USING PERFORMANCE STANDARDS TO EVALUATE SOCIAL PROGRAMS WITH INCOMPLETE OUTCOME DATA: GENERAL ISSUES AND APPLICATION TO A HIGHER EDUCATION BLOCK GRANT PROGRAM

Charles F. Manski
Department of Economics, Northwestern University
(Benjamin Meaker Visiting Professor, Department of Economics, University of Bristol)

John Newman
The World Bank

and

John V. Pepper
Department of Economics, University of Virginia

April 1999

CMPO Working Paper No. 99/008

Centre for Market and Public Organisation
University of Bristol
Department of Economics
Mary Paley Building
12 Priory Road
Bristol BS8 1TN
Tel: 0117 954 6943 Email: cmpo-office@bristol.ac.uk

1. Introduction

Agencies operating social programs often use performance standards to evaluate success in achieving outcomes of interest (e.g., see Cave and Hanney, 1992). Program outcomes are measured and compared with the standard, a threshold deemed to separate acceptable outcomes from unacceptable ones. An evaluation using a performance standard should specify not only the threshold to be used but also the action to be taken to be taken if outcomes do not meet the threshold. Discussions of performance standards are often disappointingly vague about this critical matter. However the idea usually seems to be that the threshold should be set equal to an outcome level thought achievable by some alternative, perhaps a change in the management of the program being evaluated or perhaps an entirely different program. Then a possible action is to replace the program being evaluated with the alternative if the program yields an outcome below the threshold.

Evaluation using performance standards is clearly appealing in principle. The hard questions concern implementation. This paper examines two problems of outcome measurement that confront efforts to implement standards. These are the *problem of auxiliary outcomes* and the *problem of counterfactual outcomes*.

The problem of auxiliary outcomes arises whenever considerations of timeliness or cost make it infeasible to measure the program outcomes of ultimate interest. With data on these outcomes unavailable, performance standards must be stated in terms of auxiliary outcomes that can be measured. The evaluator's problem is to set appropriate standards in terms of these auxiliary outcomes.

To illustrate, consider the problem of evaluating a pre-school program for children of age three and four. The lasting effects of such a program can be determined only by observing the treated persons as they grow from infants into adults, but some early effects may be observable soon after the administration of treatments. For example, a child's cognitive status may be measured at age five when the child enters regular schooling. The evaluator's problem is to use the available data on early outcomes to set standards, when the evaluator's real interest is in the lasting effects of the program.

The problem of counterfactual outcomes concerns the alternative serving as the standard of comparison for the program being evaluated. Whereas the program being evaluated is operational and so its outcomes are at least observable in principle, the alternative is not in operation and so its outcomes are counterfactual. To appropriately set the threshold defining a performance standard, an evaluator must somehow predict what outcomes would occur if the alternative were in operation.

Consider again the problem of evaluating a pre-school program. The alternative may be an incremental change in the way that pre-school programs are accredited and funded. Or, more drastically, it may be the replacement of publicly funded pre-schools with a child-allowance program giving families with young children cash payments that they may spend for any purpose, including but not restricted to pre-schooling. In the absence of an operational child-allowance program, the evaluator cannot observe even the auxiliary outcomes of such a program, never mind its lasting outcomes.

This paper examines how performance standards should be set and applied in the face of these problems in measuring outcomes. Our central message is that the proper way to

implement standards varies with the prior information that the evaluator can credibly bring to bear to compensate for incomplete outcome data. If this prior information is sufficiently strong, the traditional practice of using a single threshold to separate acceptable outcomes from unacceptable ones is appropriate. An evaluator having weaker prior information however, should set two thresholds rather than one. The performance of the program should be deemed acceptable if the observed auxiliary outcomes meet the higher *acceptance threshold* and unacceptable if they fall below the lower *nonacceptance threshold*.

If the auxiliary outcomes lie between these two thresholds, the performance of the program is indeterminate. In this case, there is insufficient basis for deciding whether the program being evaluated should be continued or replaced by the alternative. Decisions to continue the program or to replace it are both defensible given the available information. Efforts to obtain more information before making a decision may be justified.

We develop these ideas in two stages. Sections 2 through 5 consider the evaluation problem in some generality. Section 2 formalizes basic concepts: treatments, outcomes, programs, and treatment effects. Sections 3 and 4 use these concepts to address the problems of auxiliary outcomes and counterfactual outcomes respectively. Section 5 asks what the evaluator should do when the available information yields an indeterminate finding about the performance of the program being evaluated. Throughout Sections 2 through 5, we use the problem of evaluating a pre-school program to illustrate ideas as they are introduced. However these sections aim to make general points, so most of the discussion is necessarily abstract.

In Sections 6 through 8, we shift from generalities to the specifics of an actual evaluation problem. Here we explore in some depth how the problems of auxiliary and counterfactual

outcomes arise and may be addressed in the new World Bank sponsored Quality of Undergraduate Education (QUE) program administered by the Indonesian Board of Higher Education. Section 6 describes the QUE program, which awards competitive five-year block grants to university departments to improve the quality of their undergraduate curricula. Sections 7 and 8 examine two distinct ways in which performance standards will be used. In the short run, the progress of QUE grantees in achieving specified auxiliary outcome targets will be monitored. Then, at the end of the five-year grant period, the QUE program as a whole will be evaluated.

QUE is representative of a large class of programs that use block grants and similar decentralized decision making mechanisms to achieve social objectives. Our examination of the QUE program has lessons for the evaluation of other block grant programs. In particular, the analysis of Section 8 shows the need for integrated micro evaluation of particular grantees and macro evaluation of the program as a whole.

2. Concepts of Formal Evaluation

The usual formalization of a program evaluation assumes that each member j of a population J receives one of several mutually exclusive and exhaustive *treatments*. Each member of this population experiences a scalar outcome-of-interest that may depend on the treatment received. The possible treatments will be numbered t = 1,...,T. The outcomes associated with these treatments are y(t), t = 1,...,T. Supposing that the outcome-of-interest is scalar does not rule out the possibility that a person experiences multiple outcomes following

treatment. The outcome-of-interest transforms these multiple outcomes into a single measure that expresses the overall value of the treatment.

The treatment that a person receives depends on the set of treatments available to this person and on the person's choice of a treatment from this set. Social programs help determine the set of available treatments and thus influence the treatments that people receive. It will suffice to consider two programs. One of these is the operational program being evaluated, labeled program A here. The other is the alternative with which the operational program is to be compared, labeled program B. Let $z_{jA} \in T$ indicate the treatment that person j actually receives under program A, and let $z_{jB} \in T$ indicate the treatment that this person would receive under program B. Then the outcomes this person does experience under program A and would experience under program B are $y(z_{jA})$ and $y(z_{jB})$ respectively.

The objective of the evaluation is to determine which yields the better outcomes, program A or B. The usual practice is to compare programs in terms of their mean outcomes across the population. Let $E[y(z_A)]$ and $E[y(z_B)]$ denote the mean outcomes under programs A and B. The quantity

$$(1) \quad d(A, B) \equiv E[y(z_A)] - E[y(z_B)]$$

is the *average treatment effect* of program A relative to program B. If $d(A, B)$ is positive, the performance of program A may be deemed acceptable. Thus the mean outcome of program B provides the threshold relative to which program A's outcomes are judged.

Implementing the performance standard is straightforward if the evaluator observes the

outcomes $y(z_A)$ and $y(z_B)$ of the members of the population, or at least those of random samples of the population. Then the evaluator may learn the mean outcomes $E[y(z_A)]$ and $E[y(z_B)]$ and determine whether the former exceeds the latter.

Our concern is with evaluation in the absence of complete outcome data. The problem of auxiliary outcomes arises when the evaluator observes a vector of auxiliary outcomes of program A, say $w(z_A)$, but not the outcome-of-interest $y(z_A)$. The problem of counterfactual outcomes is that, program B not being in operation, its outcomes are unobservable in principle.

To illustrate these concepts, consider the problem of evaluating a pre-school program. In this case, the set T of possible treatments may index different types of pre-schooling; for example, some pre-schools may emphasize basic reading and arithmetic skills, others may encourage higher order thinking, and still others may seek to promote children's social and emotional development. Program A is the pre-school program being evaluated and program B is an alternative. The outcome-of-interest y may measure cognitive status at age twenty. The observed auxiliary outcomes w may measure a child's cognitive status at age five.

The performance of the pre-school program may be deemed acceptable if mean cognitive status at age twenty is higher under the pre-school program than under the child-allowance program. The problem of auxiliary outcomes is that only cognitive status at age five is observed under the operational pre-school program. The problem of counterfactual outcomes is that no cognitive measurements at all are possible under the counterfactual alternative to program A.

3. The Problem of Auxiliary Outcomes

In this section we investigate the problem of auxiliary outcomes while abstracting from the problem of counterfactual outcomes. We suppose that the evaluator sets a threshold that the mean outcome of program A must meet to be deemed acceptable. We do not ask here how this threshold is determined, but this question will be addressed in Section 4. Let c denote the threshold set by the evaluator. Then the criterion for judging the performance of program A is this:

(2) Program A is acceptable if $E[y(z_A)] \geq c$.

The evaluator observes only the auxiliary outcomes $w(z_A)$ of the population and not their outcomes-of-interest $y(z_A)$. The problem is to use the data on auxiliary outcomes to learn about $E[y(z_A)]$. A common practice is to judge Program A to be acceptable if the expected value of a chosen scalar function of the auxiliary outcomes meets a specified threshold. Thus performance criterion (2) is replaced by one of the form

(3) Program A is acceptable if $E\{f[w(z_A)]\} \geq d$.

Here f(·) is the chosen function and d is the specified threshold.

In general, criteria (2) and (3) are distinct and so may yield different conclusions about program performance. However the two criteria are equivalent if f(·) and d are chosen in a particular way. Suppose that the auxiliary outcome vector w can takes S possible values,

numbered s = 1,...,S for convenience. Use the law of iterated expectations to write

$$(4)\quad E[y(z_A)] = \sum_{s=1}^{S} E[y(z_A)\,|\,w(z_A) = s]\cdot P[w(z_A) = s].$$

Here $E[y(z_A)\,|\,w(z_A) = s]$ is the mean value of the outcome-of-interest among the people who realize value s of the auxiliary outcome, and $P[w(z_A) = s]$ is the fraction of the population who realize value s. Compare equation (4) with $E\{f[w(z_A)]\}$, which has the form

$$(5)\quad E\{f[w(z_A)]\} = \sum_{s=1}^{S} f(s)\cdot P[w(z_A) = s].$$

It follows that $E\{f[w(z_A)]\} = E[y(z_A)]$ if f(·) is chosen to be the function

$$(6)\quad f(s) = E[y(z_A)\,|\,w(z_A) = s].$$

With this choice of f(·) and with d set equal to c, performance criteria (2) and (3) are equivalent.

We have just shown how auxiliary outcome data should be used to judge performance if the evaluation is to remain focused on the outcomes-of-interest $y(z_A)$. Application of criterion (3) with other choices of f(·) and d may yield distorted conclusions about the acceptability of the program being evaluated. The practical problem, of course, is that auxiliary outcome data alone do not reveal the conditional means $E[y(z_A)\,|\,w(z_A) = s]$, s = 1,...,S. Thus implementation of the

appropriate version of criterion (3) is possible only if the evaluator can bring to bear other information that reveal $E[y(z_A)?w(z_A) = s]$, s = 1,...,S.

The possibilities explored here all assume the existence of some historical period in which data were collected on both the auxiliary outcomes w and the outcome-of-interest y. These historical data may pertain to an environment that is different in some respects from that of program A. They may nevertheless be used to inform the evaluation of program A, provided that the historical period for which (w, y) data are available shares some common features with the environment under program A. Sections 3.1 and 3.2 make this explicit.

3.1. The Equal-Conditional-Means Assumption

Let the observable historical distribution of (w, y) be denoted $P_H(w, y)$. Assume that each value of w realized by a positive fraction of the population under program A was also realized by a positive fraction of the population in the historical period. That is, for s = 1,...,S,

(7) $P[w(z_A) = s] > 0 \Rightarrow P_H(w = s) > 0.$

Now assume that, for each s such that $P[w(z_A) = s] > 0$, the conditional mean outcome $y(z_A)$ under program A equals the conditional mean historical outcome y. That is,

(8) $E[y(z_A)?w(z_A) = s] = E_H(y \mid w = s).$

This *equal-conditional-means assumption* and the law of iterated expectations (4) yield

$$(9)\quad E[y(z_A)] = \sum_{s=1}^{S} E_H(y \mid w = s) \cdot P[w(z_A) = s].$$

By assumption (7), the historical data on (w, y) reveal $E_H(y \mid w = s)$ whenever $P[w(z_A) = s] > 0$. The auxiliary outcome data on program A reveal $P[w(z_A) = s]$ for all values of s. Hence the evaluator can use the right side of equation (9) to learn $E[y(z_A)]$ and so judge the performance of program A.

The credibility of the equal-conditional-means assumption must be assessed on a case-by-case basis. The identity of the measured auxiliary outcomes may be critical, the assumption being credible conditional on some specifications of the auxiliary outcomes but not others. Note that treatments and covariates can serve as auxiliary outcomes. To formalize treatment as an auxiliary outcome, we simply define $w(z_A) \equiv z_A$. A covariate___ e.g., race or sex__ is simply an auxiliary outcome whose value varies across the population but not across treatments; that is, $w(z_A)$ does not vary with z_A. Thus treatments and covariates are two polar forms of auxiliary outcomes.

In the example of the pre-school program, the equal-conditional-means assumption states that the unobserved mean cognitive status at age twenty of persons who have measured cognitive status s at age five equals the observed historical mean cognitive status at age twenty among persons who had measured cognitive status s at age five. Is this a reasonable assumption? It is if one thinks that the pre-school program influences adult cognitive status

through its effect on measured cognitive status at age five, but not otherwise. The assumption is less reasonable if one thinks that the program may influence adult cognitive status through a developmental process that does not entirely manifest itself in measured cognitive status at age five.

3.2. Bounded-Conditional-Means Assumptions

An equal-conditional-means assumption is sufficient but not necessary to determine if the performance of program A is acceptable. Whereas this assumption identifies $E[y(z_A)]$, we only need to learn if $E[y(z_A)]$ meets the threshold c.

A flexible way to weaken an equal-conditional-means assumption is to use knowledge of $E_H(y \mid w = s)$ to bound $E[y(z_A) \mid w(z_A) = s]$. Supposing that y takes positive values, a particularly simple *bounded-conditional-means assumption* is

$$(10) \quad a \cdot E_H(y \mid w = s) \leq E[y(z_A) \mid w(z_A) = s] \leq ß \cdot E_H(y \mid w = s),$$

Here a and ß are constants such that $0 \leq a \leq ß \leq \infty$. These constants, specified by the evaluator, express the strength of the association that the evaluator feels comfortable asserting between $E_H(y \mid w = s)$ and $E[y(z_A) \mid w(z_A) = s]$. If $a = ß = 1$, we have the equal-conditional-means

assumption. If $\alpha = 0$ and $\beta = \infty$, measurement of $E_H(y \mid w = s)$ reveals nothing about $E[y(z_A) \mid w(z_A) = s]$.

Assumption (10) and the law of iterated expectations (4) imply this bound on $E[y(z_A)]$:

$$(11) \quad \sum_{s=1}^{S} E_H(y \mid w = s) \cdot P[w(z_A) = s] \leq E[y(z_A)] \leq \sum_{s=1}^{S} E_H(y \mid w = s) \cdot P[w(z_A) = s].$$

If the lower bound on $E[y(z_A)]$ meets the threshold c, the evaluator can conclude that the performance of program A is acceptable. If the upper bound on $E[y(z_A)]$ is less than c, he can conclude that the program's performance is unacceptable. Otherwise, the status of program A is indeterminate given the available data and prior information.

Consider the pre-school program. There are many reasons why an evaluator may not be willing to make an equal-conditional-means assumption. It may be that schooling norms have changed between the historical period and the present, with consequent changes in the association between pre-schooling and adult cognitive status. Or it may be that the very act of evaluating the pre-school program has incentive effects that alter the association between childhood and adult cognitive status. Administrators of the pre-school program, knowing that measured cognitive status at age five will be used to evaluate program performance, may choose to emphasize forms of pre-schooling that have measurable effects on cognitive status at age five rather than ones whose effects become measurable later on.

Concerned with these and other possibilities, the evaluator may find a bounded-conditional-means assumption to be more credible. The evaluator may be willing to assume (10),

perhaps with $\underline{a} = 0.8$ and $ß = 1.2$. That is, he may be willing to assume that the unobserved mean cognitive status at age twenty among persons with cognitive status s at age five is between eighty percent and one-hundred-twenty percent of the observed historical mean cognitive status at age twenty among persons with cognitive status s at age five. This assumption may suffice to judge whether the pre-school program is acceptable.

4. The Problem of Counterfactual Outcomes

Discussions of performance standards often exhibit considerable lack of clarity on how the threshold separating acceptable from unacceptable performance should be set and what action should be taken if performance is deemed unacceptable. Much of the difficulty that evaluators have in specifying thresholds and actions stems from the problem of counterfactual outcomes. In principle, the threshold should be set equal to an outcome level known to be achievable by an alternative feasible program, and this alternative should replace the operational program if the threshold is not met. However the outcomes that would occur under counterfactual alternatives are not observable. Hence, even abstracting from the problem of auxiliary outcomes, evaluators inevitably find it hard to specify what constitutes acceptable program performance.

The rich econometric literature on the analysis of treatment effects teaches that there is no unique resolution of the problem of counterfactual outcomes. The conclusions that can be drawn about the outcomes of counterfactual programs depend critically on what historical data are available and what prior information the evaluator can credibly bring to bear.

The dominant concern of the econometric literature has been to predict the outcomes of mandatory treatment programs__ ones giving the same treatment to all members of the population__ when the available historical data pertain to an environment in which treatment varies across the population. In this context, the problem of counterfactual outcomes is known as the *selection problem*. Analyses of the selection problem show that if historical data on the outcome of interest is combined with sufficiently strong assumptions, the counterfactual mean outcome $E[y(z_B)]$ may be identified, implying a well-defined threshold for judging the performance of program A. In practice, the most common assumption is that treatments are statistically independent of outcomes in the historical data, as they would be in a classical randomized experiment. An alternative route to identification is to assert a parametric latent variable model jointly describing how treatments are selected and outcomes determined. Another alternative is to assume that treatment effects are constant across the population and that there exists some covariate, termed an *instrumental variable*, that is mean independent of outcomes but not of treatments. See Maddala (1983), Heckman and Robb (1985), and Manski (1995) for reviews of the literature.

Concern with the credibility of the strong assumptions needed to identify treatment effects has led to the recent development of a literature imposing weak assumptions that yield bounds on the counterfactual mean outcome $E[y(z_B)]$. The starting point is to ask what can be learned about $E[y(z_B)]$ from the historical data if no assumptions at all are made about the process determining treatment selection and outcomes. The result is a "no-assumptions" bound on $E[y(z_B)]$. From this base, the evaluator may impose weak assumptions that have identifying power in the sense that they yield narrower bounds. One set of results illuminates the identifying power of instrumental

variable assumptions when imposed alone, treatment effects not being assumed to be constant across the population. See Manski (1990, 1994) and Manski and Pepper (1997). Another set of results shows the identifying power of various assumptions about the treatment selection process when nothing is known about the process determining outcomes. For example, one may assume that each member of the population was assigned the treatment yielding the better outcome for that person. See Manski (1994, 1995) and Manski and Nagin (1998). Yet another set of results shows the identifying power of assumptions about the process determining outcomes when nothing is known about the treatment selection process. For example, one may assume that treatment response is monotone, in the sense that the outcome of one treatment is always at least as good as the outcome of the other. See Manski (1995, 1997a) and Pepper (1997).

When the available historical data and assumptions suffice to bound but not identify $E[y(z_B)]$, the conventional idea of using a single threshold to separate acceptable from unacceptable outcomes needs revision. Suppose that the available historical data and credible assumptions imply that $c_0 \leq E[y(z_B)] \leq c_1$, for known constants c_0 and c_1. Suppose that the available historical data, auxiliary outcome data, and credible assumptions imply that $d_0 \leq E[y(z_A)] \leq d_1$, for known constants d_0 and d_1. Then the evaluator may conclude that

(2) Program A is acceptable if $d_0 - c_1 \geq 0$ and unacceptable if $d_1 - c_0 < 0$.

Otherwise, the performance of program A relative to B is indeterminate.

The same considerations apply when the alternative program B does not mandate a single treatment but rather permits treatment to vary across the population (see Manski, 1997b). The

general point remains that application of a conventional performance standard with a single threshold to separate acceptable from unacceptable outcomes is appropriate only if the evaluator can bring to bear sufficiently strong data and assumptions. In other settings, the performance of program A has three possible states: acceptable, unacceptable, or indeterminate.

Consider the pre-school program. Suppose that the alternative is the child-allowance program, which is not in operation. How might the evaluator predict what its outcomes would be? A possible approach would be to examine the historical association between family income and child cognitive outcomes. The evaluator might assume that child allowances are equivalent to increases in family income and that the historical distribution of child cognitive status conditional on family income would not be changed by the child-allowance program. These assumptions would enable determination of $E[y(z_B)]$.

Yet the credibility of these assumptions may well be questioned. Child allowances lower the price of children and so may influence families to have larger families than they would with the same increase in income. These increases in family size may disrupt the historical distribution of child cognitive status conditional on family income. Uncertain of the magnitude of the fertility and induced cognitive effects, it may be that the evaluator cannot credibly determine $E[y(z_B)]$ but can perhaps bound its magnitude.

5. Should Indeterminacy be Tolerated or Resolved?

Suppose that an evaluation yields an indeterminate finding about the program's acceptability. There are potentially two ways to resolve the ambiguity. One can always impose stronger assumptions. One can sometimes collect richer auxiliary outcome and/or historical data.

It is tempting to impose assumptions strong enough to yield a definitive finding. Whereas data collection can be costly and time-consuming, imposing assumptions requires only a leap of faith. The problem, of course, is that strong assumptions may be inaccurate and yield flawed conclusions. Even if an evaluator personally considers an assumption to be plausible, he must be concerned about the credibility of his findings to policymakers and the public. These may be a diverse group some of whose members may not share the evaluator's beliefs about what are and are not plausible assumptions. The evaluator must keep in mind that the weaker are the assumptions imposed, the more widely credible are the reported findings.

If stronger assumptions are not imposed, the only way to resolve an indeterminate finding is to collect richer outcome data. In Sections 3 and 4, we examined the evaluation problem given specified data, without saying anything about how these data came to be available. In practice, evaluators play a role in determining what outcome data should be collected. Evaluators may be able to influence the collection of historical data on auxiliary outcomes and outcomes of interest, thus enabling application of the ideas developed in Sections 3.1 and 3.2. Evaluators may also be able to influence the collection of outcome data in program A, thus reducing the distance between the available auxiliary outcomes and the outcomes of interest. So let us suppose that it is feasible to collect richer outcome data, either historical data or outcome data on program A. Then the evaluator must decide whether the benefits of new data collection exceed the cost.

Let us explore this matter through a simple illustration which makes some general points.

Assume that the outcome y and the cost of data collection are measured in the same units, say in monetary terms. Assume that an evaluation with the available data reveals that the average treatment effect $\Delta^d(A, B)$ lies in the interval $[b_0, b_1]$, where $b_0 < 0 < b_1$ are known constants. Thus the evaluation does not reveal whether program A or B is preferable.

Suppose that, for a known cost K, it is possible to collect new data that enable the evaluator to shrink the interval within which $\Delta^d(A, B)$ lies from $[b_0, b_1]$ to, say $[b_0', b_1']$, the latter interval being a subset of the former one. The benefit of additional data collection takes the form of an improved choice between the two programs, so we need to specify how the choice between these programs is made if the evaluation yields an indeterminate finding. Assume that policymaking is conservative in the sense that, given an indeterminate finding, the operational program is chosen over the alternative. In this setting, should the new data be collected?

Examination of the situation shows that the benefit of new data collection is zero if $b_1' \geq 0$, In this case, the evaluation is still indeterminate after collection of the new data. Given the assumption of a conservative policymaker, the same decision is made both with and without the new data, namely to retain the operational program. The benefit of new data collection is positive if $b_1' < 0$. In this case, the new data reveal that $\Delta^d(A, B)$ is negative, implying that program B is preferable to program A. The resulting gain is $-\Delta^d(A, B)$.

Unfortunately, the evaluator can only learn the value of b_1' after collecting the new data. Ex ante, he knows only that $b_0 \leq \Delta^d(A, B) \leq b_1$, implying that the benefit of new data collection falls in the interval $[0, -b_0]$. If the cost K of new data collection is larger than $-b_0$, new data collection definitely is not worthwhile. If K is less than $-b_0$, new data collection may or may not be worthwhile. In the latter case, ambiguity about the performance of program A relative to B

generates ambiguity about the desirability of new data collection.

The Bayesian literature on optimal statistical decisions proposes that ambiguity about the desirability of new data collection, whether in the above illustration or in more complex settings, can be resolved by asserting a subjective probability distribution on the quantities that the evaluator neither observes empirically nor knows a priori. See, for example, Spencer (1985) and Spencer and Moses (1990). When application of this subjective distribution yields the conclusion that new data should not be collected, the subjective distribution is then used directly to decide between programs A and B. A subjective distribution is an assumption and, as such, requires a leap of faith. The need to make this assumption is the price that Bayesian decision theory pays to resolve indeterminate evaluations.

6. The "Quality of Undergraduate Education" Program in Indonesia

In the remainder of the paper, we use an actual program evaluation to study in more depth how problems of incomplete outcome data arise in practice and how they affect the implementation of performance standards. In this section, we describe the established features of the new Quality of Undergraduate Education (QUE) program in Indonesia and call attention to important unresolved questions. With this as background, Sections 7 and 8 examine the monitoring and evaluation problems associated with this program.

6.1. Basic Description of the QUE Program

The QUE program was recently initiated by the Indonesian government's Board of Higher Education (BHE) as a component of a portfolio of educational programs supported by the World Bank. All academic departments in public universities were invited to submit proposals for block grants to improve the quality of the undergraduate education they provide. The first round of the competition for these grants was carried out in 1997. Pre-proposals were received from 317 departments, 45 of which were invited to submit proposals. Eventually, 16 five-year grants were awarded with funding levels averaging 400,000 U.S. dollars per year. The grants are meant to provide new funding to the recipient departments, supplementing their regular budgets.

Departments submitting proposals were required to provide self-assessments of their strengths and weaknesses and to propose action plans detailing the use they would make of BHE funding. However the terms of the grants give recipients full discretion in the use of the new funds. By agreement between the BHE and the World Bank, the performance of the QUE program is to be judged by the program's effects on student outcomes, not by the particular ways in which the grantee departments use their funds.

The outcome of interest to the BHE is, broadly speaking, the value to Indonesian society of having high quality university graduates, both of the departments that receive QUE grants and of those which do not. In practice, the BHE and the World Bank have agreed that the program will be first monitored and then evaluated using data to be collected on at least these seven auxiliary outcomes, which are officially termed *performance indicators*:

w1. NEE Score - average score of the department's students on the National Entrance Examination. (The NEE is used to admit students to departments.)

w2. GPA - average Grade Point Average of students enrolled in the department.

w3. TOEFL Score - average score on the Test of English as a Foreign Language, administered to graduating students.

w4. Time to Degree - average length of time that students are enrolled in the department en route to graduation.

w5. Time to Employment - average length of time that students take to secure employment following graduation.

w6. GRE Score - average score on the subject-area Graduate Record Examination, administered to graduating students.

w7. Peer Evaluation - a rating of department quality by international peer reviewers.

6.2. Monitoring and Evaluation

The BHE and the World Bank have agreed to monitor the auxiliary outcomes experienced by the current grantees during 1997 - 2002 and then to evaluate the QUE program in 2002 at the end of the five-year grant period. Monitoring means that the BHE will assess the performance of grantees in achieving auxiliary outcome targets agreed upon by the grantees and the BHE. If a department's performance in meeting its targets is deemed to be inadequate, the BHE may take limited corrective actions depending on the particulars of the case. It may, for example, provide technical assistance to a department with inexperienced personnel. It may also delay the release or reduce the size of a payment. The presumption, however, is that barring an incident of gross

negligence or fraud, the grantee will continue to receive its annual funding throughout the five-year grant period. See Section 7 for further discussion.

Although monitoring has some of the character of an evaluation, the BHE usefully maintains a distinction between monitoring and the evaluation of the QUE program that will take place in 2002, when the BHE must decide whether to continue the QUE program or to replace it with an alternative. At this point, we need to confront the fact that the QUE program is a work in progress rather than a fully-articulated funding program. The BHE and World Bank have not yet stated what it would mean to continue the QUE program after 2002. There are numerous reasonable possibilities. Each implies a different definition for the QUE program and, consequently, each implies different feasible alternatives.

Here are three possibilities, all of which maintain a constant level of funding for the QUE program:

Indefinite Funding - One interpretation of the QUE program is that the sixteen grants awarded in 1997 would be continued indefinitely, with no new grants being awarded to other departments.

Open Re-competition - A second interpretation is that a new grant competition would be held every five years, all university departments being eligible to compete as in the initial competition in 1997. Present grantees would be eligible to submit new proposals but would enjoy no special status when the grants are re-competed.

Performance-Based Grant Renewal - A third possibility is that present grantees would have their

grants renewed for an additional five-year period if their measured auxiliary outcomes are judged to be acceptable, but not renewed if their auxiliary outcomes are judged non-acceptable. Every five years a new grant competition would be held to re-allocate those QUE funds that become available when some grantees do not have their grants renewed.

It is easy enough to think of variations on these possibilities, as well as other options that become feasible if the funding level of the QUE program is itself considerable variable.

Each definition of the QUE program implies different alternatives to QUE. Suppose, for example, that the BHE should interpret the QUE program to mean indefinite funding of the present grantees. Then the open re-competition and performance-based renewal designs described above are alternatives to QUE. Other alternatives might retain the competitive funding idea of QUE but alter the number of grants or the award per grantee. Yet another alternative would be to abandon competitive funding entirely and return to the non-competitive "baseline" funding system that was used until 1997.

In Section 8, we select one version of the QUE program and one alternative for further study. We suppose that in 2002 the BHE will interpret the QUE program to be the performance-based grant renewal design and will take the relevant alternative to be the baseline non-competitive funding system used until 1997. Performance-based renewal is a particularly interesting interpretation of QUE because it encompasses indefinite funding and open re-competition as special cases. If the threshold for grant renewal is set so low that all existing grants are renewed, performance-based renewal is equivalent to indefinite funding. If the threshold is set so high that no existing grants are renewed, performance-based renewal is

equivalent to open re-competition.

7. Monitoring The QUE Grantees

Grants from government agencies commonly carry provisions for monitoring grantees during the periods of their grants. Monitoring often focuses on matters of process -- how the grant is managed, the nature of the expenditures made, etc. In contrast, the QUE program calls for monitoring certain outcomes realized by grantees.

Each of the sixteen QUE grants specifies target changes in performance indicators w1 through w5 to be achieved 2.5 years and five years after grant initiation. These midterm and final targets, which vary across the departments receiving grants, were established by negotiation between the BHE and the departments. The BHE has yet to determine how it will use the targets to assess departments' performance. We consider this question here, restricting attention to the midterm targets. As will become evident in Section 8, evaluation of the QUE program at the end of five years involves distinct considerations.

Consider the situation of one QUE grantee, say university department j. Let w_{Tj} and w_{Rj} denote this department's midterm target and realized values of the performance indicators w1 through w5. The discussion of Section 3 suggests that the BHE should view these performance

indicators as auxiliary outcomes which may be used to predict the outcome-of-interest, namely the value to Indonesian society of having high quality university graduates. To cast this idea in conventional economic terms, we might interpret the BHE as wanting to maximize the difference between the life-cycle discounted earnings of university entrants and the cost of providing their education.

Formally, let the QUE program be designated as program A, Let $I_j(A)$ denote the average life-cycle discounted earnings of enrollees in department j under the QUE program. Let $N_j(A)$ denote the number of university entrants who enroll in department j. Let $C_j(A)$ be the budget that this department receives under the QUE program. Then we take the outcome-of-interest $y_j(A)$ to be the difference between the earnings of department j's enrollees and the cost of operating the educational component of this department, namely

(13) $y_j(A) \equiv N_j(A) \cdot I_j(A) - C_j(A)$.

Let $E[y(z_A) | w(z_A) = w_{Tj}]$ and $E[y(z_A) | w(z_A) = w_{Rj}]$ be the mean values of the outcome-of-interest conditional on the performance indicators taking the values w_{Tj} and w_{Rj} respectively. Then the BHE might use this criterion to monitor the midterm performance of department j:

(14) Midterm Performance is acceptable if $E[y(z_A) | w(z_A) = w_{Rj}] \geq E[y(z_A) | w(z_A) = w_{Tj}]$.

To implement this criterion as stated requires that the BHE know the conditional means $E[y(z_A) | w(z_A) = w_{Rj}]$ and $E[y(z_A) | w(z_A) = w_{Tj}]$. As discussed in Section 3.1, these quantities are

knowable if historical data on (w, y) are available and if the BHE is able to credibly assert the equal-conditional-means assumption. Under weaker bounded-conditional-means assumptions of the form discussed in Section 3.2, the BHE can conclude that midterm performance is acceptable (unacceptable) if the lower (upper) bound on $E[y(z_A)|w(z_A) = w_{Rj}]$ lies above (below) the upper (lower) bound on $E[y(z_A)|w(z_A) = w_{Tj}]$. If neither of these conditions hold, then midterm performance is indeterminate.

There are other assumptions that the BHE might want to bring to bear. It may, for example, be credible to assume that the mean value of the outcome-of-interest varies monotonically with each of the five performance indicators. In particular, the value of university graduates may be thought to be increasing in their test scores (w1, w2, w3) and decreasing in the times (w4, w5) required to obtain their degrees and find employment. Under this assumption, the BHE can conclude that midterm performance is acceptable (unacceptable) if all of the five realized values of the indicators are better (worse) than the corresponding target values. If some realized indicators are better than their target values and others worse, then midterm performance is indeterminate.

8. Evaluation of QUE: Comparison of Performance-Based Renewal and Non-competitive Funding

In this section we examine the BHE's decision problem in 2002, at the end of the five-year grant period. In particular, we consider here how the BHE might compare performance-based QUE grant renewal with the alternative of baseline non-competitive funding. Our discussion is

intended to develop some important points, but not to cover all of the difficult issues that the BHE may need to consider. Hence we shall make some simplifying assumptions that we would not necessarily endorse in practice. These are

(A1) In 2002, the BHE is only concerned with the next round of five-year QUE grants. It does not commit itself to the QUE program beyond 2007 nor otherwise consider how departments should be funded beyond that date.

(A2) Departments that receive QUE grants continue to receive their baseline non-competitive funding as well. The size of QUE grants is not a decision variable for the BHE. All QUE grants have the same pre-determined size, denoted G.

(A3) Should a department receiving a 1997 QUE grant have its grant renewed in 2002, students who enroll in this department in the period 2002 - 2007 realize the same average outcomes as do students in this department in the period 1997-2002. Students who enroll during 2002 - 2007 in departments that receive new QUE grants in 2002 realize the same average outcomes as do students who enroll during 1997 - 2002 in the sixteen departments receiving QUE grants in 1997.

(A4) Continuation of the QUE program from 2002 to 2007 only affects the sixteen departments that receive grants in 2002. Departments that do not receive grants at that time have the same funding and student outcomes under programs A and B.

Assumptions (A1) through (A4) greatly simplify the BHE's evaluation problem. We caution, however, that these assumptions should not be taken lightly. The BHE should, in principle, think beyond the next round of grants and so (A1) may not hold. University administrations may seek to use QUE funding to substitute for departmental baseline funding, thereby violating (A2). Moreover, the BHE may give QUE grants of different sizes to different departments, also violating (A2). Assumption (A3) is plausible if relevant aspects of the higher education environment__ the characteristics of university students, the mix of departments applying for QUE grants, the BHE's decision process in awarding grants, the state of the Indonesian labor market, etc.__ do not change between 1997 and 2002. However changes in the environment may occur and make this assumption suspect. For example, the mix of departments applying for new QUE grants in 2002 may differ from the mix that applied in 1997.

As for Assumption (A4), there are several reasons why the QUE program may affect departments that do not receive grants. QUE funding may allow the departments that receive grants to compete more effectively for students, thus altering the student bodies at non-recipient departments. QUE funding may allow students in departments that receive grants to compete more effectively for a limited supply of jobs after graduation, thus altering the job prospects of the graduates of other departments. Moreover, the process of writing proposals for QUE funding may lead departments to critically appraise and improve their educational programs, even if they do not receive funding.

With these caveats in mind, we lay out general features of the evaluation problem in Section 8.1 and then develop the implications of Assumptions (A1) through (A4) in Sections 8.2

and 8.3. In Section 8.2, we abstract from the problems of auxiliary and counterfactual outcomes and consider how the BHE should act if it were somehow to have complete outcome data. In Section 8.3, we consider how the BHE should act given the outcome data that are likely to be available.

8.1. General Features of the Evaluation Problem

Let us suppose that there is a population J of university departments in Indonesia. In general terms, the QUE program affects the funding of these departments. Abstracting from QUE, let F denote a program for funding university departments. The mean outcome of funding program F is

$$(15)\quad E[y_j(F)] \equiv \frac{1}{|J|} \sum_{j \in J} N_j(F) \cdot I_j(F) - C_j(F),$$

where $|J|$ is the number of university departments. We shall interpret the BHE as wanting to choose a funding program that maximizes $E[y_j(F)]$.

By assumption, the feasible options are the performance-based renewal version of the QUE program and the baseline noncompetitive funding mechanism. In the notation of Sections 2 through 5, QUE is program A and the baseline alternative is program B. Applying equation (15), we suppose that the BHE would judge QUE to have acceptable outcomes if

$$(16)\quad \sum_{j \in J} S \left[N_j(A) \cdot I_j(A) - N_j(B) \cdot I_j(B) \right] - \left[C_j(A) - C_j(B) \right] \geq 0.$$

8.2. Evaluation With Complete Outcome Data

From this point on, we maintain Assumptions (A1) through (A4). Let J_1 denote the sixteen departments that received QUE grants in 1997. Let

$$(17)\quad d_1(A, B) \equiv \frac{1}{16} \sum_{j \in J_1} [N_j(A) \cdot I_j(A) - N_j(B) \cdot I_j(B) - G]$$

be the average difference between the outcomes that these departments realize and those that they would have experienced if they had not received QUE grants.

Let J_2 denote a hypothetical set of sixteen departments that would receive grants in 2002 if QUE is continued. Some of these, denoted J_{21}, would be members of J_1 that have their grants renewed. The remaining $16 - |J_{21}|$ members of J_2 would be new grant recipients. Assumption (A1) through (A3) imply that the average difference between the outcomes that the departments in J_2 would realize with their QUE grants and those that they would experience in the absence of the grants is

$$(18)\quad d_2(A, B) \equiv \frac{1}{16} \sum_{j \in J_2} [N_j(A) \cdot I_j(A) - N_j(B) \cdot I_j(B) - G]$$

$$= \frac{1}{16} \{ [16 - |J_{21}|] \cdot d_1(A, B) + \sum_{j \in J_{21}} [N_j(A) \cdot I_j(A) - N_j(B) \cdot I_j(B) - G] \}.$$

The term [16 - $|J_{21}|$].d_1(A, B) on the right side of (18) reflects the second part of Assumption (A3), which states that students in departments that receive new QUE grants in 2002 realize the same average outcomes as do students in the sixteen departments who received QUE grants in 1997.

By Assumption (A4), the QUE program does not affect departments that do not receive grants. Hence, in 2002, the BHE should use a two-stage process to decide which department should have their grants renewed and whether the QUE program should be continued. First, the BHE should choose J_{21}, to maximize d_2(A, B). This is accomplished by renewing the grants to departments whose outcomes are better than the group average d_1(A, B). Second, the BHE should continue the QUE program if the resulting value of d_2(A, B) is greater than or equal to zero. Formally,

Decision Stage 1: Selection of J_{21}

Let $j \in J_1$. Subject to continuation of QUE, renew the grant to department j if

(19) $N_j(A) \cdot I_j(A) - N_j(B) \cdot I_j(B) - G \geq d_1(A, B)$.

Decision Stage 2: Continuation of QUE

With J_{21} determined in Stage 1, continue the QUE program if

(20) $d_2(A, B) \geq 0$.

Given Assumptions (A1) through (A4) and the availability of complete outcome data, this two-stage decision process provides a complete prescription for BHE evaluation of the QUE program. The prescription employs performance standards at both macro and micro levels. At the macro level expressed in Stage 2, the BHE judges QUE to be acceptable if its outcomes are at least as good as those that would be achieved under the alternative of baseline non-competitive funding. To determine whether QUE meets this macro criterion, the BHE employs performance standards at the micro level expressed in Stage 1. Here the BHE judges each current grant recipient, deciding that performance is acceptable if the grantee's outcomes are at least as good as the average outcome realized by all departments currently receiving grants. Observe that this micro criterion differs from the one discussed in Section 7, in which each grantee's performance is judged relative to its own target values of specified performance indicators.

8.3. Evaluation With Incomplete Outcome Data

Implementation of the two-stage decision process developed in Section 8.2 requires that, in 2002, the BHE know the values of $N_j(A)$, $I_j(A)$, $N_j(B)$, and $I_j(B)$ for each of the departments $j \in J_1$ receiving a QUE grant in 1997. The only one of these quantities that is directly observable is $N_j(A)$, the number of students who actually enroll in department j in the period 1997 - 2002. We shall assume for simplicity that $N_j(B)$, the counterfactual number of students who would enroll if department j were not to receive the QUE grant, equals $N_j(A)$. This done, we may focus attention

on what seem the two central problems of incomplete outcome data faced by the BHE, namely that $I_j(A)$ and $I_j(B)$ are not observable.

THE PROBLEM OF AUXILIARY OUTCOMES: The absence of data on $I_j(A)$, the average life-cycle earnings of students who actually enroll in department j during 1997 - 2002, is a problem of auxiliary outcomes. With the passage of time, the value of $I_j(A)$ in principle becomes observable. In 2002, however, the BHE will only observe the outcomes w1 through w7 and, perhaps, other yet-to-be determined *performance indicators* for department j.

Applying the discussion of Section 3, the BHE should do what it can learn about $E[I(A)|w]$, the mean life-cycle income of students in departments with observed auxiliary outcomes w. Presumably the BHE can collect historical data on the life-cycle earnings and auxiliary outcomes of students enrolled in Indonesian universities. Such data may be used to infer $E[I(A)|w]$ under the equal-conditional-means assumption described in Section 3.1, or at least to bound $E[I(A)|w]$ under a bounded-conditional-means assumption of the type described in Section 3.2.

We must point out that the discussion of Section 3 considered a simpler one-stage evaluation problem than the two-stage problem faced by the BHE in comparing performance-based QUE renewal with baseline non-competitive funding. The discussion of Section 3 would apply fully if the BHE were comparing the indefinite funding version of QUE with baseline non-competitive funding. In that case, performance standards would need to be applied only at the macro level described in Decision Stage 2 above. However the micro level evaluation called from in Decision Stage 1 requires knowledge of each department's value of I(A), not of $E[I(A)|w]$. Using $E[I(A)|w]$ in place of I(A) in equation (19) is strictly correct only if w is a perfect predictor of

I(A), in which case the problem of auxiliary outcomes is completely solved.

THE PROBLEM OF COUNTERFACTUAL OUTCOMES: The absence of data on $I_j(B)$, the average earnings that students in department j would experience in the absence of the department's QUE grant, is a problem of counterfactual outcomes. Department j does have the QUE grant so it is impossible to observe what would have happened otherwise.

There are various assumptions that the BHE might bring to bear in this situation. The BHE might make the *fixed-effects assumption* that, in the absence of the QUE grant, department j's students would experience the same outcomes as the students in this department actually did experience in the pre-QUE period before 1997. This assumption of historical continuity is plausible if there have been no changes in department j's environment over time.

The BHE might make the *comparison-group assumption* that, in the absence of the QUE grant, department j's students would experience the same outcomes as the students in departments similar to department j, but not having QUE grants, actually experience in the period 1997 - 2002. This assumption is plausible if the BHE can credible identify a comparison group for department j__ departments similar to department j except that they do not have QUE grants.

It may be that the fixed-effects and comparison-group assumptions both have some plausibility, as do certain other assumptions, but that no one assumption stands out as clearly correct. In this situation, which we regard as likely in practice, the BHE should bring to bear all of the plausible assumptions, thus yielding a bound on $I_j(B)$.

THE DECISION PROCESS WITH INCOMPLETE OUTCOME DATA: If the BHE, by combining

extensive auxiliary outcome data and historical data with strong assumptions, is able to infer the unobserved values of $l_j(A)$ and $l_j(B)$ for $j \in J_1$, then the BHE can implement the two-stage decision process described in Section 7.1. It may well be, however, that the available data and assumptions only suffice to bound the values of $l_j(A)$ and $l_j(B)$, $j \in J_1$. Then, as described in Sections 3 and 4, the BHE should retreat from the traditional idea of using a single threshold to separate acceptable outcomes from unacceptable ones.

Bounds on $l_j(A)$ and $l_j(B)$ for $j \in J_1$ imply bounds on the group average outcome $d_1(A, B)$. Taken together, the various bounds imply that Decision Stages 1 and 2 cannot be implemented in the simple manner of Section 8.2. Instead, each stage must allow the possibility that outcomes are judged acceptable, unacceptable, or indeterminate.

In the micro-evaluations of Stage 1, the performance of each department $j \in J_1$ might be judged acceptable if its predicted outcomes meet a high acceptance threshold, determined by applying the lower bound on $l_j(A)$, the upper bound on $l_j(B)$, and the upper bound on $d_1(A, B)$. Similarly, department j's performance might be judged unacceptable if its predicted outcomes fail to meet a low nonacceptance threshold, determined by applying the upper bound on $l_j(A)$, the lower bound on $l_j(B)$, and the lower bound on $d_1(A, B)$. If the predicted outcomes lie between the two thresholds, then the acceptability of department j's outcomes is indeterminate and the BHE must use some auxiliary rule to decide whether this department should have its QUE grant renewed.

Bounds on the performances of individual departments aggregate into bounds on the performance of the QUE program as a whole in the macro-evaluation of Stage 2. The mechanics of aggregating the micro-level bounds may be somewhat complex but the underlying idea is

simple enough. The performance-based renewal version of the QUE program should be judged acceptable if the lower bound on its predicted outcomes is sufficiently high and unacceptable if the upper bound on its predicted outcomes is sufficiently low. Otherwise, the overall performance of the program is itself indeterminate and the considerations raised in Section 5 become relevant. Much as a definitive answer to the evaluation problem may be desired, we must emphasize that there is no clear escape from the ambiguity of the situation.

References

Cave, M. and S. Hanney, “Performance Indicators,” In B. Clark and G. Neave (editors), The Encyclopedia of Higher Education, Volume 2, Oxford: Pergamon Press, 1411-1423.

Heckman, J. and R. Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer (editors), Longitudinal Analysis of Labor Market Data, Cambridge: Cambridge University Press.

Maddala, G.S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," American Economic Review Papers and Proceedings, 80, 319-323.

Manski, C. (1994), "The Selection Problem," in C. Sims (editor), Advances in Econometrics, Cambridge: Cambridge University Press.

Manski, C. (1995), Identification Problems in the Social Sciences, Cambridge, Mass.: Harvard University Press.

Manski, C. (1997a), “Monotone Treatment Response,” Econometrica, 65, 1311-1334.

Manski, C. (1997b) "The Mixing Problem in Program Evaluation," Review of Economic Studies, 64, 537-553.

Manski, C. and D. Nagin (1998), "Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism," Sociological Methodology 1998, forthcoming.

Manski, C. and J. Pepper (1997), “Monotone Instrumental Variables,” Department of Economics, Northwestern University.

Pepper, J. (1997), "The Intergenerational Transmission of Welfare Receipt," Department of Economics, University of Virginia.

Spencer, B. (1985), “Optimal Data Quality,” Journal of the American Statistical Association, 80, 564-573.

Spencer, B. and L. Moses (1990), “Needed Data Expenditure for an Ambiguous Decision Problem,” Journal of the American Statistical Association, 85, 1099-1104.