PSYCHOLOGICAL FUNCTION	
Cognitive function	Alertness, disorientation; problems with reasoning
Emotional function	Psychological attitudes and behaviors
Mood/feelings	Anxiety, depression, happiness, worries
PHYSICAL FUNCTION	
Mobility	Acute or chronic reduction in mobility
Physical activity	Acute or chronic reduction in physical activity
Self-care	Acute or chronic reduction in self-care
IMPAIRMENT	
Sensory function/loss	Vision, hearing
Symptoms/impairments	Reports of physical and psychological symptoms, sensations, pain, health problems or feelings not directly observable; or observable evidence of defect or abnormality

Source: Gold et al. 1996, 95. (Adapted from Patrick and Erickson 1993) scales are an example of a *psychometric* approach to determining preferences. Respondents are not asked to make trade-offs, nor are they asked to make decisions under uncertainty.

The time-trade-off (TTO) approach asks respondents to make trade-offs between outcomes, where the outcomes occur with certainty. They are asked how many healthy years of life they would be willing to give up to forgo specific symptoms. If an individual is indifferent between living an additional 25 years with the symptoms and 20 years without the symptoms, the QALY rating is determined from the ratio of these two values (4/5), with an adjustment to account for discounting effects. Both practitioners and critics appear to agree that the TTO approach (as well as the SG approach discussed below) is better suited to the evaluation of chronic conditions than acute symptoms. Imagine the difficulty of identifying the length of time one would give up in one's life to forgo a headache!

The person trade-off (PTO) approach asks respondents to choose between helping a number of individuals in a certain health state and a different number of individuals in another health state. The approach varies the number of individuals in one of the classes, elicits the point at which the respondent is indifferent between the two choices being offered, and derives a weight from this point (Murray 1994). The PTO can be framed in terms of saving the lives of different numbers in the two groups, which is the approach used to elicit disability weights in constructing DALYs (Murray and Lopez 1996) or in improving the health of two different groups (Nord 1999). From a policy perspective, this method seeks information analogous to that required as a basis for policy decisions (Kaplan 1995). It is analogous to WTP studies that define the benefit in terms of the community rather than the individual.

The standard gamble (SG) approach incorporates trade-offs and uncertainty over health states and is based on expected utility theory because respondents are asked to make choices under uncertainty. Respondents are asked for the probability of death that would make them indifferent to having the described condition. They are sometimes given the option of remaining in the condition or undergoing a treatment with a probability *p* of attaining perfect health and a probability (1-*p*) of death. However, because some people may be distracted by or turned off by the mention of a treatment, the framing of the questions used to elicit scores usually omits reference to treatment and focuses on "*n*" years in the intermediately ranked state for sure versus a lottery with probability *p* of the more favorable outcome and 1-*p* of the less favorable outcome.

The choice of *p* provides a *cardinal* measure of the utility of being in the described condition. The assumption that utility can be measured cardinally is more restrictive than that underlying WTP measures, which require only an ordinal measure. A number of other assumptions are also needed for this measure to represent utility. More will be said about these in Chapter Four. According to Neumann et al. (1997), only 10% of QALY applications score health states using weights derived from the SG approach.

Protocols for the appropriate conduct and application of cost-effectiveness analysis with QALYs are contained in Gold et al. (1996) who make recommendations towards the establishment of a "Reference Case." (See, in particular, Gold et al., page 122, which lists ten recommendations for the components of the Reference Case.) One element, for example, is the recommendation that sensitivity analysis be used where an analysis is affected by characteristics such as age, race, or gender in ways that might be "ethically problematic."

Monetary Valuation

The monetary value of health improvements can be estimated in two broad ways: (i) through measures of what individuals would be willing to give up to obtain health improvements, for example, willingness to pay (WTP) and, less commonly, what individuals would be willing to accept for a health decline (WTA); and (ii) through measures of monetary outlays and forgone wage compensation—termed the cost-of-illness (COI) approach. Another potential method of estimating monetary values is through considering jury awards. As such awards address specific individuals (rather than the nameless individuals usually covered by social policy), take an *ex post* perspective (rather than the *ex ante* perspective appropriate to policy actions), may be more to compensate the household for their suffering (that is, a willingness to pay of the household to reduce the risk of one of its members dying), and are also subject to limitations and special forces associated with the courtroom setting, they are not discussed further.[10]

Willingness to pay

The WTP approach is based on the trade-offs that individuals must make between health and wealth or income (or other goods). Such trade-offs in daily life are easily recognized and sometimes observed. For example, if we are running late to a meeting we may drive faster, knowing that the increased speed carries with it a slightly increased chance of accident and possibly death. Or we may take a riskier job if we know the pay will be higher to compensate us for the greater risk (or the converse: we may be content with a less risky job making lower wages).

WTP values can be divided into those measuring preferences for reductions in the risk of death and those measuring preferences for reductions in morbidity. The resulting estimates of the WTP for mortality risk reductions are converted to a value of a statistical life (VSL) by dividing the WTP by the risk change being valued. Morbidity can be divided into acute effects and incidence of chronic disease. For valuation purposes, the acute effects are usually modeled and estimated as though they are certain to be avoided, whereas the chronic effects are usually treated in the same way as for mortality—probabilistically—that is, as a reduction in the risk of developing a chronic disease.[11] Values to reduce acute effects, the probability of chronic effects, and the probability of premature death are usually added up (with some minor adjustments to avoid obvious double counting).[12] In contrast, the QALY approach integrates all types of effects, in the same simple algorithm, that is, multiplying the preference weight on each type of effect by its duration and adding the products (although there is considerable agreement that most health indices do not handle minor health effects very well).

Some WTP studies of acute and chronic effects explicitly incorporate measures of severity and average duration; others leave these measures implicit but ask subjects to describe the nature of the health effects they value. Beyond the direct effects of the illness, there are less obvious benefits that may or may not be measurable, for example the value of reductions in anxiety about getting sick or the value of reduced effort needed to avert risk and the associated health effects.

WTP and WTA health valuation studies attempt to make preferences explicit, either by uncovering the trade-offs people actually make—revealed preference (RP)—or presenting them with hypothetical choices—stated preference (SP). The revealed-preference method involves examining behavior, either in the marketplace or elsewhere, to discern willingness to pay. There are a wide variety of revealed-preference approaches used. The most developed techniques for estimation of health and mortality risk reduction benefits are probably the hedonic-labor-market approach and the property-value approach. Under the stated-preference approach, two approaches are in use. Contingent valuation (CV) studies pose questions about the willingness to pay or willingness to accept compensation for a change in risk of an adverse health outcome. A newer alternative to CV is conjoint analysis, which is used extensively in marketing to elicit preferences for combinations of product attributes. When such analyses involve the attribute of a price, the value placed on other attributes can be estimated.

SP and RP methods have been most extensively used to estimate WTP for reductions in risks of death. The SP methods involve placing people in realistic, if hypothetical, choice settings and eliciting their preferences. In CV surveys, individuals are not asked how much they value life, because WTP to avoid certain death is limited by wealth, while WTA could be infinite. However, as has been observed in many cases, people are willing to make trade-offs between marginal changes in risk and wealth. These choices might involve alternative government programs or specific states of nature, such as a given reduction in one's risk of death in an auto accident associated with living in one city instead of another, riskier city (see Krupnick and Cropper 1992) or choosing between two bus companies with different safety records when deciding to ride a bus (Jones-Lee et al. 1985). Therefore, attempts are made to ascertain WTP to reduce the chance of death by some small probability. Framing the question in this way highlights an important point: a WTP estimate for mortality risk reduction does not provide an inherent value for human life. Rather it illuminates the choices and trade-offs that individuals are willing to make and converts those choices into a value of a statistical life (VSL) by aggregating over many people their WTP for small changes in risk.

Imagine the difficulty of identifying the length of time one would give up in one's life to forgo a headache!

The most common RP approach—and the approach whose studies have traditionally undergirded VSL estimates used by the government in cost–benefit analyses—is the hedonic-labor-market approach. This approach involves estimating the wage premiums paid to workers in jobs that have high risks of death (Viscusi 1992, 1993, Viscusi and Aldy, 2003).

Calculating the implied value of health outcomes from WTP studies is usually straightforward. Using the damage-function approach, the unit values for the different endpoints are multiplied by the expected change in the incidence of the effect, taken from physical response functions in the literature. However, it is also possible to determine total WTP without going through the step of applying values to expected outcomes.[13]

An important issue bearing on the validity of monetary valuation is its applicability to the context in which it is used. Most studies are site-specific and coverage of all possible sites and situations is impossible. Therefore, it is often necessary to transfer the results of a study that focuses on one specific situation to another study with a different location or setting of interest. This procedure is known as benefit transfer, and there are occasions when the reliability of valuation estimates can be questioned. For example, hedonic wage studies provide VSLs based on accidental deaths of prime working-age individuals. It can be argued that this context is inappropriate for estimating the benefits of pollution control, where older and ill individuals are most at risk.

As with QALYs, there is a vast literature and supporting pronouncements from expert committees on appropriate protocols in the area of estimating WTP for health outcomes. The National Oceanic and Atmospheric Administration (NOAA) Panel (Arrow et al. 1993), made up of several Nobel laureate economists, survey researchers, and others, developed recommendations about how to conduct credible stated preference studies on the valuation of natural resources, recommendations that generally carry over to health valuation.[14] Major books and articles on WTP methods include Freeman (2003); Mitchell and Carson (1989); Carson, Flores, and Meade (2001); Carson et al. (1996); Cummings et al. (1986); Alberini and Krupnick (2003); and Champ, Boyle, and Brown (2003). In addition, a variety of computer models and modeling efforts have codified the health valuation literature. See, in particular, U.S. EPA (1997, 1999), Rowe et al. (1995), Farrow et al. (2001), and European Commission (1998).

Cost-of-illness (COI)

Cost-of-illness estimates typically include direct medical expenditures and forgone wages associated with illness and premature death. Often, the value of lost household services is included as well. This approach, also known as the human capital approach, does not purport to be a measure of individual or social welfare, since it makes no attempt to include intangible but real losses in well-being, such as those associated with pain and suffering. Its advantage is that it is a relatively simple to calculate and understand. Historically, this is an important approach used to calculate monetary costs associated with illness and death. The U.S. Department of Agriculture (USDA) and the Centers for Disease Control and Prevention (CDC), in particular, feature this measure in their cost–benefit analyses (for example, Buzby et al. 1996). And the USDA has recently issued a Cost of Illness Calculator (Economic Research Service 2003) for application to foodborne illnesses. Cost-of-illness measures are generally at least several times lower than WTP measures for the same health effect, because of their exclusion of intangible values (Kuchler and Golan 1999). There is no conceptual reason for COI estimates to be lower than bounds on WTP measures however.

Dollars per QALY

Another technique used to analyze the benefits of policies (or more often medical treatments) is to monetize the QALY estimate of effectiveness. QALYs are converted to dollars generally using a single $/QALY factor and then the resulting monetary estimate of benefits can be used in a CBA. Several researchers have attempted to develop $/QALY factors (Mauskopf and French 1991; Gyrd-Hansen 2003). An alternative approach is to use a set of conversion factors, tied to the particular composition of health effects embedded in the QALYs being estimated. The latter, more complicated, approach could be seen as functionally equivalent to using a willingness-to-pay model, as it implies that a specific factor ("price") would be developed for each type of health effect.[15]

Conversion factors appearing in the literature range from $25,000 to $100,000 or more. These numbers are generally taken from studies that place ceilings on the cost-effectiveness of various medical interventions. For example, if an intervention costs more than $50,000 per QALY gained, it could be judged as inefficient or ineffective (Gyrd-Hansen 2003). These cost-effectiveness benchmarks have often been based on cost-effectiveness ratios for dialysis and treatment for hypertension (Mark et al. 1995; Siegel et al.1996). Others have attempted to use VSL estimates to derive a value of a QALY. Hirth et al. (2000) derive QALY values from four types of VSL estimates (human capital, contingent valuation, revealed preference/job risk, and revealed preference/nonoccupational safety). They find that human-capital studies yield the lowest median value of a QALY estimate ($24,777), while revealed preference/job risk, studies yield the highest ($428,286). Note that James Hammitt of Harvard University, in a very recent presentation at the American Economics Association Annual Meeting, held January 3–5, 2004, examined the conditions under which the WTP/QALY ratio would be a constant and found them very limiting.

■ ■ ■

CHAPTER THREE

Policy-Level Choices

In interpreting results with different metrics, decisionmakers need to be aware of the assumptions, both explicit and implicit, that they are accepting. In this chapter, we consider assumptions that are not amenable to technical arguments by practitioners and experts in economics, health science, and similar disciplines, but may be more in the domain of ethicists and philosophers.

Efficiency and Equity

The two broadest effects of a policy are efficiency and equity, that is, whether the allocation of society's scarce resources is enhanced by the policy and whether the distribution of these effects across relevant dimensions such as income groups or race is viewed as equitable.

The practice of CBA is very clear about splitting these two concepts. The analysis itself addresses efficiency by computing net benefits as an absolute measure, while one can do further analysis to describe the equity effects along appropriate lines, such as race, income, gender, age, and so on. With this description, independent judgments can be made about whether a particular policy is equitable or more equitable than another policy.

With CEA, the metric for measuring efficiency is only relative rather than absolute (net benefits). The distribution of costs and the effectiveness measure among gainers and losers could presumably be described as in CBA. However, this is not usually the practice with CEA. Of course, as for CBA, much of the postanalysis discussion of a CEA involves the distributional effects.

Maximizing Health or Utility

What measure is appropriate to maximize—social welfare (utility) or health (in light of costs)? Utility obviously includes health, but also captures other effects of a policy, including nonhealth effects and "qualitative attributes" of health states such as the dread associated with cancer.

A decision to use CBA with a WTP measure implies a focus on utility, although whether the qualitative attributes are included depends on the specific estimation approach. Using CEA with a physical-effects measure obviously will not represent social welfare effects. CEA with a health index has less clear implications for the choice of utility versus health (as we shall see in Chapter Four). If QALYs don't represent utility, then choosing a QALY measure implies embracing what Brouwer and Koopmanschap (2000) call the "decisionmaker's approach, " or, less judgmentally, the "extra-welfarist" approach. This approach is described as maximizing health improvements from a given budget.[16] As for WTP measures, depending on the method used to measure health, these qualitative attributes may well be reflected or included in the utility weights. For instance, a multiattribute measure might capture such anxiety. More relevantly, the SG or some other direct-preference measure may well capture these concerns.[17]

Another issue affecting the choice of measure is whether purchasing power or life years should be the preferred standard for comparing the gains and losses of a policy. In the first instance, a preference for purchasing power would favor WTP while a preference for life years would favor this effectiveness measure, as well as any of the health index measures (where life years are a major component). A preference for purchasing power is justified, according to Hammitt (2003), because purchasing power drives many of our social decisions as they play out in markets, particularly in markets for health and even the allocation of some government services. Yet, as he notes, there are many areas of public life where we use other standards, such as decisions over abortion or prayer in the schools. And, even when we use purchasing power to allocate public goods, we sometimes ignore how purchasing power varies by income or race.

As Feeny notes (personal communication), when the welfare economics model is overridden by ignoring the effects of incomes or other factors on values—something that happens often and happened last year when then EPA Administrator Christine Todd Whitman barred varying the VSL by age—what is left of the normative claims? In such an instance it could be the case that the net benefits would be positive when actual WTP values are used but the net benefits would be zero or negative when income or age effects are ignored.

Undeniably, in the first instance, health indices appear to be free of this conundrum. Indeed, CUA has been embraced in the Canadian health care system, perhaps because it was designed to allocate services with little regard to financial ability to pay.

Yet, to complicate matters, QALYs actually may not be free of income effects. Gafni notes that if preferences over health and consumption cannot be separated, weights based on such preferences are likely to depend on the individual's income and wealth (Blomqvist 2002; Klose 2003; Donaldson, Birch, and Gafni 2002).

Normative Guidance or Relative Rankings

How much normative guidance does the decisionmaker want? CBA offers a possibility of rejecting an alternative course of action on efficiency grounds if it shows this action generates net social losses. CEA can only provide this service in cases where net costs are negative. However, as the most prominent benefits (those to health) are usually in the effectiveness measure, this is likely to be only a small subset of activities that would otherwise be rejected by a CBA.

At the same time, both techniques permit activities to be compared to one another along a relevant dimension. With CBA, activities can be ordered from the activity with the largest net benefit to the smallest. With CEA, activities can be ordered from the activity with the smallest cost-effectiveness ratio to the largest.

Solving the Interpersonal Utility Problem

In adding up the gains (or losses) from a policy, aggregation over individuals is required. In principle, CBA and the associated benefits measure (WTP) are based on ordinal utility and address the Impossibility Theorem of an aggregate social welfare function by creating a potential compensation test: could the gainers from the policy fully compensate the losers with something left over? If so, everyone is potentially better off and the social utility has expanded. This is what economists call the Potential Pareto Improvement test.

CEA is not utility based. CUA (with a utility-based health index as the effectiveness measure) does not attempt to solve the Impossibility Theorem but it does permit aggregation over health effects. The use of the health index is based on the principle of cardinal utility, which assumes that individual utilities can be added by simply summing quality-adjusted life years over people irrespective of their (nonhealth) characteristics. Thus, for any given intervention, some will gain, others will lose, and the choice of intervention will be ordered for cost-effectiveness based on the net effect on QALYs and cost.

Public Goods versus Private Goods

Valuation can occur from one of three perspectives: individual preferences for their own private benefits, individual preferences of the benefits to the community of which the individual is a part, and preferences that are not based on any aggregation of individuals, but are based on some other paradigm (such as prioritarianism, see Appendix II). Both WTP and health index measures are based on individual preferences, either for health improvements to themselves or to the community, however defined (for example, the family, the neighborhood, the nation). Cost-of-illness measures are not based on individual or community preferences, but are closest to community preferences as they are based on the resources cost to the community (Kuchler and Golan 1999).

Which Measure Is More Equitable?

There are many dimensions of equity. Age, health status, and income are particularly salient here. An important question is whether the incidence of a policy by any of these dimensions should influence policymaking. This question came to the fore recently (and noted above) in the contentious debate over the "senior discount" EPA was applying to average VSLs in calculating the benefits of air pollution reductions (Skrzycki 2003). Then Administrator Whitman vowed that EPA would never again adjust the VSL for age groups affected and OMB's new Circular A-4 (2003) also told agencies not to use this practice.

Without offering judgments on the merits of such decisions, it remains that each health valuation measure has different implications for equity, although changing standard practices can alter some of these implications. Health indices implicitly value extensions to younger peoples' lives and healthy peoples' lives more than those who are older or infirm, primarily because of lower life expectancies in elderly populations and because of lower health status in infirm populations. Indeed, Nord et al. (1995), in a survey of a sample population in Australia, found that on average individuals reject the goal of QALY maximization in allocating health resources. A policy of health benefits maximization received very limited support when it involved a loss of equity to the elderly or people with a very limited potential for improving their health. However, such age and health biases can be addressed. Cutler and Richardson (1997) used self-described health status measures applied to groups of different ages and health status to generate

QALY measures that were age and health-specific. Such adjustments work against the simplicity of the QALY measure.

Conversely, WTP measures make no presumption regarding the effect of age and health status on willingness to pay for life extension. That is, a conceptual model, such as the life-cycle model, may offer some predictions about such effects, but ultimately the data determine whether and to what extent such effects exist. Unlike for the health indices, such effects are not predetermined by the definition of the index.

The major concern that some people have about WTP measures is that they are affected by income. In general, it is expected conceptually (and often found empirically) that higher-income people are willing to pay more for a given health improvement than lower-income people. If true, then policies reducing health effects of the rich would be preferred over policies affecting health of the poor, other things being equal. However, the standard practice is to use an average WTP over all income groups and not to make distinctions across them.

One of the most contentious questions in the QALY field is whose preferences should be measured, those of the general public or those who have experienced specific health states. One answer is that it depends on the purpose of the analysis. Kenkel (RFF Conference 2003) pointed out that patient preferences are useful if one is interested in measuring the value attached to people living in a health state. However, if the goal is to value the reduction in risk of anyone in the population incurring that health state, it makes more sense to use the preferences of the public (unless you feel that the public cannot envision the health state at issue).

Findings suggest that individuals experiencing a certain health state may rate that state more favorably than those who do not.

Whose preferences matter is also important because of findings suggesting that individuals experiencing a certain health state may rate that health state more favorably than those who have not experienced it. Gold et al. (1996) cite Sackett and Torrance (1978), Najman and Levine (1981), Epstein et al. (1989), and Slevin et al. (1990) as examples. Other research has investigated the extent to which perspective affects preferences through "framing" effects, yielding mixed results. Richardson and Nord (1997) find that individuals tend to be more concerned with the number of people being treated than with the size of the benefit that each individual receives when placed behind a veil of ignorance, but take the opposite view when assigned the perspective of a social decisionmaker. Dolan and Cookson (2000), however, do not find a discernable difference in preferences elicited under these two perspectives.

As a practical matter, the weights for such methods as Quality of Well-Being (QWB), Health Utilities Index (HUI), and EuroQol or EQ-5D (and also the new Standard Form-6D) all come from types of community samples, as opposed to patient samples. In fact, the Agency for Healthcare Research and Quality (AHRQ) is supporting a nationally representative study at the moment to assign "U.S. population weights" to the EQ-5D. At the same time, much of the published literature on cost per QALY (see Neumann et al.1997) uses convenience samples or expert opinion to assign weights to health domains.

As a matter of best practice, WTP studies and the CBAs using them want to rely on random sampling of individuals in the population of interest and at a minimum be "community based." Convenience and small samples are of course used in developmental work and even in a few studies adopted by policy analysts for their CBAs.

The bottom line for decisionmakers is that both types of health measures embed specific equity biases. Choosing one measure over the other requires considering first these specifics and then how they line up with the appropriate belief system.

To come to some conclusions about which measure best aligns with one's view of equity, Appendix II lays out some paradigms that were addressed in the RFF conference. These paradigms address whether individual preferences should matter more than group or social preferences.

In examining these paradigms, some unanswered questions come to mind. First, is there some reason that, with equal benefits, the distribution of these to a few, but on a high per capita basis, should be valued more or less than the distribution of benefits more broadly, but with a low per capita gain? In other words, to what degree should policy interventions be egalitarian? Second, can certain groups in society be ignored because their benefits from a policy are particularly small? A prioritarian or egalitarian view (see Appendix II) would address this question by looking at the *ex ante* well-being of the groups in question. Third, how is the trade-off between maximizing benefits to the worst off and maximizing total benefits to the community to be resolved? Answering this question requires finding a balance between the utilitarian and prioritarian viewpoints. As Brock (RFF Conference 2003) implied, a balance between paradigms is necessary in practice. Practicing absolute prioritarianism, for example, could lead to a "bottomless pit" problem, benefiting few individuals while draining resources.

Should the high value placed on preventing some types of cancer lead us to policies that favor reductions in this disease at the expense of other programs affecting children, for instance?

Treatment of Risk Perceptions

If the public becomes terrified of a disease, should the chosen valuation method pick up that terror? Should the high value placed on preventing some types of cancer lead us to policies that favor reductions in this disease at the expense of other programs affecting children, for instance? Or do we want a valuation measure that is free of that fear? Another version of this conundrum is: should uninformed preferences matter or only informed ones?

Individuals are affected by qualitative attributes of risk, such as dread, controllability, and the extent to which the risk is voluntary, as much as (if not more than) its quantitative attributes (the size of baseline risk and the risk change induced by the policy under consideration). Both WTP and QALY estimates may be affected by qualitative attributes. Some researchers estimating WTP measures for morbidity and mortality risk reductions have attempted to eliminate the qualitative influences from their estimates; others are attempting to value these attributes directly. To the author's knowledge, QALY researchers have not investigated this phenomenon, although Sackett and Torrence (1978) found that QALY estimates differ depending on whether a disease is labeled.

Avoiding Controversy

Other things being equal, public decisionmakers, like most people, want to avoid controversy. There is a perception that using monetary measures of health effects, particularly where premature mortality is at issue (Sagoff 1993), will be more controversial than using either physical effects or QALYs as a measure of health outcomes. One could even argue that the initial interest in QALYs came out of concern for "placing a value on life." Such concern was perhaps warranted when the human-capital approach to valuing lives was in use. Under this approach, a life lost was valued as the productivity forgone. Modern WTP practices techniques do not include this method, however, as was discussed in Chapter Two of this report.

Clearly, using physical measures, such as lives saved, as an effectiveness measure is less controversial—but may be less useful depending on the policy context, having no normative significance and not permitting aggregation with other types of health effects or nonhealth effects. Using QALYs eliminates the need to express benefits in monetary terms, but this merely postpones the need to put a dollar value on risks to health and life, since eventually a decision must be made about whether it is worthwhile to spend a given amount of money to save a specified number of QALYs.

Time Preferences

The debate about how to account for the incidence of benefits and costs over time—termed discounting—does not really drive a wedge between WTP and QALY measures because both use discounting. Yet, this topic is so important that it is included anyway.

The issue of how costs and benefits occurring over time are to be accounted for in CBA and CEA is as old as these techniques and has not been fully resolved. Welfare economics is clear that discounting the future is appropriate because of the common observation that getting something today is worth more than getting it later. This seems plausible even from a social perspective. Doesn't it make sense for society to invest in reducing 100 deaths today over a program that would reduce 100 deaths in 20 years? Under this paradigm, both costs and benefits should be expressed in terms of present discounted value, which involves applying discount rates to future costs and benefits. In QALY terms, the QALY gains would also be discounted depending on their time of realization.

The traditional consensus among economists is that the same rate should be used to discount both costs and benefits because both of these effects involve resource changes and preferences, so there is no presumption for their differing. The Panel on Cost-Effectiveness in Health and Medicine held at the National Institutes of Health in 2002 and a number of other CEA proponents also support this position. The fact that a given individual may not wish to discount his own future health states at an exponential rate, or at all, is a legitimate position to take. Similarly, the fact that an individual, taking now a community or societal perspective, would not wish to discount—and thus seemingly devalue—the future health states of the representative member of society is equally legitimate. But, it does not follow from that assertion that future health improvements should be treated equally with present health improvements. One must discount QALYs, and in fact at the same rate as cost, if one wishes to secure the result that a given QALY gain achieved at a given real cost is accorded the same "standing" in the CEA calculus, regardless of when in time it occurs.

However, in recent years, an increasing number of health economists have concluded that there is no compelling basis in economic theory to discount health benefits the same as costs (James Hammitt remarks at RFF Conference 2003). Indeed, a number of European countries have mandated that economic assessments either not discount health benefits or use a lower discount rate for health benefits than for costs. Furthermore, in the OMB guidelines (2003), a proposal was offered to discount costs using the real market rate of interest (on the theory that this is the price of abatement investments) and benefits using the rate of time preference, which could be anywhere above zero, but more generally from 1% to 5%.

Some ethical paradigms support discounting; but others do not. For health, the argument is a simple question—from a social view, should health in the future be worth less than health today? Brock (1998) argues that no adequate ethical justification has been offered for discounting health and well-being. He further argues that health interventions often achieve their benefits years into the future, and if these benefits are discounted inappropriately, potentially important policies may be overlooked. Lipscomb et al. (1996), citing Fuchs and Zeckhauser (1987), suggest that if interventions with future payoffs to health are truly undervalued, adjustments should be made to the benefits directly—not to the discount rate. In so doing, the researcher "confronts the allocative implications of such a choice squarely and executes the differential weighting (if there is a compelling case for it) in a precise and transparent fashion" (Lipscomb et al. 1996, p. 222).

Even where discounting is accepted, some argue that the discount rate for health should be different from the discount rate for utility if individuals' rate of time preference for health is different from that for utility (Julie Hewitt remarks at RFF Conference 2003).

■ ■ ■

CHAPTER FOUR

Evaluation of Health Valuation Measures

This chapter examines from a technical perspective the advantages, disadvantages, and underlying assumptions associated with QALYs, WTP, COI, and $/QALY measures. Most of the chapter is organized around Table 4.1, which compares WTP and QALY measures according to a list of attributes identified as desirable during the conference. The criteria for judging these measures include different types of validity, comprehensiveness, ease of application, costs of developing estimates, how well uncertainty is addressed, whether averting behavior is captured, whether qualitative risk attributes are included, and whether these measures bias choices towards certain groups. The COI and $/QALY measures are discussed in a brief separate section of this chapter, rather than being compared in every section to the WTP and QALY measures. The validity of WTP and QALY measures are compared in the WTP section.

Validity

This section is organized around five criteria for judging validity:

- *Criterion Validity:* "The validity of a measure is the degree to which it measures the theoretical construct under investigation," or the true measure (Mitchell and Carson 1989). Ideally this task would involve comparing the measure in question to the "true" measure. Since the truth can't be observed, addressing this criterion has come to mean establishing benchmark measures estimated using a construct that is as close as possible to "truth." One measure of truth is observed choices. For the stated-preference WTP measures and the QALY measures, the question is how well do actual choices reflect rankings or hypothetical choices implied by the measures?
- *Context Validity:* Context validity is defined in a policy context, that is, how close does the construction of the measure mirror the context for public policymaking, for example, are choices involving risk changes, *ex ante*, applicable to individuals and the community as a whole?
- *Convergent Validity:* This is a comparison of the measurement in question with some other measurement by a different approach. This criterion also involves comparison of measurements from the technique or study in question to those from other techniques or studies, the only difference from criterion validity being that here there is no assertion that one measurement technique or study is necessarily better than another. If different measures of the same thing yield similar results, there is a presumption of convergent validity—a kind of safety in numbers. The

TABLE 4-1

Comparisons of Health Valuation Measures for Technical Attributes

Attributes	QALY	WTP
Criterion Validity		
Tested against conditions for preferences to represent utility	Key assumptions violated by individuals, but many perform better in the aggregate.	Performs well.
Comparison to actual choices	SG scores predict treatment choices.	Concern over hypothetical bias for SP studies; difficult to make head-to-head comparisons of SP with actual choices.
Context Validity	SG does fairly well in invoking trade-offs, but not in context of reduced health risks; PTO reflects community-level choices; health domains/states defined on medical interventions may not match health outcomes relevant for policy interventions.	Performs well; however, most health valuation studies are for individual preferences rather than community preferences.
Convergent Validity	Differences in preference weights by approach; SG is the only utility-consistent approach and depends on cardinal utility assumption, but is insensitive to changes in health status.	Differences in RP and SP; SP has potential to better match choice context.
Construct Validity	Focus is more on testing validity of indices than validity of weights; Weights are sensitive to duration of effect, violating independence assumption; Difficult for people to make trade-offs as in SG; Duration estimates often unreliable or ad hoc; Yet, QALY indices can predict the initial degree of medical consumption.	Performs well, except proportionality to scope/scale for CVM.
Content Validity	Critics charge little attention given to "weights" surveys; except in the construction of health state descriptions. Proponents say there is extensive work on this topic.	Major thrust of SP literature.
Comprehensiveness	More comprehensive than WTP, but for health only. Combines mortality and morbidity.	Less comprehensive than QALYs, but covers more than health; doesn't combine mortality and morbidity.
Ease of Application	Easy	Easy
Cost	Cheap to apply, but getting weights is expensive (but a one-time effort).	Cheap to apply, but getting unit values is more expensive per endpoint than QALYs. Presumption is that measures have to be estimated for each health effect–duration combination and by context. But research approaches are changing.
Address Uncertainty in Weights (QALYs)/Prices (WTP)	Relatively little attention here, only in sensitivity analysis. Uncertainty in duration of health states not addressed.	Yes
Recognizes Avoidance Behavior	No	Yes
Inclusion of Qualitative Elements of Risk	Embedded in preferences to unknown degree.	Embedded in preferences to unknown degree; beginning to be an object of research.

problem with this criterion, however, is that just because different approaches or studies come to the same conclusion doesn't mean that they approximate truth. Researchers could be subject to herd instincts, following similar basic paradigms and protocols that may guarantee similar results.

- *Construct Validity:* Does the estimate have properties that are consistent with the theory underlying the construction of the measure? In practice this means examining whether the variables that are expected to influence the measure in question actually do have this influence.
- *Content Validity:* Do the design and execution of the study yielding the estimate conform to the generally accepted best practice or "state of the art?" Defining the state of the art makes this a challenging criterion, particularly when it is unique to the different measures. Thus, the state of the art for QALYs includes issues such as whether the measure includes the items, dimensions, or domains of health status that seem to be relevant and important. For WTP, a relevant element of state of the art would be similar—whether the "commodity" being valued is appropriate to the context at hand. Another more substantive example for WTP is whether a study is designed to test for sensitivity to scope (such as the size of the risk reduction).

Validity of QALYs

QALYs and Criterion Validity. Testing against this criterion requires a judgment about what measures "truth." We take the measurement of utility as the "gold standard," rejecting the idea advanced by some that the standard gamble approach is itself the gold standard, although more is said on this point below. In the case where utility is judged the standard, there seems to be little debate about the assumptions underlying a claim that QALYs measure utility. A QALY represents utility if an individual's preferences satisfy the following conditions:

- *Mutual utility independence.* This complicated condition (see Hammitt 2002) is necessary for utility to be a product of separate health and duration terms.
- *Constant proportional trade-off of longevity for health.* The proportion of an individual's remaining life span he or she is prepared to give up for better health does not depend on remaining life span. Thus, if an individual with 50 years left to live is willing to give up 10 of them (20%) to improve his health state from fair to good, he is also willing to give one year to get the same improvement when he only has five years left to live (20%).
- *Risk neutrality over life span.* This means that an individual favors choices with the longest life expectancy (holding health state constant). Thus, the individual will prefer a situation with a 50/50 chance of living 40 more years or 20 more years (with a life expectancy of 30 years) to a situation of living 29 more years with certainty.
- *Additive independence of utility for health states across time periods.* When health states vary across one's life, preferences for choices on health for any period do not depend on health in any other period. This implies that the sequence in which health states are experienced is irrelevant.

Freeman, Hammitt, and DeCivita (2002) add another condition:

- *Income Independence.* An individual's preference for health and longevity must be independent from his or her income, wealth, and future income.

One could also add that this type of independence assumption would also apply to any other nonhealth factors.

Without delving further into these conditions, they permit one to measure utility by simply multiplying the duration of a health state (including death) by its weight and adding QALYs across different health states in time and across groups of people.

It is generally agreed (Hammitt presentation, RFF Conference 2003) that "these conditions are violated by any individual at any point in time." For instance, there is evidence that health status and duration are not independent. WTP studies, for instance, find marginal utility (as measured by WTP) diminishing for reduced symptom days. Thus, the WTP to avoid a 14-day episode is much less than 14 times the WTP to avoid a one-day episode. Introspection might also lead to questioning the idea that utility for living longer is unrelated to one's health state. As Reed Johnson said at RFF's Valuing Health Outcomes conference: "Is there equivalency between a short and fun life and a dull and long one?" These conditions also imply that people are indifferent to sequences of health states that are improving, constant, or declining. However, studies show people prefer improving health sequences.

To the extent that these assumptions are violated, the use of QALYs to rank prospective government policy actions may well deviate from the "true" measure of utility. A key issue—and one impossible to resolve—is whether QALYs are monotonically related to utility.

Nevertheless, a number of studies show that violations of at least some of these utility assumptions tend to average out over the population (Hammitt 2003). For instance, some people are risk-averse about health states and others are risk-loving. Aggregated, the population may look reasonably risk-neutral about uncertain health states, a requirement for QALYs being a utility measure. In short, while many of the utility assumptions are violated at the individual level, they may not be significantly violated at the population level.

We have also defined criterion validity to also involve comparison of predictions derived for indices to actual choices. Feeny (personal communication) notes that in some cases preference scores predict choices very well. At one level this involves whether an index score can predict choices in a weighting survey out of sample. For instance, the HUI3 predicts SG scores out of sample, with a correlation of 0.88 for 73 health states. More important, however, is whether weighted choices correlate with actual treatment choices. In this regard, there is evidence that SG scores predict choices of treatment. For instance in prenatal diagnosis, several studies compare *ex ante* preference scores to choices and find congruence (see Heckerling, Verp, and Albert 1997). As another example, preference scores for states associated with influenza and vaccination for it were associated with decisions to be or not be immunized (see, for example, Carter et al. 1986). Note, however, that these studies are in the context of medical decisionmaking, not in a public policy context.

Is there equivalency between a short and fun life and a dull and long one?

QALYs and Context Validity. QALYs are built up from weights on health states and duration estimates. Thus, we must subject the weights to the same type of scrutiny about validity. In evaluating the four approaches for estimating health state weights used to build the QALY index, several elements of that context appear salient: (i) the risk of being in a health state is changing;

(ii) there are trade-offs (that is, choices with consequences) in the policy and these choices may involve health and cost trade-offs, as well as health–health trade-offs; (iii) the risk changes apply throughout the community, since *ex ante*, for many policies, it is unclear who will be affected; and (iv) the fact that risks are changing throughout the community also implies that altruism may be an important concern.

Ratings-scale (RS) approaches do poorly against these criteria. The individual places a health state on a line, with no notion of choice and consequence and no notion of risk or uncertainty evoked, for instance. TTO and PTO approaches evoke trade-offs, but not risks. One strength of the PTO approach is that it is community-based. That is, individuals are asked to make trade-offs in effects within a community, not just themselves.

The standard gamble (SG) approach does involve trade-offs and evokes uncertainty to some extent, in that individuals are asked to find a probability of one state occurring versus another state that leaves them indifferent to two health states—call this an indifference probability. For instance, a trade-off might be between living an additional two years of life with a chronic disease and living one additional year of life in perfect health. The respondent would be asked to find the probability that the first would happen versus the other to leave them indifferent to these two outcomes, and that probability is the weight for that chronic disease relative to perfect health used to build the QALY index number.

Indeed, the standard gamble appears problematic if it involves trading off an acute effect (say, pain for a day) for a lottery where sudden death might occur.

However, the health states themselves are not presented as a probability. In fact, there is a baseline probability of living one or two more years, as well as a probability of developing a chronic disease. Feelings about these risks do not (except perhaps in an unobserved way) enter into the choice of indifference probability. Further, the public policy context involves a reduction in risk of a given health state in exchange for higher costs of commodities. The SG approach involves neither a reduction in risk of a health state or costs. Unlike PTO, the SG approach involves decisions about individual health states, not the community. As noted above, only 10% of QALY scoring studies use indices based on the SG approach (Neumann et al. 1997). Prominent among them are the multiattribute systems, such as HUI2, HUI3, and SF-6D.

Another broad weakness in the practice of policy (applicable to all the approaches) is that many studies in the QALY literature take an *ex post* realized approach while policy analysis, in general, takes an *ex ante* perspective. That is, a policy decision is made on the basis of hypothetical risk reductions, while policy evaluations and many QALY studies are performed using data on actual, realized health improvements.

A third concern in using these indices for policy is that the health outcomes being evaluated may be defined too narrowly to match with the epidemiological or other health science literature used to estimate the outcomes. Because many of the indices are developed for use in a medical setting to examine the cost-effectiveness of alternative medical interventions, the health outcomes are often defined for this match. But in the policy setting, health outcomes are very broad, such as in the case of respiratory-related activity days, asthma attacks, or chronic bronchitis. Mapping the specific outcomes evaluated to develop preference weights to such broad health outcome measures may be difficult.

QALYs and Convergent Validity. The convergent validity test involves examining how well a QALY index compares to other indices or measures, without making a claim that these other measures necessarily represent truth any more than the QALY measure in question. In this section the comparison is restricted to QALY indices, leaving to a later section the comparison of WTP and QALY measures.

One set of comparisons involves weights obtained through the standard survey methods for direct utility elicitation. Research into how the choice of technique used to elicit weights affects resulting values tends to show that while ordinal rankings of health states remain unchanged, there is often a wide discrepancy in the cardinal values. Health state scores based on weights from the standard gamble method tend to be much higher than those elicited through TTO and RS, with the relationship among these methods SG>TTO>RS. For instance, Bleichrodt et al. (1997) find that compared to the SG-based score of a particular health state of 0.67, the TTO-based score is 14% lower and the RS-based score is 40% lower.

For instance, a trade-off might be between living an additional two years with a chronic disease and living one additional year in perfect health.

Another set of comparisons examines scores on different indices over a set of health effects. Sung et al. (2003) compare the Health Utilities Indices (2 and 3) and Child Health Questionnaire (CHQ) with each other and VAS-based scores for evaluating quality of life in children undergoing chemotherapy. The CHQ has been proposed as a widely applicable health status measure for children, covering both physical and psychosocial domains (Raat et al. 2002). Moderate correlations were observed between certain domains of the CHQ and HUI2 or HUI3. For example, CHQ body pain was moderately correlated with HUI3 pain (r=0.51), CHQ physical functioning scale was moderately correlated with HUI2 mobility (r=0.58), and CHQ mental health was moderately correlated with HUI2 emotion (r=0.53). Interestingly, only the CHQ psychosocial subscale was correlated with the VAS.

Mauskopf (presentation at RFF Conference 2003) sums up this literature as follows:

- Weights derived from RS, SG, or TTO are less sensitive to clinical changes in a disease condition than health status measures common in the psychometric literature that are not based on these approaches.
- Weights using RS are more responsive to changes in health status than weights using SG or TTO.
- Correlation of weights and psychometric measures of health and functional status is not very high.

Comparisons can also be made indirectly using a multiattribute utility approach. Several scales are available, including different versions of the HUI, the Quality of Well-Being (QWB), and the EuroQol or EQ-5D. Since each of these scales was calculated independently, using different numbers of domains and different methods of preference elicitation, and with different anchors, it is not surprising that they yield different estimates. Also, as might be expected, PTO indices, which ask about preferences for benefits to a community, are poorly correlated with the other indices, which relate to preferences regarding benefits of an individual to himself.

QALYs and Construct Validity. The conference yielded insight into a major source of confusion in the analysis of construct and content validity for QALYs. QALY practitioners focus almost exclusively on the problems and successes in applying their indices to estimate QALY gains in response to a medical intervention, such as the QALYs gained from a treatment of chemotherapy. Testing construct and content validity from this perspective would involve measuring how well the various indices score these health state changes, how the rank or cardinal order of different scores from different indices compare to one another, and how much care was taken in applying the index to score new situations.

Yet, in the context of a comparison of QALY to WTP measures, this comparison is misplaced. The relevant analogy if such a focus were applied to WTP measures would be to judge whether the right monetary unit values were being applied to the right health effects, and whether they were being summed appropriately to avoid double counting. While important, this procedure should not be the main focus.

Rather, the main focus should be to test the construct and content validity of the weights underlying the index. These weights are analogous to the "prices" or the WTP measures applicable to monetary measures of given changes in a health state. For testing construct validity this means examining whether the weights make sense according to the special assumptions made in utility theory, how much they vary over different studies using the same basic method (such as SG), the ease and consistency with which people answer the trade-off questions in these surveys, and so on. For content validity the focus is on whether surveys designed to yield the underlying weights behind the various QALY indices are state of the art and how advanced the state of the art is.

Turning to the specifics of construct validity, a basic observation is that nearly all the QALY indices, irrespective of how morbidity weights are derived, assign a zero weight to a life year lost. Thus, in the evaluation of performance of different indices and different approaches to deriving weights, the index scores are likely to be very similar across policies involving nontrivial death risk reductions.

Because construct validity applies both to the weights and the overall index values, we consider both below.

Construct Validity of the Weights. Because the SG approach appears to be the most appropriate for use in a policy context, we consider it first. One issue with the SG approach is that the scenarios used may be unrealistic (it is difficult to imagine a treatment for infertility that would carry with it significant probability of death, for example). Indeed, the standard gamble appears problematic if it involves trading off an acute effect (say pain for a day) for a lottery where sudden death might occur. There is evidence that people are unwilling to make this sort of trade-off. This point gives rise to a standard critique of the SG approach, which is that it is not credible for use in valuing acute health effects.[18]

Another issue is that the weights are sensitive to the particular duration of the effect appearing in the question. Specifically, 75% of the sample in a study by Bala et al. (1999) gave different weights for an SG involving one year of severe pain versus 20 years of severe pain. This should not come as any surprise, of course, but the results showed that half of this group had a lower weight for the 20-year pain and half had a higher weight! Overall, however, because of this heterogeneity of responses, the average weight showed no difference between these two cases!

Still this kind of result is unsettling because QALY indices are constructed assuming separability between severity and duration.

Similar concerns arise with the use of the time trade-off approach. First, Smith (presentation at RFF Conference 2003) reported survey data suggesting that people, in general, are unwilling to trade time for health. Second, Gold et al. (1996) note that the TTO approach confounds time preferences with preferences for health states. Because the years that are being "traded off" in the TTO approach come at the end of one's life span, they may be of less value than the preceding years. Failure to adjust for time preferences may bias scores upward.

Finally, in the elicitation of weights, choice context is likely to be as important as it is in the WTP literature. For example, Kerry Smith presented results (RFF Conference 2003) showing that responses to TTO questions are influenced by age, view of future health, and recent hospitalization status. These factors are not assumed to affect weights in the application of QALYs.

There are several additional issues. First is the treatment of states viewed as worse than death. Some critics contend that QALY studies generally do not allow for this case. Feeny (presentation at RFF Conference 2003) contends that methods for assessing states as worse than dead with the SG and TTO approaches are available and have frequently been used and that HUI3 and EQ-5D provide much scope for negative scores. Second are ceiling effects, that is, indices may not be sensitive to health improvements in individuals who are already near perfect health. Third, there is evidence that weights are not responsive to "small" changes in the description of the health state.

Construct Validity of the Index Values. Because index values for health states include the weights on health effects as well as their durations, we need to address the issue of the reliability of duration estimates. One expert (Mauskopf) says that, for QALYs, uncertainty or disagreement about the weights is probably less important than the uncertainty about the estimated duration of the times spent in different health states. In particular, if the health states associated with a certain disease include death, then duration information will dominate the calculation of QALYs lost because of the condition, that is, whether you assume a preference weight of 0.6 or 0.5 for those who survive will not make much difference to total QALYs. Even for health states where mortality is not an issue, the variability in the estimates of the weights is not as important for estimating the QALYs gained by avoiding the health state as the estimate of the time that one would spend in that health state. For example, if a prospective policy would reduce air pollution to reduce asthma attacks—the QALYs gained by the intervention would be much more sensitive to estimates of the number of asthma attacks avoided than to estimates of the preference weights with and without an asthma attack. Indeed, we hypothesize that QALYs estimates are more sensitive to duration than WTP estimates because WTP shows diminishing returns with duration in a health state while QALYs do not, by construction.

Finally, some studies address construct validity by comparing the performance of health status indices across a range of health effects or populations. Raat et al. (2002) evaluate the construct validity of the Child Health Questionnaire (CHQ) relative to the Health Utilities Index. They find that six of the 11 CHQ scales clearly discriminate between children with and without chronic conditions, that three of the scales discriminated between high and low medical consumption, and that the discriminative ability of the CHQ and the HUI2 were comparable.

QALYs and Content Validity. Critics contend that the effort that goes into improving and refining methods for estimating weights is predominately on defining the health states or domains, as opposed to testing and improving the surveys to derive weights. Conference participants from the QALY "side" took issue with this statement, noting extensive work on developing multiattribute systems.

Multiattribute systems aside, evidence for this lack of attention may be found, for instance, in the absence of journal articles or books devoted to this topic. Indeed, the most recent study that our experts provided to describe survey protocols for eliciting weights based on standard gambles and other approaches was an out-of-print report dated 1990 (Furlong et al. 1990)! One specific issue taken up, however, is how weights can be affected by the elicitation procedure. Lenert et al. (1998) note that whether you use a "ping-pong" or iterative procedure to arrive at the indifference point on the standard gamble can affect QALY weights by 10% to 15%.

The task of discerning best practice in QALY index construction is not an easy one because there are so many "practices," or different indices, and such practices vary not only in their elicitation methodology, but also in the domains of health that they assess and the level of detail with which they do so. For example, QWB scores are derived using a visual analog scale (VAS) and are comprised of weights on three multilevel domains related to function (mobility, physical activity, and social activity) and a symptom–problem complex with 27 possible outcomes. This index has the potential for 1,170 different health states. The HUI2, the second version of the Health Utilities Index, is based on a combination of SG and VAS responses. This index weights seven domains: sensation, mobility, emotion, cognition, self-care, pain, and fertility, and allows for 24,000 potential health states. The EQ-5D was originally derived using a VAS, but has since been reweighted using TTO. It weights five domains and allows for 243 unique health states. Given these differences in design, it is difficult to discern a "best practice" approach.

Validity of WTP Measures[19] and Comparison with QALYs

In judging the validity of WTP measures, there are two classes of estimation approaches that may score differently with respect to such an assessment—revealed preference (RP) and stated preference (SP) approaches. Examining the relative validity of the RP and SP approaches with respect to the WTP measures is analogous to examining the validity of the SG, TTO, PTO, and RS approaches with respect to the QALY measures.

WTP and Criterion Validity. The benchmark of "truth" for judging the criterion validity of WTP estimates is unclear. Concern about hypothetical bias in SP studies—under the rubric "watch what I do, not what I say"—leads to the identification of market prices or even WTP values from revealed preference studies as benchmarks. However, no RP studies or market prices exist for valuing health in a public goods context. Moreover, the RP studies in the literature on health valuation all have significant limitations of their own, making the SP comparison to RP studies problematic and potentially misleading. See the discussion below on convergent validity.

Where comparisons between RP and SP studies are possible and desirable, researchers have addressed criterion validity by creating simulated or experimental markets, where individuals make transactions. Nearly all of these analyses are for commodities other than health effects. Researchers compare responses from SP studies of hypothetical markets with the responses from these experimental markets, which are designed to approximate the hypothetical and serve as the point of reference. One of the earliest examples of this approach was a study by Bishop and Heberlein (1979), who made hypothetical offers to purchase goose-hunting permits from a sample of hunters, while making real cash offers for a limited quantity of permits from a different sample. Mitchell and Carson (1989) provide a number of examples of this approach to assessing criterion validity. The appropriateness of such studies has been disputed. Sampling variability and the role of model specification in obtaining value measures from both the real and hypothetical transactions may imply that these studies do not actually assess criterion validity (Freeman 2003).

List and Gallet (2001) conducted a meta-analysis of 29 such studies. They found a median calibration ratio of about 3.0 for hypothetical to real transactions, though this ratio was lowered considerably when a WTP rather than a WTA question was used, or when a first-price, sealed-bid auction mechanism was used (Freeman 2003). Freeman notes that it has been suggested in the literature that CV estimates be reduced by a calibration factor as a result of such evidence.

For revealed preference studies, criterion validity is the degree to which the WTP estimates from hedonic wage or price studies truly reflect an individual's utility for a risk reduction. Regarding labor market wage-risk studies, Viscusi and Aldy (2003, p. 8) suggest that "an ideal measure of on-the-job fatality and injury risk would reflect both the worker's perception of such risk and the firm's perception of the risk. Because the market opportunity locus reflects both workers' preferences over income and risk and firms' preferences over costs and safety, information on both sets of beliefs would be necessary to appropriately characterize the risk premium." They go on to note, however, that few studies have compiled workers' subjective risk preferences, and that there has been no research conducted on firms' risk perceptions. Instead, the approach used in wage hedonics attempts to characterize these preferences indirectly, relating an employee wage rate to an estimate of mortality risk based on the mortality rate for employees in a given occupation. To the extent that this approach mischaracterizes risk perceptions and preferences, it does not achieve criterion validity.

In one sense, QALYs do much better than WTP estimates according to the convergent validity criterion, but in practice does this matter?

Another possible, if crude, test of criterion validity is to compare WTP to income. If a people say they are willing to pay an unreasonably large fraction of their income to obtain some improved health state, one might question whether the value is reasonable. Of course, determining what is reasonable is difficult. But, typically, WTP for small changes in chronic illness or mortality risks amounts to only 1% or less of individual income and WTP to avoid a day of symptoms may be only $100 or less.

WTP and QALYs Compared. Tested against conditions for preferences to represent utility, the literature seems fairly unambiguous that QALYs fail for any given individual. There is evidence that once aggregated for many individuals, some of these violations of assumptions appear to cancel out. A major question among QALY practitioners is the more philosophical one of whether a measure that doesn't represent utility exactly can still be useful in the policy process. WTP measures, on the other hand, are derived from a body of theory that is utility based. At the same time, utility is not the only possible standard for examining criterion validity. There is evidence that some scoring studies yield predictions of behavior that are borne out in reality. Revealed-preference WTP studies are based on behavior, so, by this benchmark, they of course do well tautologically. The relevant literature on SP studies and behavior and in comparison to RP results is mixed and difficult to sort out. Unexamined here is how well contingent behavior studies match up with actual behavior and choice experiments involving health choices match up with real choices.

WTP and Context Validity. As with the QALY analysis, we need to examine how well WTP studies fit the policy context: (i) the risk of being in a health state is changing; (ii) there are trade-offs (that is, choices with consequences) in the policy and these choices may involve health and cost trade-offs, as well as health–health trade-offs; (iii) the risk changes apply throughout the community, since *ex ante*, for many policies, it is unclear who will be affected; and (iv) the fact that risks are changing throughout the community also implies that altruism may be an important concern.

The fit of WTP studies to (i) and (ii) is very tight; (iii) fits as well, based on the *ex ante* nature of such studies. Only a small subset of health valuation studies have been carried out at the community level. Most studies are of individual willingness to pay for individual risk reductions rather than for community risk reductions. Exceptions in the mortality risk valuation literature include Strand (2002) and Johannesson et al. (1996), who find that the value of a statistical life (VSL) is larger for public goods than equivalent private goods. Strand (2002) finds that an increase in private safety results in VSLs between $1.2 and $2.5 million, while the public safety improvement results in VSLs of at least $2.5 million with respondents saying a large fraction of their bid was altruistic. Johannesson et al. (1996) actually come to the opposite conclusion, but this illustrates the difficulty of eliciting values for public goods. They used a tax increase as the means of paying for the public safety good, which may have introduced a downward bias to the estimate, given how negatively people feel about increasing their taxes.

WTP and QALYs Compared. The WTP metric in general fits the policy context better than the QALY metric, but both pay too little attention to estimating preferences for community health improvements.

WTP and Convergent Validity. Convergent validity is assessed by attempting to measure the same value with different methods. If a high correlation exists between the different measures, then convergent validity has been satisfied. In the WTP literature, this means comparing results from revealed- and stated-preference studies. Unfortunately, such comparisons are problematic because the valuation contexts differ. The bulk of the RP literature concerns mortality risks on the job, which involves accidental deaths (often in a vehicle accident) to healthy, prime-aged people. The context for gauging the effects of government policy on health (except for transport and occupational health policy) is often nonaccidental deaths to ill, older people (and the very young). Of the relatively few SP studies, some apply to the context of accidental death in transport; others are closer to the contexts applicable to regulatory activities at EPA, USDA, and Health and Human Services (HHS). Thus, it would not be surprising to find differences in the WTP estimates from these different techniques.

For mortality risk valuation studies, as used by EPA, a reasonable number of descriptive comparisons and meta-analyses have been carried out. These comparisons show that value of a statistical life (VSL) estimates from SP studies generally are below those from RP studies, although variances are wide. Theory suggests that the values derived from the two approaches could differ. Wage-risk RP studies estimate a local trade-off at the intersection of the market offer curve and an individual's expected constant utility locus, while SP studies approximate a movement along a constant utility locus. In particular, there should be discrepancies between the two approaches for large changes in risk (Viscusi and Evans 1990; Lanoie et al. 1995). Furthermore, Lanoie et al. (1995), in their analysis of VSLs derived for a single population using both SP and RP approaches, suggest that values from the two approaches are likely to be most similar when the study population includes few risk-averse individuals. Their explanation is that risk-averse workers are likely concentrated in jobs where there is no explicit risk premia or one that is difficult to detect statistically. At the same time, these individuals are likely to have a higher WTP for a reduction in mortality risk.

There are no RP-based estimates of acute and chronic health effects, except for on-the-job injuries, so no convergent validity tests comparing SP and RP studies are possible for these types of endpoints. Conversely, there are no SP estimates for on-the-job injuries. So, again, an RP–SP comparison is precluded.

For the reasons discussed above, assessment of convergent validity of RP studies is also problematic. Some insight can be gained by comparing estimates across the different types of RP studies on mortality risk, which, while dominated by labor market studies, also include product and housing-market studies. The value of such an exercise is limited, however, because the risk contexts from which values are derived differ. Market studies have examined price-risk trade-offs for seatbelt use, cigarette smoking, home fire detectors, automobile safety, bicycle helmets, and housing price responses to hazardous risk sites. In general, these studies have found an implicit VSL on the same order of magnitude as labor-market studies, though on average slightly lower (Viscusi and Aldy 2003). The fact that estimates from all three types of studies seem to fall in the same order of magnitude provides some confidence in these techniques.

Mitchell and Carson (1989) provide a review of findings of early studies of convergent validity relating primarily to environmental resources. In general, there has been a correlation between the estimates from SP studies and RP studies. Freeman (2003) cites Carson et al. (1996), who conducted a study of convergent validity on WTP estimates for quasipublic goods. The study found a SP/RP ratio of 0.89 for the entire sample (95% confidence interval: 0.81–0.96).

Another test of convergent validity is to compare WTP estimates with cost-of-illness estimates. As noted, cost-of-illness measures can be expected to be less than WTP measures for the same change in health state. And empirical studies bear out this relationship, with typical ratios in the 2-to-1 or 3-to-1 range.

WTP and QALYs Compared. In one sense, QALYs do much better than WTP estimates according to the convergent validity criterion, primarily because the weights are confined to the [0,1] interval for QALYs and are not so constrained for WTP. In another sense, with so many different QALY indices defined on so many different sets of health domains, any given condition could receive a wide range of scores, perhaps wider than monetary values using the WTP metric. A key question in interpreting performance with respect to convergent validity is if in practice these differences, whether in QALYs or in WTP estimates, matter much for altering policy prescriptions.

WTP and Construct Validity. Construct validity assesses the extent to which WTP estimates are consistent with the expectations of theory. The NOAA Panel, which was convened to examine the stated-preference literature in light of the Exxon *Valdez* compensation study, ratified many of the practices already in use by the best SP studies for testing construct (and content) validity and developed or sanctioned several more.

One of these is the common method of regressing the WTP estimate on independent variables believed to be determinants of the WTP value based on the theory underlying these estimates (welfare economics). The reporting of results from such regressions has become common in SP studies (Mitchell and Carson 1989).

For health outcomes, WTP from SP studies should be related to income and to the size of the risk change, duration of effects, or severity of effects being valued, while the effects of age and health status (two important variables in a comparison with QALYs) have ambiguous effects in theory.

While it is difficult to simply characterize the large SP literature, it is generally true that income affects WTP and that longer durations, greater severity, and greater risks of effects result in higher WTP for health improvements, as predicted by theory. For instance, in totally separate studies, WTP for reductions in acute illness has been found to be far less than the WTP to reduce a statistical case of chronic illness, which has been found to be far less than the WTP to reduce a statistical death. Table 4.2 gives a sense of these relationships as well as comforting information on convergent validity when looking across valuation efforts in different countries or by different groups. As for age and health status, in one set of new SP studies on WTP for reduced mortality risks (Krupnick et al. 2002), age had an effect only on the WTP of people over 70, while ill people, if anything, were willing to pay more than healthy people to reduce mortality risks.

Results are not as comforting on sensitivity to the size of the risk change or, as economists—and the NOAA Panel—call it, "sensitivity to scope." The theory predicts that WTP should be proportional to the risk change being valued, other things being equal, but most SP studies find that WTP for greater risk changes increases less than proportionally to the risk change.

Assessment of construct validity for RP studies is similar to that for SP studies. However, because most of the studies focus on mortality and are typically conducted in the labor market, some of the key variables differ. In hedonic-wage studies, for example, the wage rate should be a function of risk, income, age, union status, education, and job experience. Theory would suggest that the wage–risk trade-off and implicit VSL are related positively to both risk and income, negatively to age (different from SP, where the relationship is ambiguous), and positively to union status, because of implied bargaining power and improved workplace safety.

As with SP studies, RP estimates of WTP are affected by the size of the risk, though not proportionately. Mrozek and Taylor (2002) find a nonlinear relationship between baseline occupational mortality risk and VSL. They find that the estimated VSL is 75%–110% higher at mean risks of 1.5×10^{-4} relative to the VSL based on a mean risk of 0.24×10^{-4} . Above a risk of

TABLE 4.2

Comparison of Unit Values Used in Several Major Studies or Models (IN 1990 U.S. DOLLARS)

	United States EPA[a]			United States TAF[b]			Canada AQVM[c]			Europe ExternE[d]
VALUES	LOW	CENTRAL	HIGH	LOW	CENTRAL	HIGH	LOW	CENTRAL	HIGH	CENTRAL
Mortality *in millions*	1.560	4.800	8.040	1.584	3.100	6.148	1.680	2.870	5.740	3.031
Chronic Bronchitis	—	260,000	—	59,400	260,000	523,100	122,500	186,200	325,500	102,700
Cardiac Hosp. Admissions	—	9,500	—	—	9,300	—	2,940	5,880	8,820	7,696
Resp. Hosp. Admissions	—	6,900	—	—	6,647	—	2,310	4,620	6,860	7,696
ER Visits	144	194	269	—	188	—	203	399	602	218
Work Loss Days	—	83	—	—	—	—	—	—	—	—
Acute Bronchitis	13	45	77	—	—	—	—	—	—	—
Restricted Activity Days	16	38	61	—	54	—	26	51	77	73
Resp. Symptoms	5	15	33	—	12	—	5	11	15	7
Shortness of Breath	0	5.3	10.60	—	—	—	—	—	—	7
Asthma	12	32	54	—	33	—	12	32	53	36
Child Bronchitis	—	—	—	—	45	—	105	217	322	—

a. U.S. EPA (1999). The Costs and Benefits of the Clean Air Act Amendments of 1990. Low and high estimates are estimated to be one standard deviation below and above the mean of the Weibull distribution for mortality. For other health outcomes they are the minimums and maximums of a judgmental uniform distribution.

b. Tracking and Analysis Framework (Bloyd et al. 1996), developed by a consortium of U.S. institutions, including RFF. Low and high estimates are the 5% and 95% tails of the distribution.

c. Air Quality Valuation Model Documentation, Stratus Consulting (1999) for Health Canada. Low, central, and high estimates are given respective probabilities of 33%, 34%, and 33%.

d. ExternE report (1999). Uncertainty bounds are set by dividing (low) and multiplying (high) the mean by the geometric standard deviation.

1.5×10^{-4}, however, VSL estimates begin to decline, falling about 10% between risks of 2.0×10^{-4} and 1.5×10^{-4}. Regarding income, Hamermesh (1999) finds that workplace safety is income elastic. Viscusi and Evans (1990) estimate an income elasticity of occupational-injury risk valuations between 0.6 and 1.0. Furthermore, Viscusi (1978) finds that the VSL is increasing in worker wealth (Viscusi and Aldy 2003). Such findings are consistent with economic theory and provide some evidence of construct validity. Regarding age, findings on the relationship between age and VSL in RP studies have been more consistent than in SP studies, although there are few workers above 65. A number of studies have found that the magnitude of the VSL is a decreasing function of age. Presumably, this could be explained by the fact that hedonic functions do not account for wealth and the possibility of a greater ability to pay for a reduction in mortality risk. Finally, regarding union status, studies have found that union membership does increase the wage–risk premium and the implicit VSL, as would be expected if unions exert pressure on wages and seek to compensate their membership for added occupational risk.

A significant threat to construct validity for RP studies is omitted variable bias. In labor markets, variables such as “coolheadedness” or individual productivity levels may influence wages and are difficult to capture empirically. Furthermore, mortality risk could be influenced by factors such as injury risk, the physical exertion required for the job, or environmental factors. An omission of injury risk, for example, has been found to cause a positive bias in mortality risk measures for certain samples (Viscusi 1981; Cousineau et al. 1992; Viscusi and Aldy 2003).

WTP and QALYs Compared. The focus among QALY practitioners is more on testing the validity of indices than that of weights. This is in contrast to the WTP studies, where construct validity is a major area of analysis, although insensitivity to scope dogs these studies.

WTP and Content Validity. Content validity assesses the degree to which the survey instrument presents information accurately and appropriately so as to elicit unbiased and accurate statements from respondents. This involves an assessment of such elements as survey design, question wording, and structure and extent of participant buy-in (such as acceptance of the scenarios and a feeling that their response is consequential to the outcome of the debate in question) to identify potential misspecification of the scenario, value cues, or anything else that may lead to a biased result. Methods such as the pretesting of surveys in focus groups with debriefings and the inclusion of follow-up questions within the survey to capture respondent confidence and understanding are two common practices in SP studies for addressing these issues. As noted above, the NOAA Panel on contingent valuation issued requirements for such studies in order to make the assessment of content validity somewhat standardized and transparent. These call for detailed reporting of the size and nature of the population sampled, the sampling frame used, and response rates, as well as making available survey questionnaires and data. State-of-the-art CV studies routinely follow these protocols, although some recent literature has begun questioning them.

Bishop et al. (1997) propose a more comprehensive set of questions and a point system for rating content validity. These questions address the adequacy of the theoretical definition of the true value, the identification of environmental and economic attributes and the implications of an intervention for these, participant acceptance of the scenario and understanding of their budget constraints, the appropriateness of the survey mode, the sufficiency of survey pilots and pretests, and the appropriateness of the study population sample, among other factors.

It should be apparent that an enormous amount of effort goes into the proper design of an SP survey. This effort has revealed a variety of results that lead one to accept and trust the findings from such studies and other results that are more troubling. For instance, much work on the appropriate way of asking WTP questions on surveys shows that dichotomous choice questions are more reliable than open-ended questions. At the same time, there is a literature that describes a yea-saying behavior in those taking such surveys and offers approaches to ameliorate it.

Another very positive aspect of SP studies is the now-standard practice of asking extensive "debriefing" questions at the end of such surveys. These questions are used to test for consequentiality, acceptance of the baseline and scenario, and the identification of people who tend to vote "no" because they don't like the scenario used to elicit WTP (such as a tax increase scenario turning off a "tax hater," or people who vote "yes" because they want to protect the environment or people's health, irrespective of the costs).

For RP studies, assessment of content validity is more straightforward. The hedonic regressions that serve as the basis for these studies are derived from economic theory and thus less open to question than SP content and design. Content validity for RP studies is addressed primarily through improvements and refinements in estimation techniques and improvements in data quality. For instance, there is debate in the literature about the appropriate functional form for hedonic-wage studies. Both linear and semi-log specifications are common in the literature, and an appropriate functional form cannot be derived from theory (Viscusi and Aldy 2003). While a few studies have tested the explanatory power of the two functional forms (examples include Moore and Viscusi 1988 and Shanmugam 1997), they have not found strong evidence in favor of either.

Regarding data quality, the reliability of mortality-risk data has been questioned for some time. Viscusi and Aldy (2003) note that the Society of Actuaries data, commonly used in U.S. labor-market studies from the 1970s and 1980s, reflected the overall fatality rates of individuals in a job category, and thus were not appropriate for characterizing "on-the-job" risk. In fact, some surprising conclusions could be drawn from the data, such as that being an actor is one of the highest-risk occupations! Different concerns arise with the use of mortality-risk data from other sources, such as those collected by Bureau of Labor Statistics (BLS), which are used in the majority of more recent studies. In particular, there is concern over the aggregate nature in which the data were reported through the early 1990s, which fails to properly account for interindustry wage differentials, thus biasing VSLs upwards. More recently, there has been concern over substantial differences in mortality counts among data sets, such as that between BLS and National Institute of Occupational Safety and Health (NIOSH) data. Mrozek and Taylor (2002) note that VSL estimates from NIOSH mortality data are approximately 75% larger than those from BLS data.

Debriefing questions are designed to identify people who intend to vote "no" because they don't like the scenario.

Overall, Mrozek and Taylor (2002) suggest that differences in model specification and data sources have led to a wide range of VSL estimates, as well as a failure to find a statistically significant relationship between mortality risk and wage in numerous studies. Viscusi and Aldy (2003) note that the middle 50% of U.S. labor-market studies estimate a VSL between $5 million and $12 million ($2000). They report a median VSL of about $7 million, which they cite as being in line with the most reliable studies. Results, however, have ranged from as low as $100,000 to as high as $25 million. Mrozek and Taylor (2002) report a "best-practice" VSL of approximately $2.4 million ($2000) for populations facing the average risk of accidental death in the workplace. This is far below most estimates in the hedonic-wage literature (although closer to estimates in the SP literature), a finding that they attribute to the fact that much of the hedonic wage literature has failed to account for unobserved differences in wages at the industry level.

WTP and QALYs Compared. While there are bound to be differences in opinion about match-ups between these two types of metrics, the match-up for content validity is perhaps the most controversial. Most agree that recent studies under the WTP metric pay inordinate attention to content validity, designing studies to test for it in any number of ways, guided by the NOAA Panel protocols and other research. The disagreement is in the degree to which this same attention takes place by practitioners who estimate weights and weighting functions used to develop QALY scores.

Validity of $/QALY

The use of $/QALY factors has become common in the health policy literature as a means of implementing some form of CBA in health policy[20] and as a benchmark against which medical interventions are judged cost-effective. A number of estimates of benchmarks or "cutoff points" have emerged, ranging from $10,000 to about $200,000. A policy costing more than the cutoff per QALY gained would be deemed cost-ineffective, because it exceeds society's willingness to pay.

It was suggested at the conference that these approaches as they have been used to date are not theoretically sound. Hammitt (2002) notes that variables such as life expectancy, health state, baseline risk, and wealth affect QALY and WTP measures in qualitatively different ways. As a result, individuals cannot be expected to have a constant rate of substitution between QALYs and wealth. QALYs, for example, decline with age (holding health state constant), while VSL is constant or increasing over a large age range (holding health state constant). Over this range, then, an individual's WTP per QALY would be increasing with age. The finding of a nonconstant WTP per QALY implies that QALY-based CEA is inconsistent with welfare economic theory and thus CBA.

The current reality is that cost-effectiveness thresholds based on estimates of WTP per QALY are advocated in many quarters and, as Bernie O'Brien suggested (presentation at RFF Conference 2003), the consequence of not attempting to improve such estimates is arbitrary decision rules.

Validity of COI

The COI approach makes no pretense to measure utility. Rather it is designed to capture the costs of illness to the measured economy and includes medical expenses, forgone earnings, and productivity losses to employers. COI is often portrayed as a lower bound of the societal costs of illness (Harrington and Portney 1987) because it ignores the value that affected individuals place on feeling well, avoiding pain, taking part in recreational activities, and other welfare effects. For this reason, COI per case estimates—costs such as hospitalization, lost workdays, and doctor visits—are routinely used alongside WTP estimates to fill in gaps, implying that on a WTP basis, the resulting benefits are underestimates. However, there is no necessary relationship between the two measures, which calls into question this practice, particularly where morbidity effects dominate.

The biggest validity issue concerning COI estimates concerns content validity. Deriving such estimates has traditionally not been standardized as to data sources, types of effects to be studied, and techniques for predicting the fraction of the population to experience any given health state. This situation has recently changed. USDA's Economic Research Service (ERS) has put significant effort into developing transparent and comprehensive methodologies to value cases of illness and the particular costs involved. The costs of being ill by four pathogens found in food are embodied in their "cost-of-illness calculator" (Economic Research Service 2003). Symptom-duration "outcome trees" are used to describe the severity and duration of various effects and the percentage of those with the infection who develop them.

Comprehensiveness

Agency practitioners of cost–benefit and cost-effectiveness analyses would likely want to use a type of health valuation measure that could offer them values for the widest range of health effects. In so doing, CBAs from the agency would become more routine and consistent, and education of decisionmakers could be limited to one chosen measure instead of several. As an example, EPA's focus and refinement of WTP measures in its showcase Section 812 prospective and retrospective cost–benefit studies of the Clean Air Act (U.S. EPA 1997, 1999) have served as the basis for all of EPA's more recent regulatory-impact assessments (RIAs) regarding air pollution and have been codified in integrated assessment models used by the agency.

QALYs have been developed for a wider range of symptoms and health states than have WTP estimates. Furthermore, with weights available for functional limitations, it is possible for experts or groups of individuals to develop scores for new diseases and symptoms as well as changes in their frequency or severity. This exercise can be seen as analogous to benefit-transfer techniques used with willingness-to-pay estimates (see below).

Yet, this coverage comes at a price. There is a bewildering array of QALY indices available, each with its own unique protocols, strengths, and weaknesses. Also, information on symptom duration as well as changes in functioning is subject to large uncertainty and is not systematically available. Finally, health domains are defined in ways that are sometimes too specific to match the epidemiological endpoints being targeted by policy. An example is that one might be able to score various stages of bladder cancer but not the disease itself.

Ill people, if anything, were willing to pay more than healthy people to reduce mortality risks.

WTP estimates, despite their lack of comprehensiveness, do not have the same type of problem with duration. Descriptions of the health effects to be valued in SP studies usually either include duration information or permit the respondents to define duration for themselves. RP studies yield values based on average duration (and severity) of the health state in the population contributing the data to the study. If an SP study is designed carefully, a functional relationship can be derived to link marginal changes in duration to a person's willingness to pay to improve his or her health status. This approach provides much more flexibility than having to rely on WTP estimates based on duration information built into the "commodity" being valued.

Beyond conducting primary research studies, WTP practitioners attempt to address the lack of comprehensiveness through the application of benefit-transfer approaches. Benefit transfer applies a WTP estimate from an existing study to the population sample in a study at hand. It can also apply a functional relationship explaining WTP from one study to a new disease or health effect by plugging in values for the explanatory variables relevant to the new effect. The practitioner assesses the quality and appropriateness of existing estimates in choosing an estimate to apply. Furthermore, where more than one potentially appropriate study exists, the practitioner can combine estimates from existing studies, potentially mitigating the biases or shortcomings of any of the individual studies (Desvousges, Johnson, and Banzhaf 1998). Numerous studies have used benefit-transfer methods to estimate benefits from policy. EPA's prospective and retrospective studies of the Clean Air Act (U.S. EPA 1997 and 1999) both use a statistical analysis of 26 studies that estimate the value of a statistical life. The weakness of the benefit-transfer approach is that estimates it generates are likely to be less accurate than those from primary research tailored for the study area and policy question.

Another important development is that SP practitioners are beginning to turn to conjoint analysis (now called choice experiments), which asks individuals to choose among programs with different attributes, such as functional limitations and health states, as in QALY analysis. Attributes can also be embedded in a description of alternative health states over the life cycle (DeShazo and Cameron 2003). The difference between these choice experiments and "weight" surveys (say using SG methods) is that the cost of the program is treated as an attribute in the WTP study, allowing the identification of the WTP for each attribute. Widespread application of this technique holds the promise for rapidly expanding the comprehensiveness and flexibility of the WTP approach.

Although WTP studies are not comprehensive in valuing health outcomes, they do cover a wide range of nonhealth effects that are experienced jointly with health effects and associated with many different types of policies. EPA- or USDA-type regulatory initiatives, for instance, would likely improve recreational and commercial fisheries, visibility, material lifetimes, and the like. Such improvements should be captured by a cost–benefit analysis to create a comprehensive picture of the effects of a policy. QALYs do not address these types of benefits. However, if one wanted to use CEA instead, a net-cost-effectiveness analysis could be applied in this case, where monetized nonhealth benefits could be subtracted from the cost of the policy, which would then be divided by the change in QALYs.

Ease of Application

Practitioners often operate under tight deadlines and want approaches that are easy to use. Neither health outcome measure meets this concern for the valuation of morbidity. WTP unit values for mortality-risk reduction are widely available and can be easily used. What's more, as noted above, integrated assessment models exist that embody the extant WTP estimates and are often updated, making the process easier still. On the other hand, few WTP estimates for reduction in morbidity are available, and those that are available are mostly limited to air pollution and respiratory disease.

Scores for health states, as well as the underlying weights and scoring equations are not widely available for some indices. For others, the situation is better. For instance, HUI3 has been used in many Canadian population health surveys to yield scores by age, gender, education, health condition, and co-morbidities. Published tabulations are available for many diseases, but without detailed knowledge of the original source population and disease severity, it may be difficult and misleading to apply such published scores.

Costs of Developing Estimates

A WTP study generally addresses one type of health effect (although sometimes with different degrees of severity, duration, latency, and so on) while a "weights" study provides weights for a large set of health states and domains. The latter type of study is likely to be far more expensive than a given WTP study, but to cover more territory, albeit in less depth. Indeed, when a WTP analysis uses secondary data, as is often the case with RP studies, the costs of WTP studies are likely to be exceedingly low compared to a weights study. The increasing use of choice experiments to develop attribute-based WTP estimates also holds promise for reducing costs.

QALY practitioners can rightly argue that their weighting studies have already been completed and that the cost of developing scores of various changes in health states is trivial. The force of this point ultimately rests on the credibility of the original weight studies.

Characterization of Uncertainties

Ideally, practitioners want to be able to characterize the uncertainty of their estimates of health benefits. Indeed, OMB's guidance document (Circular A-4, 2003) would require agencies to do probabilistic cost–benefit analysis for major rules with costs of $1 billion or more (see Chapter Five). On this attribute, WTP measures appear to have an advantage. To the author's knowledge it is not common practice among QALY practitioners to carry over standard errors on weights into their scores, although this happens with some frequency through sensitivity analysis. Additionally, a number of studies have addressed the use of statistical methods and presentation of uncertainty in CEA (Examples include Hoch, Briggs, and Willan 2002; Briggs, O'Brien, and Blackhouse 2002; and O'Brien and Briggs 2002).

Inclusion of Averting Behavior

Government interventions to improve health will not only lead to improved health directly (if successful), but may also lead to changes in behavior for those who had taken steps to reduce their risks. For instance, if pollution was known to be a problem and people were taking steps to avoid the pollution (for example drinking bottled water, buying organic food, staying indoors on bad air days, and so on), the knowledge that pollution has fallen might induce a benefit in the form of reductions in these averting behaviors. WTP measures can, in principle, capture such effects (and any other nonclinical effect), while these effects lie outside the paradigm underlying the development of weights and their use in QALY scoring. Other things being equal, a WTP measure would be larger if it were able to capture the benefits of less-averting behaviors.

Inclusion of Qualitative Elements of Risk

It is becoming more widely recognized that nonquantitative aspects of risks, such as dread, involuntariness of exposure, and other intangible qualities (Slovic 1992) are important influences on the perception of the seriousness of a given quantitative risk and can therefore affect WTP values as well as weights for health states used to score QALYs. The main issue is that these elements may be embedded in weights and WTP values to a degree unknown to researchers using cues that the researcher may not intend. There is also the largely philosophical issue about whether such elements of risk should be included in the public policy debate, particularly if they are based on misinformation.

Economists have begun applying SP approaches to study this issue. In one study (Jones-Lee 1991), the willingness to pay for cancer mortality reductions in the general public was found to be twice that of reducing deaths by heart attack and three times that of reducing deaths in automobile accidents. Strand (2002) also found that preventing deaths from environmental causes was more highly valued than preventing deaths from heart attacks and auto accidents. Aimola (1998) found that preventing deaths from leukemia was more valuable than preventing deaths from lung cancer. And, in a more comprehensive study, Cookson (2000) found that reducing deaths from air pollution was more valuable than preventing those from food poisoning and transportation-related accidents. Yet, all of these studies have a variety of flaws and certainly exact relationships are open to debate.

The author is not aware that QALY practitioners have taken up this issue.

Effects of Choice of Measure on Values

Throughout this paper, references have been made to the effects that choices of particular valuation methods will have on values, according to various attributes of the population and the health endpoint being affected by the policy. In this section we put this information in one place, as shown in Table 4-3. Results in this table for WTP are quite tentative for most of the entries because the literature addressing these attributes is thin.

WTP and QALY measures treat the preferences for various attributes of health risk reductions in very different ways. The same could be said for preferences of different demographic groups. Table 4.3 lists only attributes where differences may be found.

TABLE 4-3

Relative Value of Reducing Health Effects/Risks

Attribute	*QALY*	*WTP*	*Notes*
Baseline Life Expectancy	Assumed proportional; increase	No assumption; ambiguous empirically	VSL gives greater weight to shorter life span
Life Extension	Assumed proportional; increase	No assumption; less than proportional or zero empirically	Fewer QALYs for older and ill people
Health Improvements (in general)	Increase	Increase	
Acute, Mild Effects	May be insensitive	No issue, but treated as certain	Unwillingness to engage in SGs involving death
Serious Chronic Effects	Integrated with mortality; double jeopardy not usually addressed	Not integrated with mortality; double jeopardy addressed	
Wealth	No effect	Increase	May be embedded in QALY
Qualitative Attributes	May be embedded	May be embedded	New WTP studies suggest an increase
Altruism	No effect	Increase	Issues with double–counting
Latency	Use assumed discount rate	Recent SP studies ask WTP directly	

Concerning utility calculations for policies improving health endpoints, QALYs are proportionally larger for people with longer life expectancies, other things being equal. The effect of life expectancy on willingness to pay is ambiguous, but most studies suggest that the relationship is not proportional and might even be zero, the latter meaning that elderly people would place a value on a given chance of a health improvement equal to that of younger people.

An increase in life expectancy from a policy proportionally increases QALYs by assumption but may not result in any increase in WTP. Further, older people (those with lower life expectancy) will get proportionally fewer QALYs from a life-extension policy than younger people. There may be no difference in WTP for older people for life extension, however. In addition, ill people (whether they have lower life expectancy or not) will get proportionally fewer QALYs from a life-extension policy, by assumption. WTP for life extension by ill people has not been found to be lower than that for healthy people.

Thus, the standard use of quality-adjusted life years as opposed to willingness to pay as an analytical tool implies that benefits to currently healthy, younger people are favored over benefits to currently elderly and, in the case of mortality-risk reduction, sick people.

QALYs (other than multiattribute approaches) are thought to have a problem in capturing preferences for reducing mild, acute effects. People are unwilling to engage in SGs or TTOs where they must consider the possibility of death in a lottery involving a mild symptom. WTP studies do not have this problem. However, such studies (all stated preference) have adopted the convention of treating the acute health reductions in a setting of certainty (asking people how much they would be willing to pay for one less symptom-day, for instance)

Issues with properly capturing chronic morbidity effects are shared by both QALYs and WTP measures. QALYs integrate morbidity measures with mortality measures as a major feature of the index. However, in the application of QALYs, a "double-jeopardy" problem occurs. According to Hubbell (presentation at RFF Conference 2003), this issue involves not taking into account the effect on mortality risk of changing the likelihood of getting a chronic illness.[21]

WTP analyses to date almost always address morbidity and mortality endpoints separately. This procedure has been adopted for practical reasons, not conceptual ones. It is recognized that individuals view their health states over the life cycle holistically and may therefore have difficulty providing WTP responses to changes in mortality risks without reference to or influence from thoughts about the morbidity preceding this change. One of the few studies of the willingness to pay to reduce risk of developing a chronic disease (Viscusi, Magat, and Huber 1991) does add the increased risk of death from the disease as an attribute, thus addressing the double-jeopardy issue. Another study lets the respondent with a relative who has a chronic respiratory disease self-define its attributes, which may include reduced life span (Krupnick and Cropper 1992). One study in process is attempting a more fundamental integration of morbidity and mortality (see DeShazo and Cameron 2003).

QALY calculations ignore income and wealth, but such factors could be affecting the way individuals answer standard gamble or other questions. Sculpher and O'Brien (2000) argue that income can influence the measurement of health status scores, while David Feeny (personal communication) stated that HUI3 and SG scores and weights have been found to be unrelated to income. Use of a WTP measure implies that benefits should be directed to wealthier people unless, as is standard practice, the willingness to pay of average income groups is used.

Weight and WTP studies do not generally explicitly address dread, controllability, and other nonclinical aspects of health improvements. From several studies directly on the issue, the WTP for reducing cancer death is likely to be far higher than for reducing other types of death, say, from an auto accident.

Weights and WTP estimates are usually derived in terms of benefits to individuals. Exceptions are weights derived by the PTO approach and a few studies that have examined the WTP for health improvements to a community or even focused directly on altruistic values. Community WTP values have generally been found to exceed individual values. There is a debate, however, over how much of such altruistic values should count in assessing the benefits of a policy.[22]

Where health improvements are latent, or occur over a period of time, the standard approach in deriving QALYs is to use a discount rate to express the measure in present value terms. There are two different approaches in use with WTP measures. The first is the same as for QALYs, where the monetary benefits of health improvements or the unit values are discounted to account for timing of benefits. The second, newer approach used in the SP literature is to ask individuals directly about their willingness to pay for either a time stream of health improvements or health improvements that begin at a certain age (Alberini, at al. forthcoming; Krupnick et al. 2002; DeShazo and Cameron 2003). With this method, no discount rate is assumed. Rather a discount rate can be derived using information on the current WTP for a future risk reduction, on the current WTP for an equal risk reduction experienced currently, the age of the people answering the two WTP questions, and their self-assessment of the likelihood they will survive to experience the future health benefit.

■ ■ ■

CHAPTER FIVE

OMB Guidance Document

OMB's regulatory guidance document (Circular A-4, September 2003) proposes a number of changes to regulatory-impact analysis procedures currently followed by agencies, with particular implications for the use of CEA and CBA, as well as WTP and QALY measures. Below, these changes are reviewed. Issues unrelated to the topic of this paper are not considered.

The main point of the document is to boost the standing of cost-effectiveness analysis. Specifically, OMB calls on agencies to present both CEA and CBA in support of all major proposed rules. More specifically, the guidelines emphasize a call for CEA for all major rulemakings where the majority of benefits are improvements in health and safety "to the extent that a valid effectiveness measure can be developed to represent expected health and safety outcomes" (p. 126). Cost–benefit analysis should be conducted for major health and safety rulemakings "to the extent that valid monetary values can be assigned to the primary expected health and safety outcomes" (p. 126). CBA is justified because it "a) provides some indication of what the public is willing to pay for improvements in health and safety and b) offers additional information on preferences for health using a different research design than is used in CEA." (p. 145). For regulations that are not focused on health and safety, CBA must be done, and a CEA should be performed when some of the benefits cannot be expressed in monetary units.[23]

With more CEAs to be done in the future and many different types of effectiveness measures available, OMB could have offered guidance on which measures to choose. But it did not, leaving such decisions to the agencies and making reports like this one, as well as other articles in the literature that examine the features of QALY measures and compare WTP and QALY measures, both of which are important for practitioners.

Given that agencies will be able to choose the measures of effectiveness they feel are most appropriate for any given rule, OMB recognizes that comparisons of cost-effectiveness across rules, both within and across agencies, will be difficult. Therefore, the guidelines require agencies to provide OMB with all relevant data, including morbidity and mortality, population data, and the severity and duration of the conditions evaluated, so that OMB can make comparisons across rulemakings that employ different measures. Whether OMB will have the funds and staff for this task and be able to derive consistent cost-effectiveness estimates remains to be seen.

For WTP measures, OMB's most dramatic change is to enjoin agencies not to use the so-called senior discount when valuing statistical lives of the elderly. Previously there were no requirements on this topic. This discount comes out of recent stated-preference research which shows that elderly people are willing to pay less than younger adults for a given mortality risk change. This finding is by no means robust, however, as other research finds the same effect but also that it is not significant. In addition, OMB suggests that agencies may want to use a statistical life year approach to valuation, which gives seniors an arbitrarily larger value per year.

OMB also recognizes more than previously that determining the value of a statistical life may involve examining many factors, rather than being "one size fits all." The guidance document notes that if a VSL is being applied in a different context from which it was derived, appropriate adjustments should be made and the differences in contexts should be outlined: "You should explain your selection of estimates and any adjustment of the estimates to reflect the nature of the risk being evaluated. You should present estimates based on alternative approaches, and if you monetize mortality risk, you should do so on a consistent basis to the extent feasible" (p. 148).

For monetizing health benefits from nonfatal health-and-safety risks when WTP estimates do not exist, the guidelines suggest that scores from health-utility studies can be used in conjunction with existing monetary values in order to estimate monetary values for a wide range of health states that vary in severity and duration. They do caution, however, that analysts should acknowledge the assumptions made in and the limitations of their approach. Whether this is a good idea should be a question for further research. The degree of success of the Johnson et al. (1997) study that applied this methodology is an open question.

Another potentially major addition to the guidelines is the requirement that relevant outcomes should be reported as probability distributions when possible, and that for rules with an economic impact greater than $1 billion, a formal quantitative analysis of uncertainty is required. This means that probabilistic analysis should be applied throughout the entire range of calculations, such that uncertainty is carried through to the endpoints. This factor is particularly relevant to the choice between WTP and QALY measures, because the former is generally expressed in probabilistic terms, while reporting of uncertainty for the latter is far less frequent.

Finally, OMB emphasizes, as it has done in previous guidelines, that the credibility of WTP measures would be enhanced if the studies generating these estimates met a series of criteria. These comments should apply equally to QALY measures, with OMB charged to identify validity tests that would enhance QALY's credibility.

■ ■ ■

CHAPTER SIX

A Research Agenda

The research agenda that emerged from the RFF conference and workshop presentations and discussions may be divided into research within the particular valuation paradigm or tool (such as WTP, QALY, COI) and research involving the integration or cross-fertilization of paradigms or tools. As the author is more familiar with the WTP/COI paradigms, this research agenda may not do justice to research opportunities within the QALY paradigm. This list is obviously not comprehensive, but is merely suggestive of the research that would productively bear on the issues taken up in this paper.

One overarching research task emerged as well. There is a need to obtain information from decisionmakers on their heuristics for addressing the policy issues posed in Chapter Three. Also, they should be asked about how much information is too much and what they need analysts to do to sort out complex and often contradictory information.

Willingness to Pay

As noted above, there are a number of developments in stated-preference techniques that hold the promise of improving WTP estimates of health effects, including

- greater use of choice experiments to develop values for attributes that can then be combined to develop WTP estimates for a wide range of conditions;
- irrespective of the SP techniques, more attempts to develop life-cycle estimates of WTP, combining morbidity and mortality;
- SP studies of the effects of qualitative risk attributes on WTP;
- more studies of age, health status, latency, and size of risk change on WTP; and
- more studies of individual WTP for community-wide health improvements.

Note that there is at least one choice experiment study that attempts to estimate transferable WTP estimates for a wide range of acute cardiovascular and respiratory symptoms that have been epidemiologically linked to air pollution exposures (Johnson et al. 2000). This study employs multiattribute conjoint methods to elicit respondent trade-offs among multiple symptoms, severities, and durations. The resulting estimated indirect utility function can be used to construct welfare estimates for a large number of health states, controlling for various demographic characteristics of respondents.

Cost of Illness

Recent strides by ERS and USDA to systematize estimation of costs of illness are extremely important. But, such research only considers health risks from foodborne pathogens. Similar scrutiny is needed for other types of risks.

Quality-Adjusted Life Years

Much of the debate between experts on the WTP side and those on the QALY side concerns the protocols for a credible study. Thanks to the NOAA Panel (Arrow et al. 1993) and a number of in-depth treatments (see Chapter Two), the protocols for WTP studies have been reasonably well articulated and addressed in the modern SP and RP literature. Similar protocols do not exist to this extent with QALY measures when it comes to the conduct of design and conduct of surveys to derive health state weights (see Gold et al. 1996, however). OMB's call in its new CBA Guidance for the Institute of Medicine to examine this issue is only the first step in building and implementing such protocols.

Other items on the agenda include:

- Develop QALY estimates that are particular to age groups and groups with different health conditions, different income levels, and other factors, both to examine the influence these factors have and to make the implementation of QALYs more sensitive to equity concerns.
- Address the double-jeopardy issue.
- Add extrinsic environment to weight surveys. This idea (Patrick presentation at RFF Conference 2003) encourages QALY researchers to broaden the choice context in which weights for alternative health states are elicited.
- Consider side constraints on QALY indices (Smith presentation at RFF Conference 2003).

Smith's suggestion needs some elaboration. His argument recasts the outcomes of the health state choices (using SG or TTO approaches, for instance) as choices explainable by a conventional preference function. Combining these responses with information derived from SP and RP studies then allows one to calibrate a preference function (see Smith, Van Houtven, and Pattanayak 2002). Once the parameters of the preference function are determined through calibration, consistent quantity indices can be derived from it for use in cost-effectiveness evaluations. The function can also be used to derive benefit measures in evaluating new policy alternatives.

The proposal can be implemented in several ways: (a) using the choice outcomes of a QALY study; (b) defining the summary statistics provided as measures for QALY indices in terms of a consistent preference function and assuming that they correspond to the average respondent; or (c) using estimation methods (such as the generalized method of moments) to "fit" several sets of responses to the equations defining QALYs that have been expressed in terms of the preference function (see Smith et al. 2003).

Integration and Cross-Fertilization

Perhaps the most exciting options for further research are in the integration and cross-fertilization of the QALY and WTP traditions. There has already been some preliminary work on the former. In particular, two meta-analyses statistically link published WTP estimates to imputed QALY weights to derive money equivalences of changes in health states (Johnson et al. 1997 and Smith et al. 2003). The primary advantage of estimating a statistical relationship between outcomes measured on QALY and WTP scales is that analysts can construct WTP for any health outcome within the range of the data in the study, even if there is no specific empirical estimate available for that particular outcome. Implementing such an approach is subject to several caveats, however, according to Johnson (personal communication):

> *(i) the statistical function that links QALY weights to WTP is strictly empirical. Neither the QALY studies from which the weights were obtained nor the WTP studies from which the money values were obtained were intended for such purpose. The conceptual and empirical basis for the two types of estimates is quite different;*
>
> *(ii) although the number of WTP estimates available for analysis increased substantially between the 1997 and 2003 studies, the number of observations available remains small, with limited and uneven coverage over the range of health outcomes that might be of policy interest. In both studies, both older and newer results were pooled with little attention to variation in quality. Some of the WTP studies employed outdated methods that have been rejected in light of subsequent validity concerns;*
>
> *(iii) in both cases the authors employ the Quality of Well Being (QWB) health utility index because it is easy to use for imputing scores based on the conditions valued using WTP estimates. However, the scale is relatively coarse scale and insensitive to small variations in health states, particularly acute health states. The imputations themselves are subject to uncertainty that is not reflected in the analytical approach.*

Another promising research area is suggested by the Johnson et al. (2000) study. In this study, an indirect utility function for health outcomes is derived, suggesting a possible bridge between ordinal utility-based welfare measures widely used in the WTP tradition and cardinal utility-based health state measures in the QALY tradition. Stated-preference studies rescale utility differences by the estimated marginal utility of money to get WTP. However, Johnson et al. (2000) suggest that utility differences can be rescaled by any continuous attribute, including time. If the reasonable assumption is made to put duration of health states in the utility function, time-equivalent willingness to wait (WTW) measures of welfare changes can be derived. WTW incorporates exactly the same preference information about relative trade-offs as WTP does, but may avoid objections to valuing health in monetary terms.

According to Johnson et al., WTW also can be converted to time-trade-off QALY weights if the indirect utility of normal or perfect health can be defined. Utility differences can be constrained to vary only by observed clinical outcomes and other restrictions required by conventional QALYs can be imposed, or all observed attributes that affect utility can be included. These generalized or "super" QALY weights incorporate full, nonlinear utility and thus improve the welfare relevance of QALY measures.

Following the logic of QALYs used in cost-effectiveness analysis, Smith has a proposal that adopts the logic underlying conventional economic index numbers instead of QALYs. This approach would construct a weighted index of cases (or days) of specific health outcomes. The weights would be defined using what might be termed equivalent health income (EHI). This EHI is defined as the sum of the cases of health outcomes (such as restricted activity days or hospital admissions) in the baseline year, each weighted by the unit benefit measure developed for morbidity benefit measures. The weight for each category is the share of the EHI due to that component in the base year. This approach has the advantage of paralleling the logic used in developing quantity indices for market goods and offering a simple interpretation for cost effectiveness measures—as the cost per "equivalent case."

As for cross-fertilization, health domains and profiles commonly used in the QALY tradition could be valued using stated-preference methods. Generally, WTP values are developed for symptoms rather than functional limitations, the latter being more in evidence in the QALY literature. Additionally, the valuation of children's health can benefit from QALY protocols. Many researchers (including Schlottmann 2001; Harbaugh, Krause, and Vesterlund 2002; and Hargreaves and Davies 1996) have shown that all but the youngest children are very capable of ranking their health states and that these rankings differ from those their parents have for them. To be sure, this finding raises important philosophical issues about whose preferences matter (see Hoffmann, Adamowicz, and Krupnick 2003) and says nothing about whether children could make credible health–money trade-offs. But it also raises the possibility of using QALY approaches to communicate such rankings to their parents, who then could use this information to make decisions about their care (and implicitly to value their child's health).

The cross-fertilization can also work in the other direction. Study-design protocols and credibility testing are more advanced in the SP WTP literature than in the QALY literature. These protocols could be adopted and adapted by QALY researchers. Further, the untested role of income and other currently unobserved heterogeneity in samples taking "weight" surveys could be explored for the purpose of improving comparability between the two types of measures.

▪ ▪ ▪

CHAPTER SEVEN

Toward Implementation

The new OMB Guidance for RIAs calling on agencies to rely more on cost-effectiveness analysis along with cost–benefit analysis will challenge experts, practitioners, and decisionmakers to make the best use of these tools in designing and approving policy. This paper takes a step towards meeting that challenge by examining in detail the appropriate use of CEA and CBA tools and the credibility and usefulness of the underlying valuation measures based on WTP, COI, QALYs (and multiattribute systems such as HUI). It begins to answer two basic questions: What does each tool and measure bring to the decisionmaker that the other doesn't? How might the decisionmaker's core beliefs about criteria for evaluating regulations influence his or her reliance on one tool or measure more than another?

What Does Each Tool or Measure Bring to the Decisionmaker?

CBA versus CEA

CEA (cost divided by some nonmonetary measure of benefits) is for all intents and purposes a subset of CBA. The cost terms are identical, and underlying whatever benefits measure is used in the "denominator" are physical effect changes that may themselves appear in cost-effectiveness analyses or be inputs to QALY indices. Therefore, in distinguishing between CEA and CBA, the real issue is the difference between the monetary benefit measure and the various effectiveness measures used in the "denominator." There is also an important, though subsidiary, issue—differences in the subcultures supporting the use of these tools.

Turning to the primary issue, the distinguishing feature of these two sets of measures is that the monetary benefits measure permits calculation of net benefits (benefits minus costs), which can claim to represent changes in social welfare, while the effectiveness measures do not produce defensible net benefit estimates.

This statement needs a number of clarifications. First, if benefits are calculated using COI values rather than, or in addition to, WTP estimates, they would lose claim to being a social welfare measure, unless it is clear that the COI is a lower bound to the WTP measures being used. Second, if the effectiveness measure is a QALY index, it can be converted to dollars using several different, essentially arbitrary techniques. In the judgment of most of the conference participants, use of such a conversion approach would not be defensible.

Nevertheless, the cost-effectiveness tool is still useful for comparing alternative regulatory designs, generally within the same regulatory effort. This qualification is offered because such comparisons demand that the effectiveness measure in each calculation be the same. Of course, CBA also permits comparisons across regulatory designs. Also, monetizing health benefits is controversial, so offering CEA comparisons along with CBA comparisons gives the decision-maker a way of limiting this particular controversy—leaving, of course, many others.

The secondary issue is that the subcultures underlying use of CEA and CBA are very different. CEA, in the context of the current debate, represents what is commonly termed cost-utility analysis and has been extensively used over the last two decades to compare the effectiveness of alternative medical interventions. It focuses on maximizing health outcomes per dollar of cost. CBA was developed much earlier and for policy applications—to compare social welfare implications of projects or policies in any sector—but has been applied to health issues only in the last several decades. Thus, it focuses on maximizing social welfare rather than health.

WTP versus COI

According to experts at the conference, WTP studies can provide reasonable and credible social welfare-based estimates of value for some health endpoints but not for others. The labor market studies provide a particularly robust set of studies on the VSL, with a growing body of SP studies on this—generally the most important—health endpoint.

The COI approach does not purport to be a measure of individual or social welfare, since it makes no attempt to include intangible but real costs, such as those associated with pain and suffering. Its advantage over willingness to pay is that the concept of cost-of-illness is relatively transparent and easy to estimate. While COI measures are generally at least several times lower than willingness to pay estimates, this relationship is not a necessary one. If it were, one could perform benefit analysis with COI measures, or a mixture of WTP and COI measures, and then make comparisons among regulatory alternatives knowing that a finding that benefits were greater than costs would not be reversed if WTP replaced COI.

QALY Indices and Weighting Approaches Compared

This paper has not been able to do justice to the great variety of indices available for effectiveness analysis. We used the term QALY index to stand for all these indices, where each one is developed from specific surveys that provide the preference weights for health states (or domains), which together with duration estimates permit calculation of the QALY index.

QALY indices are in use throughout the world, primarily to examine the effectiveness of medical interventions. Whether they are ready for and appropriate for use in a policy setting is an open question.

As measures of social welfare or utility, in the sense that economists mean by the term, none of the indices pass the test, but some have more utility-based content than others. In particular, indices based on the standard gamble to develop weights are favored because they incorporate the notion of trade-offs and some notions of risk. Indices based on the person-trade-off approach to weighting also have some appeal, as this weighting approach is the only one that may address individual preferences for effects to the community, in contrast to the standard gamble, which is concerned with individual preferences for one's own health.

Indices that use weights derived from standard gamble, time-trade-off, or rating scale approach are typically considered "preference-based" within the QALY literature, though some use the term "utility-based," which due to the reasons emphasized above, is a term the author hesitates to use. Under this umbrella, there are differing approaches. For example, the Quality of Well-Being (QWB) index scores health states using an additive formula with weights derived using a visual analog scale, a type of rating scale. The Health Utilities Index II and III scores health states using a multiplicative formula, with weights from visual analog scale responses transformed into SG values. Additionally, the Healthy Years Equivalent (HYE) has been proposed as alternative to QALYs that measures utility with fewer restrictive assumptions. There is currently no consensus in the QALY literature (defined at its broadest) regarding the best or most appropriate index to use for different types of analysis. Interestingly, as a follow-up to the issue of its new guidelines, OMB is asking the Institute of Medicine to assemble a panel of experts this spring to provide guidelines for using health status measures, such as which measures are the best and in what circumstances.

The literature on health effects valued with QALY indices is extensive and has been catalogued at Harvard University. Not surprisingly, most studies use the simpler, less defensible indices because they are easier to understand and more useful to score aspects of disease.

There are also the psychometric measures, such as those developed from Short Form-36, which describe aspects of health functioning. The health status questionnaire for SF-36 does not involve trade-offs or changes in health status, which are important elements of the policy context.

WTP versus QALYs

The paper has devoted the most space to this comparison, primarily because agencies tend to use and be familiar with one measure to the exclusion of the other. There is no simple conclusion regarding this comparison. The paper defines the appropriate comparison as one between the credibility of the underlying weights on health states in calculating QALYs and the credibility of the WTP estimates for individual health conditions. On this dimension, the literature on refining WTP approaches to improve their credibility is more advanced than that for refining weights underlying the QALY indices. However, on another salient dimension—the comprehensiveness of estimates—QALYs do better, generally because any given survey to develop weights covers many types of health states that can be repackaged as a particular health effect is redefined. WTP methods generally apply to one health effect at a time, but newer studies taking the “choice experiment” approach promise to develop “prices” for a variety of health attributes.

There appears to be consensus that WTP measures are better at capturing preferences regarding acute health effects and can, at least in theory, capture qualitative attributes of risk (voluntariness, dread, and so on) that are not captured in standard surveys to derive weights used to derive QALY indices. WTP measures can also be applied in consistent fashion to nonhealth effects, say to the effects of acid rain on fish and trees as well as the effects of fine particulate aerosols (created by the same emissions) on health. For QALYs to incorporate the former effects in a CEA would require use of WTP techniques to monetize them and then subtraction of this amount from costs. This hybridization of QALYs and WTP in one analysis may be more confusing than enlightening.

There also appears to be consensus among our conferees that many of the shortcomings of QALY literature could be remedied by using best practice or even reforming some practices. Two examples are the treatment of uncertainty and credibility of the valuation measures. The WTP literature places great effort on both of these features of WTP estimates, reporting uncertainty of such estimates and performing a variety of content validity tests as a matter of course. The QALY literature pays attention to these factors when comparing scores of different indices or their components through use of sensitivity analyses and other more heuristic techniques. However, according to critics of the tool, little attention is currently paid to analyzing uncertainty and credibility of weights.

Decisionmakers' Criteria and Their Effect on Choices of Tools and Measures

Short of a major study of decisionmakers in federal agencies, it is hard to know what criteria they use to judge the weight to be applied to all the disparate results contained in a typical RIA. Thus, the strategy has been to list various criteria that could be used and then to show how they might influence the choice of measures and tools. This discussion was covered mostly in Chapter Three and is summarized here.

Efficiency

Efficiency has two dimensions in the context of regulatory analysis.[24] The first is the normative dimension, that is, does the regulation generate positive public benefits and does this particular regulatory design generate the largest public benefits? The second is a relative dimension, that is, does this regulation rank highest in incurring the lowest social cost per chosen measure of effectiveness. Under the first perspective, CBA using WTP measures of value is favored. Under the second perspective, the choice of type of analysis is unclear because both CEA and CBA rank alternative regulatory designs.

Equity

The tools themselves have no equity implications; these implications are indirect through their use of a theory to value costs (which this paper has ignored) and benefits.

One of the major equity concerns with valuation is about the role income plays. Because WTP measures are developed with preferences constrained by income, this measure is expected to affect values, with those with higher incomes holding higher values for health improvements, other things being equal. In practice, such relationships are not always found. Further, in government policy analysis, the standard practice is to use average WTP estimates, hence the resulting CBA does not discriminate against any income groups. Or, said another way, WTP values differentiated by income class are not generally given regulatory standing.

But income is not the only socioeconomic factor that may influence WTP. Because WTP measures are derived from individual preferences, differences in values are typically searched for and may be found across many different types of groups in society. For example, older people may exhibit a lower value for reducing their mortality risk than younger people, African Americans may exhibit a higher value for health risk reductions than Caucasians, women might hold higher values than men, and so on. Whether or which of these preferences should have standing in regulatory analyses is another matter (see below). To the extent that such preference differences are ignored in a regulatory analysis, the welfare basis of WTP measures is impinged.

QALY measures are also derived from individual preferences, but have more constraints on how these preferences are expressed and aggregated than the WTP measures. In standard practice, whether or not weights for different health domains are affected by group characteristics—such as age, gender, race, and other factors—is not usually examined but could be. In the algorithm for computing QALYs, it is generally the case that the disabled, ill, and old are discriminated against when the regulation promises increases in life expectancy (other things being equal). Adjustments could be made to eliminate such biases, as has been suggested by OMB in Circular A-4 with respect to the bias against the disabled and ill.

Individual versus Social Perspective

Welfare economics places individual preferences at the center of the "story," but the storyline is mostly about a market economy unfettered or minimally constrained by government. In a market context, consumers are "sovereign" and their preferences create demand for goods. Government should not interfere with this demand if it wishes to maximize social welfare—as the sum of individuals' welfare. Where government is setting regulations to address market imperfections, this presumption of consumer sovereignty is weakened somewhat.

Whether individual preferences should be at the core of government decisionmaking is an open question. Individuals have preferences for their own health states; they also have preferences for what happens to others—in their household, in the community, and so on. Either or both of these objects of individual preferences may be important for decisionmakers to take into account. They may also want "social preferences" to play a role in their decisions. To do this they may want to rely on equity paradigms, such as those listed in Appendix II.

Which approach is chosen affects how QALY and WTP measures are viewed. WTP values are based on individual decisions about their own health. QALY indices based on PTO weights might better address valuations where individual preferences for improving the health of others in a community are deemed important.

Underlying the choice between an individual and social perspective may be, at least in part, beliefs about the reliability of individual preferences. Certainly, there is ample evidence of gaps between individual perceptions and scientific estimates of risk and other factors relevant to valuing health. Some of these differences may be cognitive difficulties, say in understanding probabilities, which would argue for reducing the weight given to individual preferences. But, differences between individual and scientific assessments may be valid because of the well-known phenomenon that individuals imbue risk preferences with many qualitative attributes, such as degree of voluntariness and dread, which lie outside of the standard probabilistic treatment of risk. Both WTP and QALY approaches are conceptually indistinguishable on this perception issue, although this issue has received far more attention by economists in the WTP literature.

Health versus Utility

If an aggregate measure of health changes associated with a regulatory design is the preferred measure of effectiveness, then this favors QALYs. Needless to say, describing changes in social welfare, not just health, would seem more important in a policy context, and this view would favor WTP measures.

Avoiding Controversy

WTP approaches that value health and mortality risk have generated enormous controversy, although conferees agreed that this was unwarranted and resulted in large part from a misunderstanding of what is being valued and how it is being valued. QALYs have not received nearly this level of public reaction.

Completeness

Developing WTP measures is more labor-intensive than developing QALYs because the latter approach provides weights for many different health states or domains in a given survey while WTP studies yield values for at most a few health endpoints at a time. Attempts to develop benefit transfer approaches to extend the range of health effects valued by WTP have not generally been successful. Choice experiment or conjoint analysis techniques—where attributes of a health state are valued—may prove useful for this purpose. Scores for health states are generally more available than WTP values, although they tend to be for endpoints that are more detailed than those typically specified in epidemiological studies used in RIAs.

Credibility

A large part of Chapter Four was devoted to analyzing the credibility of the various valuation measures. The bottom line is that, under a welfare economics paradigm, WTP measures are theoretically more credible than QALYs. QALYs provide a valid utility measure only under very restrictive conditions. WTP measures using stated-preference techniques are beginning to be viewed as more credible for use in evaluation of policy interventions related to health than those using revealed preference techniques, either because of the paucity of the latter or because the market behavior being studied is too far removed from the policy context. Typically, WTP measures derived from RP techniques are based on wage-risk trade-offs, which differ in context from most environmental and health risks in the populations and the nature of the risk at hand. SP methods can be more easily tailored to a particular context, thus potentially providing more credible estimates for policy evaluation. Finally, within the literature, much attention has been devoted to validating the credibility of WTP measures. Within the QALY literature, however, treatment of the issue of credibility (at least according to critics) has been far less extensive, though this is a flaw in the literature rather than the measure itself.

Transparency

Perhaps the least transparent of the WTP measures is the value of a statistical life. VSL measures are frequently misinterpreted as representing a market value for human life, rather than their true interpretation, which is a value for a statistical life, derived by aggregating individuals' willingness to pay for small changes in risk. The metric, then, is representative of choices and trade-offs individuals make (or say they would make) in the face of risk. WTP measures for acute effects are more transparent in that they are simply the average willingness to pay to avoid an illness.

On one level, QALYs are more transparent than willingness to pay measures, as they are simply the product of two components: a health state score and its duration. What is less transparent about QALYs, however, is the information underlying the health state scores. As this paper has demonstrated, health state scores can be taken from a number of different indices, which in turn are developed using a number of different approaches to deriving preference weights (ratings scale (RS), time trade-off, and standard gamble among them) that have implications for health state scores (just as using RP or SP methods for WTP has implications for VSL values).

Consistency

Consistency can be assessed along two dimensions: consistency of the approaches used in studies and consistency in values across studies. WTP measures tend to be reasonably consistent regarding approaches—the two in use (SP and RP) are derived from the same economic theory, and both have been standardized (and improved upon) within the literature. The QALY literature also recognizes a number of approaches to calculating health status scores, but they do not descend from the same theoretical origins. While SG is based on expected utility theory, TTO and RS are not.

The value of a statistical life (VSL) estimates based on a willingness-to-pay range roughly from less than $1 million to greater than $15 million. Values from stated preference (SP) approaches, on average, are lower than those obtained in revealed preference (RP) studies. Within SP studies, variance in VSL can often be attributed to the size of the risk being studied, since individuals do not tend to adjust their WTP proportionally to the size of the risk. With RP studies, the differences are often attributed to model specification and data sources. Similar variance exists in QALYs, though, of course, the bounded nature of the index (generally between 0 and 1) limits the variances relative to that for WTP values. SG tends to yield higher scores than TTO and RS approaches. Also, while different indices may be correlated with one another, they tend to vary in their sensitivity to changes in health status.

Concluding Thoughts

Regulatory decisionmaking at agencies will always be messy as decisionmakers evaluate the incomplete and uncertain information given to them in RIAs and, at the same time, attempt to meet their legislative requirements, respond to the pulls and pushes of their staffs and stakeholders, and take their own readings about what is best for society. The new, more complex RIAs that are to be produced under Circular A-4 could lead to better decisionmaking because of the multifaceted, more complete results that would be provided. But, decisions could be slowed or their quality reduced if information overload and confusion rule. More information is not always better.

More information may be better if it can be clearly organized, underlying assumptions are made transparent, and the advantages of each tool and measure can be clearly spelled out. Unfortunately, there is no simple way to do this. For instance, WTP and QALY measures cannot be unambiguously ranked in their usefulness for policy. Such rankings would depend on what we have termed “policy” or philosophical choices set out in Chapter Three and on the relative weight given to the technical criteria (validity and so on) set out in Chapter Four. Similarly, even if one were to state that CBA was unequivocally “better” than CEA in measuring and ranking regulatory outcomes, the inevitable gaps in valuation and other practical problems posed by CBA that are not present in CEA would discourage sole reliance on that method of analysis. Development of a research agenda to: (i) improve all the tools and measures, (ii) evaluate each new RIA for how it approaches these issues, and (iii) examine how decisions are influenced by RIAs before and after the implementation of the OMB guidelines would make improvements in decisionmaking more likely.

■ ■ ■

Notes

1 Comments on this report were drafted by the author, commented on by members of the Steering Committee and other experts, modified by the author, and sent to OMB. The author's comments to OMB are included as an appendix to this paper.

2 A widely cited example of broad-based cost-effectiveness analysis is a table developed by John Morrall of the Office of Management and Budget (OMB) in the 1980s (Morrall 1986) that looked at the costs per life saved of 44 actual and proposed environmental and health regulations. The analysis found a wide variance in the efficiency of the regulations, with several being hugely expensive in terms of their cost per life saved.

3 The number of life years lost is calculated from life expectancy at a given age. For example, life expectancy for a 25-year-old may be 55 years and that of a 75-year-old may be 10 years. In addition, life years are usually discounted. See a discussion of discounting life years in Murray and Lopez (1996).

4 Calculating QALYs gained from an intervention requires the following steps. First, choose the time period of interest. Second, identify all possible health states with and without intervention within time period of interest. Third, develop weights for each health state either by mapping the health states into domains from the available indices or by using VAS, SG, TTO or another estimation approach. Fourth, determine the duration in each health state with and without intervention in the time period of interest. Fifth, weight each health state and multiply by its duration to compute QALYs in that health state. Sixth, add together QALYs for all health states over the time period of interest with and without the intervention. Seventh, calculate the difference in QALYs attributable to the intervention.

5 It is possible to have a scale where there are states worse than dead, that is, anchored on a negative number.

6 Another approach is the Healthy Years Equivalent (HYE) Index. Because HYEs are derived from each individual's utility function, they are said to fully represent individual preferences. As a result, assumptions are not necessary to equate HYEs with utility (Gafni, Birch, and Mehrez 1993). Calculation of HYEs involves the measurement of preferences over complete life paths as opposed to discrete health states and the conversion of utility measures into HYEs. Critics of this approach (Culyer and Wagstaff 1993) have suggested that HYEs are equivalent to health indices derived using the TTO approach. However, Gafni, Birch, and Mehrez (1993) argue that TTO deals with indifference curves under certainty, while HYEs use a two-stage lottery that deals with indifference curves under uncertainty. Comparisons of QALYs and HYEs in the literature are limited to date.

7 The weights can be determined directly using SG, TTO, and VAS approaches or from a multiattribute system. Different multiattribute systems are based on different functional forms (linear additive, multiplicative) and different methods for eliciting preference scores (SG, TTO, and VAS).

8 The scoring of disease states can be based on the preferences of individuals, but a recent survey of QALY studies found that this has often not been the case; in many studies, physician judgments have substituted for individual preferences (Neumann et al. 1997). The U.S. Panel of Cost-Effectiveness in Health and Medicine (Gold et al. 1996) recommended that QALY scoring should be based on individual preferences from community samples.

9 In a study of nursing home residents who evaluated health states using the VAS, TTO, and SG, Patrick et al. (1994) asked respondents to rate the difficulty of the methods. The results indicated that the SG was the easiest, followed closely by the TTO, and that the VAS was the hardest. In part this may be the result of the complexity of this particular study, where five or six states were being compared on the VAS line at one time. Recently Statistics Canada has obtained the same ranking.

10 For a recent study estimating a VSL from jury awards see Cohen and Miller (2003).

11 In this case, the value of a statistical case of chronic illness is (the WTP for a risk reduction in chronic illness)/(risk change).

12 Recently, DeShazo and Cameron (2003) have administered surveys that ask for preference rankings over life-cycle-based health effects and mortality risks, offering the possibility of monetizing preferences for mortality and morbidity holistically.

13 For instance, in measuring the WTP to reduce air pollution using housing price variation over space, the physical effects measure is embedded in perceptions of homebuyers and sellers about what would happen to their health if they lived in homes at locations with different degrees of air pollution. Because this approach uses public perceptions of dose-response relationships rather than scientifically based relationships, it has fallen into disuse in favor of the damage-function approach.

14 The NOAA Panel was convened to sort out competing claims about the credibility of CV surveys on existence value in the wake of the Exxon *Valdez* oil spill in Prince William Sound, Alaska.

15 More recent variants of this approach are gaining interest. For one, net benefits are calculated conditional on an assumed range for $/QALY factors. An allied concept is the cost-effectiveness acceptability curve where one computes the probability distribution that the cost-effectiveness estimate falls below some threshold.

16 This approach is described more grandly as transcending traditional welfare measures by not merely focusing on individual utilities, but including nonhealth aspects such as effects on relationships with individuals and whether they are happy and free of pain. This extra-welfarist measure would seem to include the very factors that make up utility but that are being rejected by such views.

17 For instance, David Feeny (personal communication) found in his study of genetic testing for carrier status for cystic fibrosis, that scores differed among states such as know you are a carrier, know you are not a carrier, uncertain—all with the same “physical” health status.

18 Feeny (personal communication) states that path states and sequences of states can be used to handle the issues with acute states. Multiattribute systems can handle acute states, as well. Feeny is more concerned about how the indices score chronic conditions.

19 In what follows, the discussion is restricted to estimates of willingness to pay. Another measure is termed willingness to accept (WTA), which is the amount of payment an individual would accept to be placed in a worse health state. Aside from some empirical issues with estimating WTA, this choice context appears to be generally unrealistic when applied to government programs. However, the construct of willingness to accept compensation to forgo an improvement in health or environmental quality is arguably a policy-relevant value measure. Consider the decision to site a waste treatment plant in a neighborhood. One way of conceptualizing the CBA for this choice is to ask what is the cost of siting the plant there versus somewhere else, of which the compensation needed to make people indifferent to whether the plant was sited or not (a WTA measure) would be important. Another is to ask what the WTP of people is to keep the plant out of the neighborhood. The appropriate context depends on who has the property right.

20 See Mauskopf et al. (1988) for one of the first comprehensive attempts to derive such estimates (in the food safety setting).

21 In the cost-effectiveness literature, there is a current controversy concerning the appropriate costs in the numerator. The issue is whether future health care costs resulting from a person’s living longer should be attributed to the policy intervention (Garber et al. 1996).

22 This debate revolves around the concepts of paternalistic and nonpaternalistic altruism. If people

care about other people's health, rather than their utility, this preference is called paternalistic altruism and should be added to individual WTP. Otherwise, there is double counting. See Hoffmann, Adamowicz, and Krupnick (2003) for a discussion.

23 Converting a CBA into a CEA is easy if you have a single health endpoint that generates the benefits part of the CBA. All of the QALY problems kick in when you start doing a CEA of multiple outcomes, conditions, durations, and so on.

24 By efficiency, allocative efficiency is meant, not technical efficiency.

25 If one wishes to delve into the debate regarding preference satisfaction one can begin with the exchange between Sagoff (1993) and Kopp (1993). More strictly philosophical discussions can be found in Williams (1985) and Scanlon (1991).

26 The term "value" as used by economists causes a great deal of confusion. For example, if one asked you what are your "values," you probably would not respond by saying $2 for a Big Mac or $30 for a round of golf at a public course. Rather, when asked about your values, you might say things like honesty or hard work. Similarly, if one asked you what value do you place on the environment, you might say the need to preserve it for future generations, or you might mention your commitment to environmental stewardship and conservation. You would probably not say $32 per day to view bald eagles along the California coast.

27 One may be inclined to say that the analyst estimates the economic value rather than constructs the value. In fact, the economic literature routinely refers to "value estimates." However, we use the verb *constructs* to underscore the notion that economic value does not exist in a freestanding fashion amenable to empirical measurement. Rather, economic value can only be measured with reference to a choice, and the characteristics of that choice largely determine the measured value.

28 To monetize economic value, the forgone alternative (defined by an individual's choice within a specified trade-off) must be expressed in dollars. Unfortunately, this monetization has sometimes created misconceptions. For example, it has been suggested that economic values are confined to prices observed in markets. These misconceptions arise because many people commonly think of the monetary measure of economic value as a *price:* if a widget sells for $6 in a market, then $6 must be its value. This view is misleading, however. When a person buys a widget, the analyst only learns that it is worth *at least* $6 to the buyer. He or she might be willing to pay much more than $6 if necessary to get the widget. Markets do offer opportunities for people to make choices, but it is these choices and the circumstances relevant to them that permit construction of the underlying economic values, not the existence of markets and market prices *per se*.

29 Policies rarely affect only one good or one price. Most often they affect many goods and many prices. But if we knew the economic value of all the goods affected by the policy and the effects on those goods of the policy, we could aggregate the monetary measures of well-being gains or losses across all the affected goods to capture the full impact of the policy on individual well-being.

30 Indeed, with QALYs, the young are already advantaged over the old proportional to their life expectancies. Yet a prioritarian view might imply giving the young an added boost by valuing a young person's life years saved by more than that of an older person.

31 COI measures typically include medical costs, loss in productivity from illness, and premature death losses valued as lost productivity (termed the human capital approach).

32 In this report, "QALYs" is a term used for a wide variety of indices that permit aggregation of morbidity effects in terms of health states and the duration they are experienced, as well as life years lost.

33 Use of QALYs in a cost-effectiveness analysis is sometimes termed cost-utility analysis (CUA), referencing the fact that certain types of health indices are based on Von Neumann-Morgenstern cardinal utility theory.

34 Performing a CEA if already doing a CBA is easy if a single physical health endpoint generates the benefits part of the CBA. The complications for CEA begin when multiple physical health outcomes must be incorporated

35 CBA using WTP metrics is to be distinguished from CBA using cost-of-illness metrics, which cannot make the claim of unambiguously measuring utility. More will be said about this difference below. Unless otherwise specified, CBA refers to analyses done with WTP metrics.

36 Using the "price per QALY" approach to monetize QALYs can provide a net benefits measure, but this approach is problematic. See below.

37 Consider the case of ranking alternative approaches to reducing fatality risks from automobile accidents using cost per life saved. The per life saved measure would ignore morbidity effects of automobile accidents, effects that might well dominate deaths in terms of public preferences.

38 See, for instance, National Research Council. 2002. *Estimating the Public Health Benefits of Proposed Air Pollution Regulations* (NRC, Washington, D.C.)

39 It is convenient here to refer to a score or QALY score as the number of QALYs estimated to be related to some state or intervention. The term "QALYs" is used for the technique for assigning and adding up life years and quality-adjusted life years and also the total score.

40 One potential problem with this approach is that adding up QALYs for life years gained only is transparent where each life year takes a value of one or some other constant value. If life years take on different values according to the health status of the life saved, then transparency will be harder to achieve.

41 The idea of unpacking could also apply to WTP measures. If income matters, computations could be based on different income groups with different VSLs, but also computations could be based on average income. The same could also be done for age and health status, for instance.

42 The conference and workshop participants generated many ideas for research, which will be presented in a future report. A few ideas are presented here: One would involve implementing a conjoint analysis stated preference survey, where the attributes of the survey would include dimensions that would apply to a wide variety of morbidities (see Johnson et al. 2000 for the model and results for a small Canadian sample). One of the issues with this type of study is that the many different dimensions of morbidity may exceed respondent ability to trade-off among the necessary attributes. Another idea (Johnson 2002) is to construct "super-QALYs," by estimating willingness-to-wait estimates for morbidity and convert them to time-trade-off QALY weights. Another research team (Smith, Van Houtven and Pattanayak, 2003) proposes to use results of QALY elicitation along with economic information about behavior to calibrate the health-related components of individual preference functions.

■ ■ ■

List of Acronyms

AHRQ	Agency for Healthcare Research and Policy
CBA	cost–benefit analysis
CEA	cost-effectiveness analysis
CHQ	Child Health Questionnaire
COI	cost-of-illness
CUA	cost-utility analysis
CV	contingent valuation
CVM	contingent valuation model
DALYs	disability-adjusted life years
EHI	equivalent health income
HUI	Health Utilities Index
HYE	Healthy Year Equivalent Index
OMB	Office of Management and Budget
PTO	person-trade-off
QALY, QALYs	quality-adjusted life year
QWB	Quality of Well-Being
RIA	regulatory impact analysis
RP	revealed preference
RS	ratings scale
SF-6D	Standard Form 6D
SG	standard gamble
SP	stated preference
TTO	time-trade-off
VAS	visual analog scale
VSL	value of a statistical life
WTP	willingness to pay
WTA	willingness to accept
WTW	willingness to wait

APPENDIX I

Description of Terms

Social Welfare. Social well-being is the summation of all the things that members of a society see as contributing to the quality of their lives—individually and collectively—without enumerating what those factors might be. However, to develop empirical measures of well-being in CBA, we need a concrete definition of well-being. To avoid confusing the abstract notion of well-being with its operational counterpart, we will hereafter term the latter social welfare. Unlike the components of well-being that are left vague and open to interpretation, the components of social welfare included in CBA must be clearly delineated. These components are individual welfare aggregated in particular ways that generally (but do not have to) take the existing distribution of income as given and don't assign (or for purposes of making welfare judgments distinguish between) how the benefits and costs of a particular action are distributed across society.

The individual measures are subject to two critical concerns: the appropriateness of the single measure chosen as a valid measure of an individual's well-being and the problems that one faces when attempting to quantify the components of the measure. The appropriateness of the aggregate measure depends on both the appropriateness of the individual measures and their aggregation.

Utility. Utility is another term for satisfaction, where a higher level of utility is more satisfaction. Of all possible combinations of goods and services, the one that maximizes individual utility is the one that is preferred. In the neoclassical welfare paradigm, the goal of individuals as economic agents is to maximize utility subject to a budget constraint.

Individual Preferences. Individual measures of well-being are premised on a fundamental economic assumption: that the satisfaction of individual preferences gives rise to individual well-being. Economists take this assumption as a matter of faith, and it underlies most if not all of economic theory. Others (many others) reject the assumption outright. At its base, the assumption is that individuals know what is good for them (what will enhance their well-being), their preferences for actions and outcomes reflect this knowledge, and they act in a manner consistent with these preferences in a desire to increase their well-being.[25]

If we accept the preference satisfaction assumption, we can look to people's actions as guides to their well-being. For example, if we see a person exchanging $3 for a six-pack of beer, we can state that the exchange made the person better off (increased the person's well-being) on the grounds that actions are motivated by a desire to satisfy preferences. But how much better off? The answer to that question brings us to the concept of economic value.

Economic Value. To economists, the term "value" has a specific meaning, so we use the term "economic value."[26] The most important, but often overlooked, features of economic value are that it is a theoretical construct and that monetary measures of it are inferred by analysts from the actions that people make in accordance with their preferences. Economic value cannot be independent of an action, in particular, a type of action that requires a person to make a choice whereby something is given up and something gained.

For microeconomists, the study of choice allows economic values to be defined and quantified.[27] Choice implies that a person is confronted with a selection of alternatives and that the consideration of the alternatives defines a trade-off. Contemporary economic theory of individual behavior, based on the assumption of preference satisfaction, suggests that when a person is confronted by choices, the alternative that is chosen must be at least as desirable, from the perspective of that person, as the alternatives that were not chosen. The value of the alternative chosen is thus defined in terms of the alternatives forgone. For example, if a person chooses to relinquish three apples to gain a peach, an analyst can state that under the circumstances of the choice (perhaps known in their entirety only to the person), the economic value of the peach to the person is at least three apples. If the choice were to give up $1 for the peach and the person chose the peach, the analyst would conclude that the value of the peach to that person was at least $1.[28]

Returning to the problems of measuring changes in individual well-being, suppose that a policy is being considered that would lower the price of peaches by 25% and have no other consequences. From the perspective of the person who is willing to pay at least a dollar for a peach, the policy enables him or her to pay only $0.75. The difference between the amount given up and the economic value of the peach is a monetary measure of the increase of the person's well-being—in this case $0.25.[29]

Willingness to Pay (WTP). An individual's willingness to pay reflects the trade-offs that he or she is willing to make between the consumption of a market or nonmarket good and wealth or income.

Value of a Statistical Life (VSL). The value of a statistical life is calculated by taking the willingness to pay for a mortality risk reduction of a given magnitude and dividing it by this risk reduction. For example, if the average WTP for a 1/10,000 reduction in mortality risk is $400, then the implied VSL is 400/(1/10,000) or $4 million. VSL estimates are often misinterpreted as representing a market value for human life. However, the metric is representative of choices and trade-offs individuals make (or say they would make) in the face of risk.

Costs. Although this paper is about describing the policy and technical properties of alternative effectiveness and benefits measures, some comments on the cost side are appropriate. The cost of a good or service is the value of the opportunities forgone in obtaining that good or service. More precisely, the cost of a regulation is equal to "the change in consumer and producer surpluses associated with the regulation and with any price and/or income changes that may result" (Cropper and Oates 1992).

This definition of cost is grounded in the welfare theory underlying neoclassical economics; it includes losses in satisfaction that do not result in monetary outlays as well as direct monetary outlays. The typical conception of cost limits itself to these latter, direct costs, such as the capital and operating expenditures associated with regulatory compliance, and may also include costs to the government in operating programs. There are other categories of costs—disrupted production, spillover effects to other industries, and the so-called "tax interaction" effect (Parry 1995)—that are plausibly important but are often excluded from cost analyses because of conceptual questions, a lack of understanding of the concept, or difficulties in obtaining credible data. The exclusion of these categories is often the basis for claims that the true costs of regulation are understated. The effect of technological change on costs is also often excluded. It is generally expected that technological change reduces costs of regulation, in which case omitting this factor from the cost analysis would bias costs upwards, other things being equal. In any event, while most controversies and uncertainties in CBA are thought to center on the benefits side, there are also significant uncertainties about costs.

Health States. Health states are conditions of health at a point in time that are generally (but not always) multidimensional (death being an exception).

Preference Weights (or Scores). Preference weights apply to health states, measuring how far they deviate from dead (0) and perfect health (1). Gold et al. (1996, p. 404) define this term as "a numerical judgment of the desirability of a particular outcome or situation."

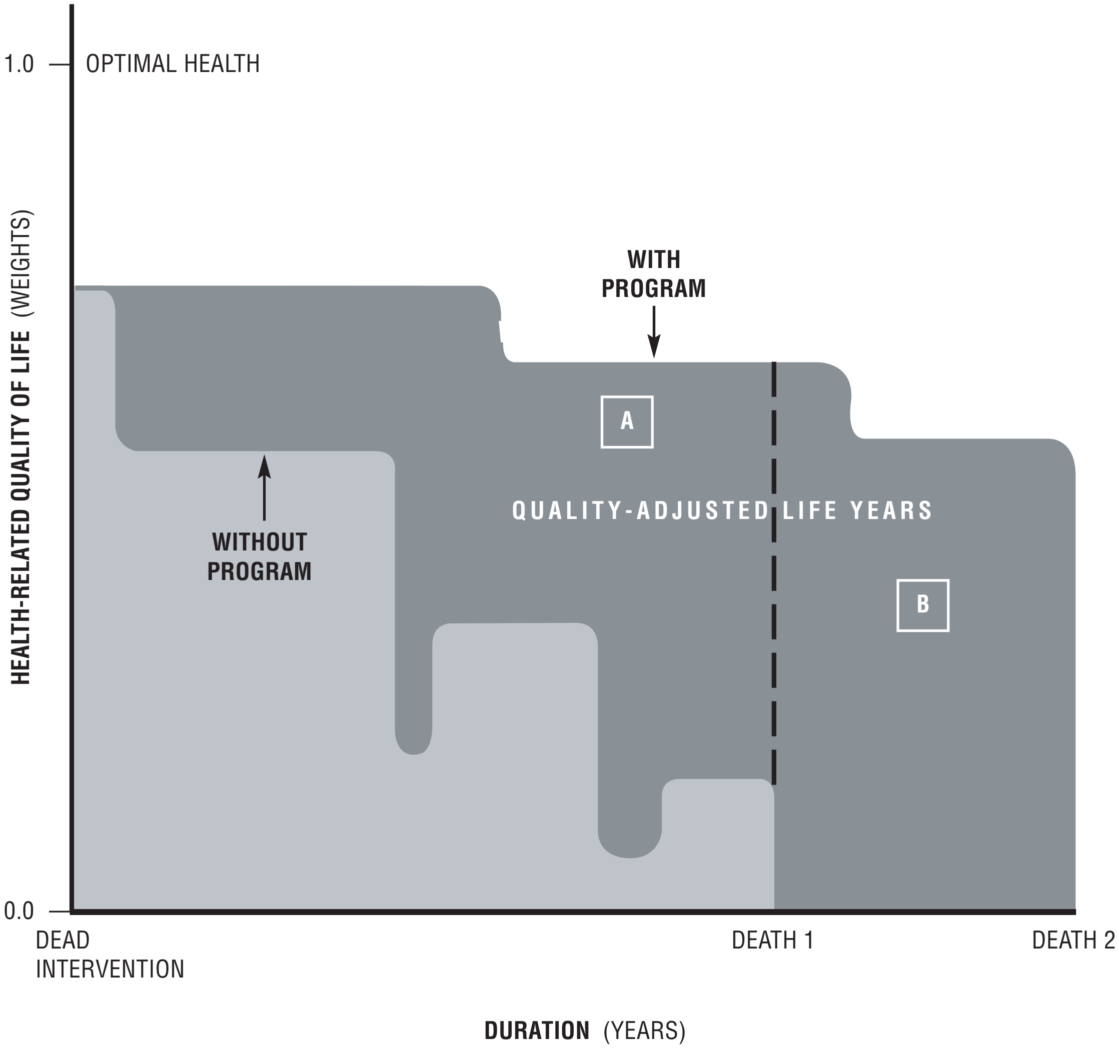
Domains. Domains are the components of a health state. As defined here, they could be aggregate, such as physical and emotional health, or they could be very detailed, describing functionality (ability to dress oneself, and so on).

Cost-utility analysis (CUA). Cost-effectiveness analyses that use QALYs as an effectiveness metric are often referred to as cost-utility analyses, because of the fact that many QALY indices purport to measure utility. We consider such analyses under the broader term of cost-effectiveness analysis in this paper, because QALYs can claim to measure utility only under very restrictive assumptions.

Quality-Adjusted Life Year (QALY). The QALY is a measure of health outcome that assigns a score ranging from 0 to 1 that characterizes quality of life (1 being perfect health, 0 being a health state equivalent to death) to a period of time. A QALY, then, is the product of the health score and some duration. It represents the number of years of healthy life that are equivalent to the time spent in the actual health outcome.

FIGURE A-1.

Estimating a QALY Loss or Gain



Estimating QALY loss or gain. The above figure depicts the estimation of QALYs gained for an individual under a policy intervention. In the absence of an intervention, the individual's health-related quality of life would deteriorate along the lower curve and the individual would die at Death 1. With the intervention, the individual would deteriorate at a slower rate, meaning live longer, and die at Death 2. The area A shows the QALYs without the intervention, and B shows added QALYs because of the intervention. As can be seen in the figure, there are gains in terms of both quality of life and quantity of life. It should be noted that the time of the intervention may not necessarily coincide with the divergence of the two curves. Where there is latency involved, the curves will remain identical beyond the point of intervention, diverging when the effects of the intervention are apparent.

APPENDIX II

Equity Paradigms

The paradigms include:

- Utilitarianism
- Rawls: "veil of ignorance"
- Sen: capability approach
- Prioritarian view

Utilitarianism is the idea that an equitable policy is one that best fits the aggregation of individual preferences. It does not make a priori distinctions between particular groups in society, but lets the aggregation of individual views on equity make up the social view on equity. Consumer sovereignty is consistent with utilitarianism.

The other paradigms present rationales for interposing a wedge between private and social views, based on the idea that individuals acting in their own interest may indicate one set of preferences, while if asked to provide preferences they think society as a whole should follow, they may express different preferences. For instance, 40-year-old respondents on WTP surveys appear not to value reducing risks to themselves any more highly than 70-year-olds value equivalent risk reductions. Should these preferences direct policy or should they be trumped by other survey results that may (or may not!) reveal that 70-year-olds would prefer to see society spend its scarce resources on saving younger people's lives? That is, should the distinction between social preferences and individual preferences be given credence by policymakers and do the social preferences trump the private ones, other things being equal? Or would having private preferences for social outcomes be enough?

The **veil of ignorance,** credited to Rawls, is the idea that the social preferences that matter are those that would be expressed if one did not know what state one would be in the future. Thus, although I might favor spending more resources on healthy people because I am healthy today, I might favor helping chronically ill people under a veil of ignorance if I were forced to give up my own view of the future and replace it with complete uncertainty.

The **capability approach,** credited to Sen, is the view that society should maximize people's capability to achieve their potential. Under this paradigm, well-being is comprised of "functionings" and "capabilities." Functionings are achievements, and capabilities are opportunities to achieve. Capabilities represent a person's potential well-being, or one's maximum potential health. Functionings are what individuals can achieve given their potential health. This paradigm would imply extending measures of health status to capture opportunity or the disparities between function and opportunity (Patrick remarks at RFF Conference 2003). One means of doing this might be to include people's perceptions about their health states on questionnaires. Under this approach, one would not want a valuation paradigm that would lead to assigning lower priority to improving the health of the infirm. The idea is that an individual's own view of his or her perfect health matters, more than a social standard or expert definition. (See Cutler and Richardson 1997, and others who have tried this approach.)

In the **prioritarian view,** those who have had least of what is to be distributed have priority. Under this paradigm, a year of life extension has a greater moral importance the younger its recipient. This view also gives preference to the sick and disabled. Interestingly, the way QALYs work in standard practice is not fully consistent with the prioritarian view. QALYs, by design, give greater weight to health improvements to the young because health improvements occur over a longer life expectancy.[30] However, they give less weight to health improvements to the infirm or disabled, who may have a lower life expectancy. WTP measures have no direct relationship with the prioritarian view because they are the sum of individual preferences. To the extent prioritarian views are embedded in such preferences and people are being asked their WTP for improvements in community health, the prioritarian view would be embedded in the measure.

APPENDIX III

Comments on OMB Draft Guidelines for the Conduct of Regulatory Analysis and the Format of Accounting Statements

Alan J. Krupnick
Resources for the Future
May 30, 2003

Acknowledgments

I would like to acknowledge the insightful and careful detailed comments on drafts of this paper from Kelly Maguire and Chris Dockins, EPA; Clark Nardinelli, FDA; Peter Belenky, DOT; and Carol Scotton, Scott Grosse and Phaedra Corso, CDC, all members of the Steering Committee on Valuing Health Outcomes, as well as Reed Johnson, RTI and David Feeny, Institute of Health Economics. Thanks also to Elise Golan, USDA; Ted Miller, Pacific Institute for Research and Evaluation; my colleague at RFF, Mike Taylor; Laura Blanciforti, NIOSH (also on the Steering Committee); and Dan Brock, NIH, for their very helpful comments. In spite of their help, the content of this report is solely the responsibility of the author.

Readers should note that the author is an environmental economist producing studies that estimate the willingness-to-pay for health improvements and mortality risk reductions as well as studies comparing the benefits and costs of environmental regulations. The comments and recommendations in this paper come out of this professional experience and context, as well as the intense debates of the last few months on the relative merits of various approaches to valuing health outcomes—an unsettled and relatively unresearched debate.

Introduction

Regulatory and non-regulatory activities by governments at the federal, state, and even local level affect the economic and physical health of their constituents. Analytical tools and concepts are available and are often used to evaluate and compare the effectiveness, efficiency, and distributional implications of these proposed or actual activities—namely cost-benefit analysis (CBA) and cost-effectiveness analysis (CEA).

Once the decision is made to use these tools, a further decision is needed about the measures to use for health outcomes. These measures include simple counts of cases (e.g., the number of deaths) prevented and indexes to aggregate over the variety of health effects that might accompany any particular course of action contemplated or taken by the government. Broadly speaking, there are two types of indexes: monetary measures (e.g., willingness-to-pay (WTP) and cost-of-illness (COI)[31]) and indexes over health states, referred to by the general term Quality-Adjusted Life-Years (QALYs).[32] The former are used in CBA, the latter (as well as physical measures) are used primarily in CEA.[33]

The choices of tools and health benefit measures are often controversial. Some statutes (such as the Clean Air Act) forbid use of cost-benefit analysis to make certain types of decisions (for example, on the stringency of the National Ambient Air Quality Standards), while others (such as the Safe Drinking Water Act) mandate its use. Irrespective of statutory requirements, federal agencies are required under Bush's Executive Order 13258—and Executive Orders from the Clinton to the Reagan Administrations—to evaluate the costs and benefits of major regulations, defined as those expected to have an effect on the economy of at least $100 million dollars annually. For agencies whose activities are not regulatory or whose regulations don't exceed the $100 million threshold, the choice of tools may be a matter of agency culture (although OMB has several other criteria for what actions require a CBA). While the U.S. Environmental Protection Agency (EPA) typically performs cost-benefit analyses, and may or may not supplement them with cost-effectiveness analyses, other agencies, such as the Occupational Safety and Health Administration (OSHA), generally perform cost-effectiveness analysis and do not monetize either morbidity or mortality benefits.

The Office of Management and Budget (OMB) is now attempting to change how government agencies do their analyses with its "OMB Draft Guidelines for the Conduct of Regulatory Analysis and the Format of Accounting Statements" (termed Draft Guidance here), which is Appendix C in its *Draft 2003 Report to Congress on the Costs and Benefits of Federal Regulations.* This Guidance calls for expanded use of cost-effectiveness analysis for regulatory analyses. The effectiveness measures are not spelled out, but it is thought by some observers that this is a call for more analyses based on cost per QALY or cost per life-year gained.

The primary purpose of this paper is to provide commentary to OMB on its Draft Guidance. Practitioners in government and stakeholders interested in agencies' Regulatory Impact Analysis (RIA) procedures may also find this document useful.

The responsibility for the ideas and commentary in this paper lies wholly with the author as a member of RFF. In large measure the intellectual underpinnings for this paper can be attributed to a conference and workshop held at the behest of an Interagency Steering Committee headed by the author and Michael Taylor, also of RFF. The Steering Committee members include: Kelly Maguire and Chris Dockins, EPA; William Lawrence, AHRQ; Elise Golan, USDA; James Schuttinga, NIH; Clark Nardinelli, FDA; Jim Schaub, ORACBA, USDA; Joseph Lipscomb, NIH; Debbie Aiken, OSHA; Peter Belenky, DOT; Greg Rodgers, CPSC; Laura Blanciforti, NIOSH and Carol Scotton and Phaedra Corso, CDC.

The conference on *Valuing Health Outcomes* was held February 13–14, 2003 at the RFF Conference Center in Washington, D.C., and involved some 25 speakers and 175 participants who contributed a great deal of information describing the conceptual and empirical foundations of the health valuation measures and their use in CBA and CEA. The follow-up workshop was held April 29, 2003, and involved a subset of this group in detailed discussions on a major paper written by the author, which analyzed CEA, CBA, WTP, QALY and COI measures in detail. At this workshop it was decided that a separate, more focused paper—this paper—would be written by the author to provide comments to OMB on its Draft Guidance.

The recommendations made in this paper are not necessarily consistent with views of any of the steering committee members or participants in the conference and workshop. Indeed, some members and participants disagree with some of these recommendations. Also, some topics in this paper lie at the fringes of topics discussed at the conference or workshop.

The remainder of the paper describes the provisions of the OMB Draft Guidance that are relevant to the debate over CEA and CBA analyses and appropriate health outcome measures and provides commentary and recommendations associated with these provisions.

Note that none of these particular OMB provisions speak to the decision context in which CBA or CEA are placed or to the use of these and other analyses to address the distribution of effects across populations. Both of these factors are vitally important in understanding how regulatory decisions should and do get made. Further, the measures of health value discussed below embed implications for equity effects that must be understood to choose among them. A paper discussing these and many related technical issues will be released soon, drawing on the conference and workshop presentations and discussions. However, these issues are largely absent from the paper below.

OMB Draft Guidance, Commentary and Recommendations

OMB's Draft Guidance proposes a number of changes to current procedures (or a reemphasis of current procedures) for the conduct of RIAs. There are seven provisions in the Draft Guidance that overlap the interest of our conference and workshop. With each is included a brief summary of the provision, a commentary, and recommendations.

The OMB provisions are:

1. Use CEA and CBA in the same regulatory impact analysis.

2. Give agencies discretion on the choice of effectiveness measures in CEA, but require justification.

3. Use a value of statistical life (VSL) that matches the case under study when monetizing mortality risk reductions.

4. WTP measures based on revealed preference approaches or passing scope tests (if based on stated preference approaches) are favored for their enhanced credibility.

5. Agencies must provide OMB with data to permit cross-rule calculations and comparisons.

6. Rules with costs over $1 billion must use Monte Carlo simulation approaches to represent uncertainties.

7. Draw on a literature that combines WTP and QALY estimates where gaps in WTP values exist.

1. Use CEA and CBA

The revised guidelines, with some exceptions, call for agencies to present both CEA and CBA in support of major proposed rules. The guidelines emphasize a call for CEA for all major rulemakings where the majority of benefits are improvements in health and safety. In addition, CBA should be conducted for major health and safety rulemakings "to the extent that valid monetary values can be assigned to the expected health and safety outcomes" (p.5516). CBA is justified because it "a) provides some indication of what the public is willing to pay for improvements in health and safety and b) offers additional information on preferences for health using a different research design than is used in CEA." (p. 5520). For regulations that are not focused on health and safety, CBA must be done, and a CEA should be performed when some of the benefits cannot be expressed in monetary units.[34]

Implementation of the proposed guidelines would have far-reaching effects. EPA rarely performs CEA, for instance, and the agency would need to start doing so. At the same time, some of the health agencies do not routinely perform CBA. Our reading of the OMB Draft Guidance is that it seeks to place both of these tools in a more equal position at all agencies whose regulatory decisions affect public health. Nevertheless, OMB is not expecting agencies to perform CBA where "valid monetary values" do not exist.

Political debates have been fierce about the use of CBA in government agencies and have resulted in suspicion that requirements for more analysis mask an intent for government paralysis. In this paper, these concerns are not addressed.

Ignoring the politics and focusing on informing decisions, it is hard to argue with the basic idea of providing additional information in a regulatory analysis, i.e., a CEA to supplement a CBA and a CBA to supplement a CEA. More information is generally desirable because it covers more perspectives. However, if the information provided is misleading or results in information overload to the point that decision makers fall back on heuristics, then more information may be undesirable. For instance, if these different analyses yield different policy prescriptions, decision makers will need to know in detail the meaning of the two analyses. If such meanings, which are complex and are based on many, often implicit, assumptions, cannot be clearly and succinctly conveyed, decision makers may simply ignore the least familiar analysis or "split the difference," or follow some other rule of thumb that may simplify or ignore vital pieces of information.

Several issues need clarification for this provision to be useful and workable, including: (i) the information provided by the two tools; (ii) the meaning of "valid monetary values," (iii) the practicality of developing an inventory of such measures; (iv) how to convey this added complexity to decision makers; and (v) the degree of harmonization implied by the draft guidance.

(I) INFORMATION PROVIDED BY THE TOOLS

CBA, with WTP-based values used to value health improvements,[35] is clearly based on social utility and, thus can in theory answer the question of whether a particular course of action has net benefits to society. CEA itself cannot do this.[36] Physical effectiveness measures, such as life years or the number of lives saved, provide no information regarding society's preferences. However, both measures can rank-order options, albeit with different indices. CBA rank-orders cardinally according to a money metric; CEA rank-orders with a cost-per-unit-of-effectiveness metric.

CBA has the ability to combine mortality and morbidity improvements into one (monetary) measure. That is, willingness-to-pay measures can capture all of the endpoints associated with a particular policy option—health and non-health. This is appealing because many (if not all) policies address multiple endpoints. In reality, willingness-to-pay measures for morbidity endpoints are often unavailable (as well as non-health endpoints).

CEA can also combine mortality and morbidity effects if a QALY metric is chosen, but this metric has not been used to aggregate non-health endpoints (although non-health benefits could be subtracted from costs and used in a net cost-per-QALY analysis). Morbidity values are more available in QALY terms than as WTP measures but the former may be defined too narrowly to match epidemiological endpoints. If CEA is used with a physical measure of effectiveness instead, the dominant measure of program output is chosen, for example, lives saved or life-years saved. In this case, other health endpoints are either ignored[37] or monetized and netted out of cost. It may be impossible, however, to find values for the monetization, and ignoring other effects may seriously bias the analysis.

Recommendations

Agencies should develop a common understanding of the advantages and disadvantages of these analytical tools and develop facility in using these tools.

(ii) "valid monetary values"

This term applies only to the use of CBA and is meant to apply to the benefits side of such an analysis, although it could apply to the cost side as well.

The term "valid" can apply in theory and empirically. Turning first to theory, WTP-based values would qualify as "valid monetary values," being based on the generally accepted paradigm of neoclassical welfare economics.

COI measures might not. Harrington and Portney (1987) and others show how COI measures fit into the neoclassical economics paradigm and conclude that they are generally considered a lower bound to WTP measures, because they leave out pain and suffering, for instance. Yet, in more complex and realistic models, this conclusion will not hold. Where individuals are insured, and do not see the marginal cost of their treatment decisions, medical costs may exceed willingness to pay, even without the COI measure capturing pain and suffering. COI could also overstate WTP because of technical inefficiencies in health care provision. The market for health care services is not competitive, meaning prices for health care services may not reflect social opportunity costs.

Another approach for possibly obtaining valid monetary values is to compute QALYs and convert them to a monetary metric using a "price per QALY" approach. This situation is problematic. First, the theory for such a conversion rests on the validity of using the revealed preferences of government or health care decision makers as a proxy for individual (public) preferences. As such decisions rest (rightly) on many factors beyond public preferences, there is no reason to think that this proxy is at all accurate. Second, CBA is meant to be used as an indicator of social welfare. If QALYs are not such an indicator (see the discussion under provision 2 below) then converting them to monetary terms does not make them any more of an indicator.

From an empirical perspective, it is helpful to begin from current state-of-the-art practice of CBA, say in the most recent EPA RIAs under the Clean Air Act (USEPA, 2003). There, WTP measures are used to value arguably the most important as well as the most numerous effects, i.e., mortality risk reductions, chronic respiratory disease risk reductions, and reductions in acute health effects. COI estimates are used to value effects not covered elsewhere, such as hospitalizations. Thus, the state-of-the-art incorporates a mix of WTP and COI approaches. Some analysts also use factors to convert COI to a WTP measure by computing the ratio of observed

WTP values to COI values for health endpoints that have these values available. WTP to COI ratios range from 2–3 for the endpoints generally used in air quality CBAs. CBAs are also performed, primarily for use in medical settings, by converting QALYs to monetary measures, generally using a single conversion factor. Such factors vary from $25,000 to $100,000 per QALY and higher. Sometimes, these factors are not actually termed "prices" or conversion factors, but are used as benchmarks to determine whether specific medical interventions are covered by insurance.

The use of factors to convert COI to WTP and to convert QALYs to monetary terms, while easy to do empirically, do not, in the author's judgment, pass the validity test. The COI conversion factor is an empirical observation over a narrow range of endpoints and might not hold up over a broader range. The QALY conversion factors are judgmental and arbitrary (as a measure of public preferences). Further, conferees agreed that there is no reason to think that a single conversion factor could be used to make such a conversion sensibly. How multiple conversion factors could be used or what they would be remains to be seen.

As for COI as a valid empirical measure, procedures for estimating these values have been refined over a long period and are routinely used in government analyses. There is reasonable consensus that COI will be lower than WTP for the types of endpoints examined for food safety and air quality improvements, particularly for the more serious illnesses. In this case the direction of the bias when using COI is known and regulations where benefits exceed costs can then be reliably described as delivering welfare improvements. But whether these relationships would hold more generally is an open question.

Empirically, the validity of both WTP and QALY measures remains highly contentious. Clearly, WTP measures are less often available for health endpoints of interest than QALY measures. A future paper will take up this issue in detail.

Recommendations

Members of the Steering Committee and other expert participants are particularly divided on the use of price per QALY factors in CBA. The author suggests that OMB discourage the use in CBA of both "price per QALY" factors and COI to WTP factors unless and until credible estimates are available.

Research is needed to develop WTP-based values for endpoints currently covered by COI estimates.

(iii) practical considerations

It will generally be easier for agencies doing CBA to do supplemental CEA analysis than for agencies doing CEA to do CBA, if the effectiveness measure is a physical one.

For agencies doing CBA already, CEA with a physical effectiveness measure is simply a subset of a regulatory analysis that estimates health effects and then monetizes them. If an agency is supplementing CBA with CEA expressed with a QALY effectiveness measure, the transformation of physical effects measures to QALYs may not be as straightforward, because there are many different health status indices to choose from and, as noted above, the degree to which one can map QALYs into specific epidemiological outcomes is unclear.

For agencies doing CEA, to do a CBA requires a means of converting the physical measures to monetary measures. In addressing morbidity effects, no WTP measures may be available. COI measures are generally available or can be estimated at fairly low cost.

With some agencies familiar with QALYs and in need of monetary measures, and other agencies familiar with monetary measures and in need of QALYs, there will be a big increase in demand for procedures and values (WTP, COI, and QALY weights) for use in such analyses. At the same time there is skepticism that the supply of appropriate and credible values is close to being adequate. For any of these measures it can be difficult to judge their generalizability. For instance, QALYs are not always based on a representative sample of the population with a particular problem.

Recommendations

It may be useful for OMB (or perhaps a special Review Board) to examine the supply of values by establishing data repositories or by endorsing integrated assessment/benefit models that serve to provide repositories and procedures for using the data. To a certain extent, existing OMB (and other agency) guidance and standard operating procedures do just this.

Models are reasonably well established for estimating WTP for the health benefits of air quality improvements and for estimating the cost-of-illness associated with foodborne disease. Possibly the best example of the former is the recently released BENMAP model developed by OAQPS, EPA. A latter example is the recently released Foodborne Illness Cost Calculator from the Economic Research Service, USDA (http://www.ers.usda.gov/data/foodborneillness/). So-called league tables of QALY weights and scores associated with various health states and medical interventions are widely available (see http://www.hcra.harvard.edu/pdf/ preferencescores.pdf), although QALYs (and cost/QALYs) associated with medical interventions are far more numerous than those for particular health states, the latter probably being more useful for national policy initiatives.

Use of such tables and models should not be in any way mandated by OMB, but rather offered by OMB as a service to agencies doing RIAs and as an attempt to encourage harmonization of methods across agencies and to give RIAs more weight with decision makers. The Environmental Valuation Reference Inventory (EVRI, http://www.evri.ca/) developed by Environment Canada (with USEPA support) is a good example of such a database. Further, agencies could add to this database in some controlled manner, perhaps through peer review, so that their work would be available to and benefit all. This approach should avoid freezing the state-of-the-art and would recognize the significant heterogeneity of contexts in which agencies write rules.

(iv) reporting

Because of the need to explain and report the results of the application of several different benefit measures, the burden on the agencies to report clearly on their methods and assumptions is great. Yet, overhaul of agency reporting protocols is needed for many reasons. First, the agencies are still not reporting uncertainties clearly[38] and will have an even larger burden once they start performing Monte Carlo simulations (see below). Second, they are not reporting enough detail for even professionals to understand how certain key estimates were derived. For instance, the discussion of estimating benefits of the EPA's very recent Off-Road Diesel rule (USEPA,

2003) omitted mention of the assumption that people benefiting from the rule would have otherwise had 6 months to live if they had chronic respiratory disease and five years to live if they died of other causes. Third, agencies are already using base case and alternate analyses with sensitivity analyses layered on top of these in appendices. Now, with multiple benefit measures and entirely separate analyses being used, these protocols will need to be revised.

Recommendations

OMB should encourage studies of decision makers at agencies concerning how they use complex information and how such information could be better reported to make its use easier.

Uniformity in reporting formats in RIAs should be encouraged. As in provision 7, below, this uniformity should include how uncertainties should be addressed (NRC, 2002).

(v) harmonization of methods

Historically, OMB's guidance documents enforce a relatively loose regime for the conduct of CBA, as evidenced by the wide variety of choices made by the agencies in their choice of specific measures for such important elements as the value of statistical life. The same may be true of agencies using CEA.

Recommendations

For several reasons a tighter regime is recommended, where the guidance is written with more specificity about the options the agencies have and what the default is to do both types of analyses. When the guidance is not followed, agencies would have a high burden of proof for showing why they were not followed.

Why this tighter regime? First, agency cultures are entrenched when it comes to use of CBA or CEA. At EPA, for instance, standard procedures are for using CBA and there is likely to be a costly learning curve for using QALYs. At some of the public health agencies the learning curve for monetization techniques is likewise high; more importantly, at these agencies there is a traditional reluctance to monetize reduced health effects. These agencies may not easily buck this culture and need a strong push from OMB to make such a change happen. Second, OMB wants the ability to conduct its own analyses of agency analyses and to compare findings across rules both within and across agencies. The only way to make this happen—and the author endorses the idea of cross agency comparisons (see provision 5)—is for the agencies to follow procedures in greater faithfulness and in greater detail than previously.

A corollary to this recommendation is that OMB consider whether this threshold is low enough to result in pressure to change agency culture (granting that performing such analyses on very minor regulations would in itself waste resources and would play into the suspicion of paralysis by analysis raised above. One option would be to drop all thresholds, but state that the degree of sophistication of the analysis match what is at stake with the rule. Defining these terms would be a challenge, of course.

These recommendations are offered with two caveats. First, because legislative mandates and requirements as well as rulemaking contexts may differ markedly across regulatory efforts both within and across agencies, such agencies should have the option of doing analyses their preferred way as a supplement to meeting the OMB guidelines and to do so without need of justification.

Second, because learning to do analyses in new ways takes time and money, and because building the inventory of values would take time as well, these requirements should be phased in over several years and special budgetary allotments should be made for such training.

2. Appropriate Effectiveness Metric and Agency Justification

The OMB Guidance draft does not endorse any particular effectiveness metric for CEA. These metrics can be in physical terms, e.g., lives saved or life-years-saved, or in terms of QALYs. If the latter, they may be drawn from particular health status indices, such as the Quality of Well Being Scale (QWB) or the Health Utility Index (HUI). The indices can be based on four preference survey approaches: standard gamble (SG), rating scales (RS), time tradeoff (TTO), or person tradeoff (PTO). Direct elicitation of preferences using these survey approaches can also occur.

(i) physical effectiveness measures

The technical requirements associated with presenting physical effectiveness measures in a CEA are far less than those with QALYs. Generally, epidemiological studies present outcomes in terms of lives saved, cases reduced, days of illness reduced, among others. These measures, when divided into costs, provide a CEA ratio directly and simply.

There are several issues associated with physical effectiveness measures. First, as noted above, only one effectiveness measure can be used at a time. For example, lives saved cannot also be combined with reduced hospital visits. Other physical effects must then be either monetized and subtracted from costs or ignored. If key effects are ignored in a CEA, there is obviously the potential for the analysis to be misleading. Without direct use of preference-based information to determine which effects are "key" and which can be safely ignored, the problems for doing a useful CEA analysis are magnified.

Nevertheless, where key output measures have been identified a priori, whether or not based on preference-based information, use of CEA with physical effectiveness measures could yield important and useful information. EPA's RIA for reducing fine particulate concentrations, for instance, could usefully be enhanced with cost-effectiveness analysis based on cost per life saved (or life-year-saved, see below), because it is widely accepted that the mortality effects dominate these analyses (based on WTP studies over the last decade).

Second, the controversy between using lives saved versus life-years saved as the physical effects measure is an issue. A life-years saved measure presumes that saving younger people's lives is "worth" more than saving older people's lives, or that saving a year of life is equivalent irrespective of one's age. Using the lives saved measure, on the other hand, presumes that age doesn't matter.

Third, really a generalization of the second point above, the quality that makes physical measures of effectiveness in CEA attractive—their simplicity—is also their weakness. People may wrongly, but understandably attach normative significance to cost-effectiveness rankings based on them, not realizing that preferences for health outcomes may depend on far more than their number.

(ii) QALYs

Where a key effect hasn't been established or monetization of key effects is either impossible or has been otherwise ruled out for a net cost-effectiveness analysis, agencies are likely to use QALYs. At the conference, and, in particular, at the following workshop, significant time was spent examining the validity, reliability and practicality of using alternative QALY indices.

At these meetings, QALYs were often termed a measure that can be "utility-based" (if based on standard gamble or time-tradeoff approaches to deriving weights) while not being a representation of utility. If this were the accepted interpretation of QALYs, then many of the conceptual objections raised about this measure (e.g., its empirical violation of necessary assumptions to be a measure of welfare) would disappear. In this case, QALYs may be regarded as a convenient way of aggregating a variety of physical health effects for the purposes of ordering regulatory options in a CEA. This CEA measure would be little different in meaning than a cost per life saved or other physical effectiveness measure, except it would account for morbidity effects as well. As such, this measure would not directly capture preferences related to risk "qualities," such as stigma, fear, controllability and the like.

But empirical issues remain. For instance, QALYs are often derived from small, specialized samples. They also are often based on survey methods that have remained largely unchanged for many years and have not been subjected to rigorous validity tests, raising the question of their robustness and unbiasedness. It was generally agreed at the workshop that indices based on SG and TTO surveys more closely mirrored the policy-relevant choice situations faced by individuals and better-measured preferences than other weighting approaches. At the same time, it was acknowledged that even these approaches had significant issues. Both of these approaches were criticized as placing individuals in unrealistic choice situations, particularly with respect to SG questions for measuring preferences to reduce acute morbidity. However, participants in the workshop cited at least one study that has captured such effects well (Feeny et al. 2002).

It is worth noting here that WTP measures face many empirical issues as well. See provision 4 below.

Recommendations

OMB should be silent on the agency choice of effectiveness measure, but laying out the tradeoffs between using physical measures vs. QALYs. OMB should note that the physical measures are very simple but should have no normative significance attached to them.

OMB should require agencies to provide an explanation for their choice of effectiveness measure. This requirement would help OMB and other agencies and stakeholders understand the agency's rationale for this important set of choices. It could also help make consistency across agencies easier to achieve.

At the workshop, there was discussion about the idea of agencies electing to use QALYs being asked to "unpack" their estimates to increase the transparency of this measure. For instance, rather than presenting one "score"[39] for the change in QALYs, the analysis would present a score for the life-years-saved and a score for the change in the morbidity portion of the QALYs. The sum total of these two categories would be the total QALYs gained from the option under consideration.[40]

A way to unpack the morbidity portion of QALYs is to separate duration from the morbidity preference weight. The weights come from the original SG, PTO, TTO or RS surveys, while duration information may have many origins and be quite uncertain. Unpacking estimates in this manner can help focus attention on these two components.[41] At the workshop this unpacking idea lacked consensus.

3. VSL Heterogeneity

OMB does not suggest a particular VSL for CBA. Rather, OMB suggests that "You should not use a VSL estimate without considering whether it is appropriate for the size and type of risks addressed by your rule" (p. 5521).

OMB is asking that VSLs be applied where they are appropriate to the context being considered. For instance, OMB does not want VSLs derived for fatal heart attacks to be applied to reductions in cancer mortality risk. The guidance document notes that if a VSL is being applied in a different context from which it was derived, appropriate adjustments should be made, and the differences in contexts should be outlined. This guidance, although not altogether new, is in marked contrast to procedures generally followed in RIAs, where one VSL is chosen, albeit after much analysis of the existing literature.

While OMB is aware of the newer VSL literature suggesting that VSLs vary by risk size and type as well as population characteristics, use of varying VSLs is highly controversial. Administrator Whitman has vowed that a "senior discount" will never be used at EPA to make decisions (National Public Radio, May 21, 2003).

Recommendations

The practical issue about encouraging VSL heterogeneity is that such context-specific values are often unavailable. In this case, analysts typically use benefit transfer and meta-analysis techniques. Thus, the guidance needs to reflect this practical reality but could require agencies to include a justification.

OMB should support additional research in valuing mortality risk reductions in contexts where estimates are currently unavailable. It is no secret that getting survey work done within the federal government or with federal funding is difficult at best because of the holdups often experienced in the Information Collection Request approval process. Addressing this significant hurdle will result in great strides in our abilities to develop WTP estimates using survey techniques.

It is worth noting that this recommendation applies just as well to any survey work, including surveillance surveys that can be used to obtain data on prevalence, incidence, duration, and severity, which are needed to estimate QALYs as well as WTP measures of morbidity. It would also apply to surveys used to develop QALYs themselves.

How alternative life-saving programs (and health-improving programs, in general) are to be ranked is an important public policy question that needs open debate. OMB should develop or support the development of a forum for such a debate.

4. Enhancing the Credibility of Measures

OMB emphasizes, as it has done in previous guidelines, that the credibility of WTP measures is enhanced when they are derived from revealed preference (RP) studies. The guidelines note, however, that RP data are often not available for health and safety risks or are derived from studies far from the appropriate regulatory context. In this case, values from stated preference (SP) studies can be used, with those passing external scope tests judged to be more reliable than values from other studies.

These statements regarding WTP measures should be revisited to place stated preference methods on an equal, if not superior, footing with RP measures as SP practices improve in quality or if research shows that the inability of RP measures to match the appropriate regulatory context is a significant impediment to their applicability.

Further, these comments regarding WTP measures should apply equally to QALY measures, with OMB setting out validity tests on such measures that would enhance their credibility. Preference weights underlying QALY calculations need to be subjected to more validity tests and guidelines, just like WTP studies.

The OMB Guidance is silent on the question of whether QALYs or WTP measures are preferable for use in RIAs. Implicitly, WTP is endorsed because CBA is endorsed and QALYs are not ruled out as an effectiveness measure in CEA.

The participants in the conference and workshop agreed that both measures have advantages and disadvantages, implying that one measure cannot be unambiguously preferred to the other. Participants also felt that an important contribution of further efforts would be to determine when a WTP measure would be useful and when a QALY measure would be useful to a policy analysis.

To give some flavor for the issues discussed in the workshop and conference, a paper summarizing what was learned is in the preparation stage. In this paper, the following points are made. Both measures provide an index useful for aggregating diverse health effects, with one index based on a monetary unit and the other based on a life-year. Both measures reflect something about preferences across health effects, with some approaches to eliciting those preferences being very simple or even simplistic and others being very complex, sophisticated and realistic. Both measures carry significant "baggage," such as the controversy over monetizing "life," and the recent controversy over the "senior discount" that would also follow from use of a QALY measure. Both sets of measures have problems demonstrating credibility and both have problems of data availability.

Recommendations

OMB should rewrite the Guidance to demand equal burden of credibility/validity for WTP and effectiveness measures such as QALYs and should revisit its bias towards revealed preference approaches to measuring WTP.

5. Data provision to OMB

The guidelines require agencies to provide OMB with all relevant data, including morbidity and mortality, population data, and the severity and duration of the conditions evaluated, so that OMB can make comparisons across rulemakings that employ different measures.

CEA comparisons may be problematic unless the effectiveness measures are standardized. Requiring that each agency use lives saved may not be enough for physical measure comparability if it is deemed that life-years saved is important. Neither is sufficient where morbidity is important (unless the morbidity effects can be monetized and subtracted from costs). Requiring the agency to use QALYs as the effectiveness measure is not enough to assure comparisons are possible because there are so many competing indices.

Recommendations

Requiring agencies to provide underlying data and models in sufficient detail that OMB can do the calculations according to its own dictates is likely to be challenging because of the many specialized data and specific decisions needed to estimate QALYs. On the other hand, to the extent life-years dominate the QALY calculations, QALY scores across alternative indices may not be much different. Therefore, more study is needed about whether OMB's recommendation is practical.

6. Monte Carlo Analysis

The guidelines note that relevant outcomes should be reported as probability distributions where possible, and that for rules with an economic impact greater than $1 billion, a formal quantitative analysis of uncertainty is required. For rules below the $1 billion threshold and without sufficient information to report probability distributions, sensitivity analysis can be used to assess uncertainty. For rules above the threshold, however, probabilistic analysis should be applied throughout the entire range of calculations, such that uncertainty is carried through to the endpoints.

Improving and standardizing the characterization of uncertainty in RIAs is of major importance, given the significant uncertainties associated with all parts of CEAs and CBAs in support of major rules. Use of Monte Carlo analysis is becoming more routine in the literature and among practitioners of these analyses, partly because the computer tools are becoming more widely available and easier to use.

At the same time, the new guidance is silent on techniques for reporting uncertainties. One suggestion is for OMB to mandate incorporation of tables listing unquantifiable uncertainties, e.g., model uncertainties as opposed to statistical uncertainties, and the direction of possible biases.

A more fundamental issue involving uncertainty and reporting concerns the procedures for defining and presenting main and alternative results and sensitivity analyses in the RIA. The recent Off-Road Diesel Engine RIA issued by EPA (2003) is a case in point. In this report, EPA presented a Base case and an Alternate case, plus sensitivity analyses in an appendix. The Base case incorporated what EPA considered their best guess at appropriate model selection (e.g., epidemiological studies) and parameter choices. The Alternate case incorporated models and parameters that were towards the lower range of plausible values. The resulting benefit estimates from the Base and Alternate cases were not offered with uncertainty distributions. Furthermore, although a reasonable case could be made for the validity of models and parameter values towards the upper range, benefits based on these choices were relegated to sensitivity analyses in the appendix. Thus, all but the most dogged reader would believe that the mass of uncertainty lay below the base case benefit estimate (which it very well may, but if so, this is not clear).

The use of Monte Carlo simulation techniques would improve upon these procedures but not eliminate them.

Beyond these "macro-uncertainty" issues are some "micro" issues. Because of the significant uncertainties in WTP, QALY and physical effects (as well as cost) estimates, it is essential to quantify such uncertainties. This is standard procedure with WTP estimates in cost-benefit analysis and is also generally followed in the better cost-effectiveness analyses. For instance, duration and the preference weights in QALYs are encoded as distributions in the better analyses.

Finally, an issue was raised in the workshop about the appropriate treatment of uncertainties from multiple links in a chain of causation. It may be that the uncertainties in one variable in the calculation are correlated with uncertainties in another variable. Not accounting for this dependence, but rather assuming that the uncertainties are independent, would result in wider confidence intervals (if the correlations are positive).

Recommendations

OMB guidance on the appropriate development of scenarios to fairly represent statistical and model uncertainties, given that Monte Carlo simulation will be used, would be a useful addition to the draft guidance document.

An explication of why a cost threshold of $1 billion would be used for probabilistic CBA/CEA would be useful, as this figure is arbitrary. Why not a lower figure, say $100 million? Because of uncertainties in the WTP and QALY measures and historic problems within the agencies in fairly and completely representing uncertainties, a lower threshold would assure that more analyses include this important added information.

As Monte Carlo analysis tends to be performed with the assumption of independence in uncertainty across variables in a causal chain, it would be useful for OMB to flag this issue and suggest that attention should be paid to modeling correlated uncertainties where there is good reason to think they exist.

7. Use literature combining WTP and QALYs

OMB offers a suggestion to agencies faced with gaps in their WTP information. The Guidance says: "If data are not available to support monetization, you might consider an alternative approach that makes use of health-utility studies . . . This health utility [QALY] information may be combined with known monetary values for well-defined health states to estimate monetary values for a wide range of health states of different severity and duration. If you use this approach, you should acknowledge your assumptions and the limitations of your estimates" (p. 5521).

With this comment OMB advances the idea of using both measures together, of combining WTP and QALY measures (“QALYfying the WTP”). Because there is as yet a small literature in this area (e.g., Johnson et al. 1997; Van Houtven et al. 2003), this suggestion serves to underline the need for more research, rather than propose something the agencies could immediately operationalize in constructing their regulatory impact analyses. Indeed, researchers at the conference and workshop talked much more about mounting different research efforts.[42]

Recommendations

OMB guidance should acknowledge that combining WTP and QALYs is not something that can be operationalized without additional research and should not limit or favor ways in which WTP and QALY concepts and measures can be combined.

References for Appendix III

Feeny, David, Marie Townsend, William Furlong, Darrell J. Tomkins, Gail Erlick Robinson, George W. Torrance, Patrick T. Mohide, and Qinan Wang, 2002. “Health-Related Quality of Life Assessment of Prenatal Diagnosis: Chorionic Villus Sampling and Amniocentesis,” *Genetic Testing* 6 (1): 39–46.

Harrington, Winston, and Paul Portney. 1987. “Valuing the Benefits of Health and Safety Regulations,” *Journal of Urban Economics* 22 (1): 101–112.

Johnson, F. Reed, M.R. Banzhaf, and W.H. Devouges et al. 2000. “Willingness to Pay for Improved Respiratory and Cardiovascular Health: A Multiple Format, Stated Preference Approach.” *Health Economics* 9: 295–317.

Johnson, F.R., E.E. Fries, and H.S. Banzhaf. 1997. “Valuing Morbidity: An Integration of the Willingness-To-Pay and Health-Status Index Literatures.” *Journal of Health Economics* 16(97): 641–665.

Johnson, F.R., J. Mauskopf, and A.B. Hauber. 2002. “Measuring Health-State Utility: Rulers, Thermometers, or Shoe Sizes?” Working Paper. RTI Health Solutions.

National Public Radio. 2003. “EPA Administrator Whitman to Step Down.” *All Things Considered*, May 21, 2003.

National Research Council (NRC). 2002. *Estimating the Public Health Benefits of Proposed Air Pollution Regulations*. Washington, DC.

Office of Management and Budget (OMB). 2003. Draft Report to Congress on the Costs and Benefits of Federal Regulations. Available online: www.whitehouse.gov/omb/fedreg/2003draft_cost-benefit_rpt.pdf

Smith, Kerry, George Van Houtven, and Subhredu Pattanayak. 2003. *Preference Calibration with QALYs*. Working Paper.

U.S. Environmental Protection Agency (USEPA) 2003. “Chapter 9: Cost-Benefit Analysis”. *Off-Road Diesel Engine RIA* (draft). Available online: www.epa.gov/nonroad/r03008j.pdf

Van Houtven, George et al. 2003. *Valuation of Morbidity Losses: A Meta-Analysis of Willingness-to-Pay and Health Status Measures*. Final Report. Report prepared for Food and Drug Administration Center for Food Safety and Applied Nutrition. March.

■ ■ ■

Bibliography

Aimola, A. 1998. Individual WTPs for reductions in cancer risk deaths. In *Environmental Resource Valuation: Applications of the Contingent Valuation Model in Italy*, edited by R. Bishop and D. Romano. Dordrecht: Kluwer, 196–212.

Alberini, A., M. Cropper, A. Krupnick, and N.B. Simon. Forthcoming. Does the Value of a Statistical Life Vary with Age and Health Status? Evidence from the United States and Canada. *Journal of Environmental Economics and Management.*

Alberini, A., and A. Krupnick. 2003. Valuing the Health Effects of Pollution. In *The International Yearbook of Environmental and Resource Economics 2002/2003*, edited by T. Tietenberg and H. Folmer. Cheltenham: Edward Elgar.

Anderson, J.P. et al. 1989. Interday Reliability of Function Assessment for a Health Status Measure: The Quality of Well-Being Scale. *Medical Care* 27: 1076–1084.

Arrow, K. 1951. *Social Choice and Individual Values.* New York, NY: Wiley & Sons.

Arrow, K., R. Solow, P.R. Portney, E.E. Leamer, R. Radner, and H. Schuman. 1993. Report of the NOAA Panel on Contingent Valuation. *Federal Register* 58(10): 4601–4614.

Bala, M.V., L.L. Wood, G.A. Zarkin, E.C. Norton, A. Gafni, B.J. O'Brien, 1999. Are health states "timeless"? the case of the standard gamble method. *Journal of Clinical Epidemiology*, 52(11): 1047–1053.

Bishop, R.C., and T.A. Heberlein. 1979. Measuring Values of Extra Market Goods: Are Indirect Measures Biased? *American Journal of Agricultural Economics* 61(5): 926–930.

Bishop, R.C., and D.W. McCollum. 1997. Assessing the Content Validity of Contingent Valuation Studies. Unpublished manuscript. University of Wisconsin, Madison.

Bleichrodt, H., and M. Johannesson. 1997. The Validity of QALYs: An Experimental Test of Constant Proportional Trade-Off and Utility Independence. *Medical Decisionmaking* 17: 21–32.

Blomqvist, A. 2002. QALYs, Standard Gambles, and the Expected Budget Constraint. *Journal of Health Economics* 21: 181–195.

Bloyd, C. et al. 1996. *Tracking and Analysis Framework (TAF) Model Documentation and Users Guide*. Argonne National Laboratory, ANL/DIS/TM-36, December.

Briggs, A.H., B.J. O'Brien, and G. Blackhouse. 2002. Thinking Outside the Box: Recent Advances in the Analysis and Presentation of Uncertainty in Cost-Effectiveness Studies. *Annual Review of Public Health* 23: 377–401.

Brock, D.W. 1998. Ethical Issues in the Development of Summary Measures of Population Health Status. In *Summarizing Population Health: Directions for the Development and Application of Population Metrics*, edited by M.J. Gold and M.R. Gold. Washington, D.C.: National Academy Press, 73–85.

Brouwer, W.B.F., and M.A. Koopmanschap. 2000. On the Economic Foundations of CEA. Ladies and Gentlemen, Take Your Positions! *Journal of Health Economics* 19: 439–459.

Buzby, J., T. Roberts, C.T. Jordan Lin, and J. MacDonald. 1996. Bacterial Foodborne Disease: Medical Costs and Productivity Losses. *Agricultural Economics Report No. 741*. August.

Carson, R.T. et al. 1996. Contingent Valuation and Revealed Preference Methodologies: Comparing the Estimates for Quasi-Public Goods. *Land Economics* 72: 80–99.

Carson R.T., N.E. Flores, and N.F. Meade. 2001. Contingent Valuation: Recent Controversies and Evidence. *Journal of Environmental and Resource Economics* 19: 173–210.

Carter, W.B., L. Beach, T.S. Inui, J.P. Kirscht, and J.C. Prodzinski. 1986. Developing and Testing a Decision Model for Predicting Influenza Vaccination Compliance. *Health Services Research* 20(6): 897–932.

Champ, P.A., K.J. Boyle, and T.C. Brown (eds.). 2003. *A Primer on Nonmarket Valuation*. Dordrecht: Kluwer.

Cohen, M.A., and T.R. Miller. 2003. "Willingness to Award" Nonmonetary Damages and the Implied Value of Life from Jury Awards. *International Review of Law and Economics* 23(2): 165–181.

Cookson, R. 2000. "Incorporating Psycho-Social Considerations into Health Valuation: An Experimental study," *Journal of Health Economics* 19: 369–401.

Cousineau, J-M. et al. 1992. Occupational Hazard and Wage Compensating Differentials. *Review of Economics and Statistics* 74: 166–169.

Cropper, M., and W.E. Oates. 1992. Environmental Economics: A Survey. *Journal of Economic Literature* 30(2): 675–740.

Culyer A.J., and A. Wagstaff. 1993. QALYs versus HYEs. *Journal of Health Economics* 12(3): 311–23.

Cummings, R., D. Brookshire, and W. Schulze. 1986. *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*. Totowa, NJ: Rowman and Allanheld.

Cutler, D., and E. Richardson. 1997. Measuring the Health of the United States Population. *Brookings Papers on Economic Activity, Microeconomics*, 217–272.

DeShazo J.R., and T.A. Cameron. 2003. Draft: *An Empirical Life-Cycle Model of Demand for Mortality and Morbidity Risk Reduction.*

Desvousges, W.H., F.R. Johnson, and H.S. Banzhaf. 1998. *Environmental Policy Analysis with Limited Information.* Cheltenham: Edward Elgar.

Dolan, P., and R. Cookson. 2000. A Qualitative Study of the Extent to Which Health Gain Matters When Choosing between Groups of Patients. *Health Policy* 51: 19–30.

Donaldson C., S. Birch, and A. Gafni. 2002. The distribution problem in economic evaluation: income and the evaluation of costs and consequences of health care programs. *Health Economics* 11: 55–70.

Drummond, M.F., B.J. O'Brien, G.L Stoddart, and G.W. Torrance. 1997. *Methods for the Economic Evaluation of Health Care Programmes.* Oxford: Oxford University Press.

Economic Research Service. 2003. Foodborne Illness Cost Calculator. www.ers.usda.gov/data/foodborne illness. (accessed on Dec. 23, 2003).

Epstein, A.M. et al. 1989. Using Proxies to Evaluate Quality of Life. *Medical Care* 27: S91–98.

Erickson P., R. Wilson, and I. Shannon. 1995. Years of Healthy Life. *Healthy People 2000: Statistical Notes*. National Center for Health Statistics, U.S. Department of Health and Human Services. No.7: 1–9

The EuroQol Group. 1990. EuroQol—A New Facility for the Measurement of Health-Related Quality of Life. *Health Policy* 16: 199–208.

European Commission-DGXI. 1998. *Economic Evaluation of Air Quality Targets for Sulphur Dioxide, Nitrogen Dioxide, Fine and Suspended Particulates and Lead.* Final Report. Luxembourg: Office of Official Publications of the European Communities.

ExternE. 1999. *Externalities of Energy, Volume 7: Methodology 1998.* Update. Brussels: European Commission.

Farrow, R. S. et al. 2001. Facilitating Regulatory Design and Stakeholder Participation: The FERET Template with an Application to the Clean Air Act. In *Improving Regulation: Cases in Environment, Health, and Safety*, edited by Paul S. Fishbeck and R. Scott Farrow. Washington, DC: Resources for the Future.

Feeny, D., W. Furlong, G.W. Torrance, C. H. Goldsmith, Z. Zhu, S. DePauw, M. Denton, and M. Boyle. 2002. Multiattribute and Single-Attribute Utility Functions for the Health Utilities Index Mark 3 System. *Medical Care*. 40(2): 113–128.

Freeman, A.M., III. 2003. *The Measurement of Environmental and Natural Resource Values*. Washington, DC: Resources for the Future.

Freeman, A.M., III, J.K. Hammitt, and P. DeCivita. 2002. On Quality Adjusted Life Years (QALYs) and Environmental/Consumer Safety Valuation. *AERE Newsletter* 22(1): 7–11.

Fuchs, V.R., and R.J. Zeckhauser. 1987. Valuing Health—A "Priceless" Commodity. *American Economic Review*, *Papers and Proceedings* 77(2): 263–68.

Furlong, W.J. et al. 1990. *Guide to Design and Development of Health-State Utility Instrumentation.* Centre for Health Economics and Policy Analysis Paper 90–9. June.

Furlong W. J., D. H. Feeny, G. W. Torrance, and R. D. Barr. 2001. The Health Utilities Index (HUI) System for Assessing Health-Related Quality of Life in Clinical Studies. *Annals of Medicine* 33(5): 375–384.

Gafni, A., S. Birch, and A. Mehrez. 1993. Economics, Health and Health Economics: HYEs versus QALYs. *Journal of Health Economics* 11:325–339.

Garber, A.M. et al. 1996. Theoretical Foundations of Cost-Effectiveness Analysis. In *Cost-Effectiveness in Health and Medicine*, edited by M.R. Gold et al. Oxford: Oxford University Press.

Gold, M.R., J.E. Siegel, L.B. Russell, and M.C. Weinstein. 1996. *Cost-Effectiveness in Health and Medicine*. Oxford: Oxford University Press.

Gyrd-Hansen, D. 2003. Willingness to Pay* for a QALY. *Health Economics*. 12: 1049–60.

Haddix, A.C., S.M. Teutsch, and P.S. Corso. 2003. *Prevention Effectiveness: A Guide to Decision Analysis and Economic Evaluation.* Second edition. Oxford: Oxford University Press.

Hamermesh, D.S. 1999. Changing Inequality in Markets for Workplace Amenities. *Quarterly Journal of Economics* 114: 1085–1123.

Hammitt, J.K. 2002. QALYs Versus WTP. *Risk Analysis* 22(5): 985–1001.

______. 2003. Valuing Health: Quality-Adjusted Life Years or Willingness to Pay? *Risk in Perspective* 11(1): 1–6.

Harbaugh, W.T., K. Krause, and L. Vestelund. 2002. Risk Attitudes of Children and Adults: Choices over Small and Large Probability Gains and Losses. *Experimental Economics* 5: 53–84.

Hargreaves, D.J., and G. Davies. 1996. The Development of Risk-Taking in Children. *Current Psychology* 15(1): 14–16.

Harrington W., and P. Portney. 1987. Valuing the Benefits of Health and Safety Regulation. *Journal of Urban Economics* 22:101–112.

Heckerling, P.S., M.S. Verp, and N. Albert. 1997. Prenatal Testing for Limb Reduction Defects. How Patients' Views Affect their Choice of CVERSUS. *Journal of Reproductive Medicine* 42(2): 14–129.

Hirth, R.A. et al. 2000. Willingness to pay for a quality-adjusted life year: in search of a standard. *Medical Decisionmaking* 20: 332–42.

Hoch, J.S., A.H. Briggs, and A.R. Willan. 2002. Something Old, Something New, Something Borrowed, Something Blue: A Framework for the Marriage of Health Econometrics and Cost-Effectiveness Analysis. *Health Economics* 11: 415–30.

Hoffmann, S., W. Adamowicz, and A.J. Krupnick. 2003. *Economic Uncertainties in Valuing Reductions in Children's Environmental Health Risks.* Draft. Washington, D.C.: Resources for the Future.

Johannesson, M., P. Johansson, and R. O'Conor. 1996. The Value of Private Safety versus the Value of Public Safety. *Journal of Risk Uncertainty* 13(3): 263–75

Johnson F.R., M.R. Banzhaf, and W.H. Desvousges. 2000. Willingness to Pay for Improved Respiratory and Cardiovascular Health: A Multiple-Format, Stated-Preference Approach. *Health Economics* 9: 295–317.

Johnson, F.R., E.E. Fries, and H.S. Banzhaf. 1997. Valuing Morbidity: An Integration of the Willingness-to-Pay and Health-Status-Index Literatures. *Journal of Health Economics* 16(97): 641–665.

Jones-Lee, M.W., M. Hammerton, and P.R. Philips. 1985. The Value of Safety: Results of a National Sample Survey. *Economics Journal* 95(377): 49–72.

Jones-Lee, M.W. 1991. Altruism and the value of other people's safety. *Journal of Risk and Uncertainty* 4: 213–219.

Kaplan, R.M. 1995. Utility assessment for estimating quality-adjusted life years, in *Valuing health care: Costs, benefits, and effectiveness of pharmaceuticals and other medical technologies*, edited by F. A.

Sloan. Cambridge and New York: Cambridge University Press, 31–60.

Kaplan, R.M. et al. 1993. The Quality of Well-Being Scale: Rationale for a Single Quality of Life Index. In *Quality of Life Assessment: Key Issues in the 1990s*, edited by S.R. Walker and R.M. Rosser. Dordrecht: Kluwer.

Klose, Thomas. 2003. A Utility Theoretic Model for QALYs and Willingness to Pay. *Health Economics* 12:17–31.

Kopp, R. 1993. Environmental Economics: Not Dead but Thriving. *Resources*, 111(Spring): 7–12.

Krupnick, A.J., A. Alberini, M. Cropper, N. Simon, B.J. O'Brien, R. Goeree, and M. Heintzelman. 2002. Age, Health, and the Willingness to Pay for Mortality Risk Reductions: A Contingent Valuation Survey of Ontario Residents. *Journal of Risk and Uncertainty* 24(2): 161–186.

Krupnick A.J., and M. Cropper. 1992. The Effect of Information on Health Risk Valuations. *Journal of Risk and Uncertainty* 5: 29–48.

Kuchler, F., and E. Golan. 1999. Assigning Values to Life: Comparing Methods for Valuing Health Risks. *Agricultural Economic Report Number 784*, ERS, USDA, November.

Lanoie, P., C. Pedro, and R. Latour. 1995. The Value of a Statistical Life: A Comparison of Two Approaches. *Journal of Risk and Uncertainty* 10: 235–257.

Lenert, L.A. et al. 1998. The Effect of Search Procedures on Utility Elicitations. *Medical Decisionmaking* 18: 76–83.

Lipscomb, J., M.C. Weinstein, and G.W. Torrance. 1996. Time Preference, in *Cost-Effectiveness in Health and Medicine*, edited by M.R. Gold, J.E. Siegel, L.B. Russell, and M.C. Weinstein. Oxford: Oxford University Press: 214–246.

List, J.A., and C.A. Gallet. 2001. What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? *Environmental and Resource Economics* 20(3): 241–254.

MacKenzie E.J., A. Damiano, T. Miller, and S. Luchter. 1996. Development of the Functional Capacity Index. *Journal of Trauma* 41(5): 799–807.

Mark, D.B. et al. 1995. Cost-effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *New England Journal of Medicine* 332: 1418–1424.

Mauskopf, J.A., and M.T. French. 1991. Estimating the Value of Avoiding Morbidity and Mortality from Foodborne Illness. *Risk Analysis* 11: 619–631.

Mauskopf, J.A., M.T. French, A.S. Ross, C.R. Hollingsworth, D.M. Maguire, R.W. Leukroth, and K.D. Fisher. 1988. *Estimating the Value of Consumers' Loss from Foods Violating the FD&C Act: Volume I & II Final Report*. Research Triangle Park, NC: Research Triangle Institute.

Mitchell, R.C., and R.T. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Washington, D.C.: Resources for the Future.

Moore, M.J., and W.K. Viscusi. 1988. Doubling the Estimated Value of Life: Results Using New Occupational Fatality Data. *Journal of Policy Analysis and Management* 7: 476–490.

Morall, J.F. 1986. A Review of the Record. *Regulation* 10: 25–34.

Mrozek, J.R., and L.O. Taylor. 2002. What Determines the Value of Life? A Meta-analysis. *Journal of Policy Analysis and Management* 21(2): 253–270.

Murray, C. 1994. Quantifying the Global Burden of Disease: The Technical Basis for Disability-Adjusted Life Years. *Bulletin of the World Health Organization* 72: 495–501.

Murray C, and A. Lopez. 1996. *The Global Burden of Disease*. Cambridge, MA: Harvard University Press.

Najman, J., and S. Levine. 1981. Evaluating the Impact of Medical Care and Technologies on the Quality of Life: A Review and Critique. *Social Science and Medicine* 15F: 107–115.

Neumann, P.J., D.E. Zinner, and J.C. Wright. 1997. Are Methods for Estimating QALYs in Cost-Effectiveness Analyses Improving? *Medical Decisionmaking*, 17: 402–408.

Nord, E. 1999. *Cost-Value Analysis in Health Care: Making Sense out of QALYs*. Cambridge: Cambridge University Press.

Nord, E., et al. 1995. Maximizing Health Benefits versus Egalitarianism: An Australian Survey of Health Issues. *Social Science in Medicine* 41: 1429–1437.

O'Brien, B.J., and A.H. Briggs. 2002. Analysis of Uncertainty in Health Care Cost-Effectiveness Studies: An Introduction to Statistical Issues and Methods. *Statistical Methods in Medical Research* 11: 455–68.

Office of Management and Budget (OMB). 2003. Circular A-4 *Draft Report to Congress on the Costs and Benefits of Federal Regulations*.

Parry, I.W.H. 1995. Pollution Taxes and Revenue Recycling. *Journal of Environmental Economics and Management* 29(3): S64–77.

Patrick, D.L., H.E. Starks, K.C. Cain, R.F. Uhlmann, and R.A. Pearlman. 1994. Measuring Preferences for Health States Worse than Death. *Medical Decisionmaking* 14(1): 9–18.

Patrick, D.L., and P. Erickson. 1993. *Health Status and Health Policy: Allocating Resources to Health Care* Oxford: Oxford University Press.

Raat, H. et al. 2002. Reliability and validity of comprehensive health status measures in children: The Child Health Questionnaire in relation to the Health Utilities Index. *Journal of Clinical Epidemiology* 55: 67–76.

Resources for the Future (RFF) et al. 2003. *Valuing Health Outcomes: An Assessement of Approaches*, Conference proceedings, Washington, DC. February 12–13. www.rff.org/rff/Events/calendardetail.cfm?eventID=254&eventyear=2003 (accessed December 2003).

Richardson, J., and E. Nord. 1997. The Importance of Perspective in the Measurement of Quality-Adjusted Life Years. *Medical Decisionmaking* 17: 33–41.

Rowe, R.D. et al. 1995. *The New York State Environmental Externalities Cost Study. Volume I: Introduction and Methods.* Prepared by Hagler Bailly Consulting, Inc., Boulder, CO, for the Empire State Electric Energy Research Corp., December.

Sackett, D.L., and G.W. Torrance. 1978. The Utility of Different Health States as Perceived by the General Public. *Journal of Chronic Disease* 31: 697–704.

Sagoff, M. 1993. Environmental Economics: An Epitaph. *Resources* 111(Spring): 2–7.

Scanlon, T.M. 1991. The Moral Basis of Interpersonal Comparisons. In *Interpersonal Comparisons of Well-Being*, edited by J. Elster and J.E. Roemer. Cambridge and New York: Cambridge University Press.

Schlottmann, A. 2001. Children's Probability Intuitions: Understanding the Expected Value of Complex Gambles. *Child Development* 72(1): 103–122.

Sculpher, M.J., and B.J. O'Brien. 2000. Income effects of reduced health and health effects of reduced income: implications for health-state valuation. *Medical Decisionmaking* 20(2): 207–15.

Shanmugam, K.R. 1997. The Value of Life and Injury: Estimating Using Flexible Form. *Indian Journal of Applied Economics* 6(3): 125–136.

Siegel, J.E., et al., 1996. Reporting Cost-Effectiveness Studies and Results. In *Cost-Effectiveness in Health and Medicine*, edited by M.R. Gold et al. Oxford: Oxford University Press.

Skrzycki, C. (2003) Under Fire, EPA Drops 'Senior Death Discount.' *Washington Post*, May 13 E01. www.washingtonpost.com/ac2/wp-dyn/A47678-2003May12?language=printer (accessed December 2003).

Slevin, M.L. et al. 1990. Attitude to Chemotherapy: Comparing Views of Patients with Cancer with Those of Doctors, Nurses, and General Public. *British Medical Journal* 300: 1458–1460.

Slovic, P. 1992. Perception of Risk: Reflections on the Psychometric Paradigm. In *Social Theories of Risk*, edited by S. Krimsky and D. Golding. Westport, CT: Praeger, 117–152.

Smith, V.K., G. Van Houtven, and S.K. Pattanayak. 2002, Benefit Transfer Via Preference Calibration: 'Prudential Algebra' for Policy *Land Economics* 78(1): 132–152.

______. 2003, Preference Calibration with QALYs, CENREP Working Paper, North Carolina State University (March).

Strand, J. 2002. Public- and private-good values of statistical lives—Results from a combined-choice experiment and contingent-valuation survey. *HERO Working Paper* 2.

Stratus Consulting. 1999. *Air Quality Valuation Model Canada (AQVM) User Guide*. Developed by Stratus Consulting for Environment Canada. Boulder, CO.

Sung, L. et al. 2003. Construct Validation of the Health Utilities Index and the Child Health Questionnaire in Children Undergoing Cancer Chemotherapy. *British Journal of Cancer* 88: 1185–1190.

Torrance, G.W., W.J. Furlong, D.H. Feeny, and M. Boyle. 1995. Multiattribute Preference Functions: Health Utilities Index. *Pharmaco Economics* 7: 503–520.

Torrance, G.W., D.H. Feeny, W.J. Furlong, R.D. Barr, Y. Zhang, and Q. Wang. 1996. Multiattribute Preference Functions for a Comprehensive Health Status Classification System: Health Utilities Index Mark 2. *Medical Care* 34: 702–722.

U.S. Environmental Protection Agency. 1997. *The Benefits and Costs of the Clean Air Act: 1970 to 1990*. Washington, DC: Office of Air and Radiation/Office of Policy.

U.S. Environmental Protection Agency. 1999. *The Benefits and Costs of the Clean Air Act, 1990 to 2010*. Washington, DC: Office of Air and Radiation/Office of Policy.

Viscusi, W.K. 1978. Wealth Effects and Earnings Premiums for Job Hazards. *Review of Economics and Statistics* 60(3): 408–416.

______. 1981. Occupational Safety and Health Regulation: Its Impact and Policy Alternatives. In *Research in Public Policy Analysis and Management*, edited by J.P. Crecine. Greenwich, CT: JAI Press, vol. 2, 281–289.

______. 1992. *Fatal Trade-offs: Public and Private Responsibilities for Risk*. Oxford: Oxford University Press.

______. 1993. The Value of Risks to Life and Health. *Journal of Economic Literature* 31: 1912–1946.

Viscusi, W.K., and J.E. Aldy. 2003. The Value of a Statistical Life: A Critical Review of Market Estimates Around the World. *Journal of Risk and Uncertainty* 28(1): 5–76.

Viscusi, W.K., and W. Evans. 1990. Utility Functions that are Dependent on One's Health Status: Estimates and Economic Implications. *American Economic Review* 80: 353–374.

Viscusi, W.K., W.A. Magat, and J. Huber. 1991. Pricing Environmental Health Risks: Survey Assessment of Risk-Risk and Risk-Dollar Trade-Offs for Chronic Bronchitis. *Journal of Environmental Economics and Management* 21: 32–51.

Williams, B.A.O. 1985. *Ethics and the Limits of Philosophy*. London: Collins/Fontana Press.

■ ■ ■

INDEPENDENT. BALANCED. OBJECTIVE.

RESOURCES
FOR THE FUTURE

1616 P Street, Northwest · Washington, D.C. 20036-1400
Telephone: (202) 328-5000 · Fax: (202) 939-3460
www.rff.org