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
and the Institute for Policy Research

John Newman
The World Bank

and

John V. Pepper
Department of Economics, University of Virginia

December 14, 1999

We have benefitted from the opportunity to present this research in seminars at the Nat
Economic Research, Northwestern University, the University of Bristol, and the Universi
We have also benefitted from the comments of three anonymous reviewers.

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

Abstract

The basic idea of program evaluation is both simple and appealing. Program outcomes ar compared to some minimum performance standard or threshold. In practice, however, evalu difficult. Two fundamental problems of outcome measurement must be addressed. The first the *problem of auxiliary outcomes*, is that we do not observe outcome of interest. The s call the *problem of counterfactual outcomes*, is that we do not observe the threshold st paper examines how performance standards should be set and applied in the face of these measuring outcomes. In particular, we consider the problem of evaluating the new World Quality of Undergraduate Education (QUE) program. This competitive block grant program by the program's effects on student outcomes, not by the particular ways in which the g departments use their funds. Our central message is that the proper way to implement st with the prior information that the evaluator can credibly bring to bear to compensate outcome data. An evaluator, confronted with the auxiliary and counterfactual outcomes combine the available data with credible assumptions on treatments and outcomes. Given information, the performance of a program may be deemed acceptable, unacceptable or ind

1. Introduction

The Quality of Undergraduate Education (QUE) program was recently initiated by th government's Board of Higher Education (BHE) as a component of a portfolio of education supported by the World Bank. As part of this program, competitive proposals for block the quality of undergraduate education in specific fields were solicited from academic across the country. In August 1997, 16 five-year grants were awarded with funding leve 400,000 U.S. dollars per year. By agreement between the BHE and the World Bank, the pe QUE program is to be judged by the program's effects on student outcomes, not by the pa which the grantee departments use their funds.

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

Consider the problem of evaluating the QUE program in 2002. The outcome of inter is, broadly speaking, the value to Indonesian society of having high quality university threshold might be set as the outcome that would be expected under the baseline non-com scheme. To cast this idea in conventional economic terms, we might interpret the BHE as determine whether the QUE program maximizes the difference between the expected life-cy earnings of university entrants and the cost of providing their education.

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

The problem of auxiliary outcomes arises whenever considerations of timeliness or infeasible to measure the program outcomes of ultimate interest. Since life-time earnin entrants will not be revealed until many years after the program evaluation in 2002, t observe the outcome of interest. With data on these outcomes unavailable, performance s stated in terms of auxiliary outcomes that can be measured. In fact, the BHE has agree on at least seven auxiliary outcomes, which are officially termed *performance indicator* evaluation problem is to use the available data on early outcomes to set standards, whe real interest is in the lasting effects of the program.

The problem of counterfactual outcomes concerns the alternative serving as the st comparison. Whereas the program being evaluated is operational and so its outcomes are observable in principle, the alternative is not in operation and so its outcomes are co data cannot reveal what would happen to university students under the baseline non-comp system. To appropriately set the threshold defining a performance standard, an evaluato predict what outcomes would occur if the alternative were in operation.

This paper examines how performance standards should be set and applied in the fa problems in measuring outcomes. Our central message is that the proper way to implemen varies with the prior information that the evaluator can credibly bring to bear to comp incomplete outcome data. Of course, the assumptions an evaluator is willing to impose case to case. If this prior information is sufficiently strong, the traditional pract single threshold to separate acceptable outcomes from unacceptable ones is appropriate. having weaker prior information however, should set two thresholds rather than one. Th the program should be deemed acceptable if the observed auxiliary outcomes meet the hig

threshold and unacceptable if they fall below the lower *nonacceptance threshold*.

If the auxiliary outcomes lie between these two thresholds, the performance of th indeterminate. In this case, there is insufficient basis for deciding whether the prog evaluated should be continued or replaced by the alternative. Decisions to continue th replace it are both defensible given the available information. Efforts to obtain more before making a decision may be justified.[1]

We develop these ideas in two stages. Sections 2 and 3 consider the evaluation p generality. Section 2 formalizes basic concepts: treatments, outcomes, programs, and t Section 3 uses these concepts to address the problems of auxiliary outcomes and counter respectively. These sections aim to make general points, so some of the discussion is abstract.

In Sections 4 through 6, we shift from generalities to the specifics involved in new World Bank sponsored Quality of Undergraduate Education (QUE) program. Section 4 d program, which awards competitive five-year block grants to university departments to i quality of their undergraduate curricula. Sections 5 and 6 examine two distinct ways i performance standards will be used. In the short run, the progress of QUE grantees in specified auxiliary outcome targets will be monitored. Then, at the end of the five-ye the QUE program as a whole will be evaluated.

QUE is representative of a large class of programs that use block grants and simi decentralized decision making mechanisms to achieve social objectives. Our examination program has lessons for the evaluation of other block grant programs. In particular, t

---

[1] From Keynes (1921) and Knight (1921) to Walley (1991), decision theorists have long struggled to credibly deal with the ambiguity inherent in program evaluations and decision making. No method of resolving ambiguity (e.g., the maximin rule (Wald, 1950) and Bayesian decision rules (Berger, 1985; Spencer, 1985, Spencer and Mosses, 1990)) can ensure that expected outcomes are maximized. In this paper, we analyze the implications of indeterminacy which arises from two fundamental identification problems; the auxiliary and counterfactual outcome problems. While these two concerns are central, they certainly do not exhaust the set of possible causes of ambiguity. For a general discussion, see Manski (1999).

Section 6 shows the need for integrated micro evaluation of particular grantees and mac the program as a whole.[2]

Section 7 concludes by considering what a planner might do when the available inf an indeterminate finding about the performance of the program being evaluated.

2. Concepts of Formal Evaluation

The usual formalization of a program evaluation assumes that each member j of a p receives one of several mutually exclusive and exhaustive *treatments*. Each member of t experiences a scalar outcome-of-interest that may depend on the treatment received. Th treatments will be numbered t = 1,...,T. The outcomes associated with these treatments 1,...,T.[3]

The treatment that a person receives depends on the set of treatments available t and on the person's choice of a treatment from this set. Social programs help determin available treatments and thus influence the treatments that people receive. It will su two programs. One of these is the operational program being evaluated, labeled program other is the alternative with which the operational program is to be compared, labeled z_{jA} 0 T indicate the treatment that person j actually receives under program A, and let the treatment that this person would receive under program B. Then the outcomes this p experience under program A and would experience under program B are $y(z_{jA})$ and $y(z_{jB})$ res

The objective of the evaluation is to determine which yields the better outcomes,

---

[2] A similar distinction is made in the literature on evaluating personnel, where both a particular job as well as the individuals who hold the job may be evaluated (Lazear, 1995, Chapter

[3] 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.

B. The usual practice is to compare programs in terms of their mean outcomes across th conventional economic terms, we assume that the planner wants to maximize a utilitarian function.[4] Let $E[y(z_A)]$ and $E[y(z_B)]$ denote the mean outcomes under programs A and B. T

(1) $*(A, B) \;/\; E[y(z_A)] - E[y(z_B)]$

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

Implementing the performance standard is straightforward if the evaluator observe $y(z_A)$ and $y(z_B)$ of the members of the population, or at least those of random samples of Then the evaluator may learn the mean outcomes $E[y(z_A)]$ and $E[y(z_B)]$ and determine wheth exceeds the latter.

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

To illustrate these concepts, consider the problem of evaluating the QUE program. the set T of possible treatments may index different types of funding mechanisms; univ might vary by the allocation scheme (competitive versus non-competitive), the levels an assistance, the restrictions on inputs and outputs, and the criteria for future funding be some version of the competitive block grant funding introduced by the QUE program, a be the baseline non-competitive allocation system. The outcome-of-interest y may measu

[4] Of course, one might consider evaluating other features of the distribution of outcomes. See, for example, Manski (1995; 1997b) and Heckman, Smith and Clements (1997).

earnings. The observed auxiliary outcomes w might measure cognitive ability in year 20

The performance of the QUE program may be deemed acceptable if mean present disco life-time earnings is higher under the QUE program than under the baseline alternative. auxiliary outcomes is that only cognitive status is observed under the operational prog of counterfactual outcomes is that no outcome measurements at all are possible under th alternative to program A.

3. Problems of Outcome Measurement

In this section, we provide a general introduction to the problems of auxiliary a counterfactual outcomes. In Section 3.1 we investigate the problem of auxiliary outcom possible solutions, both of which rely on historical data. Abstracting from the probl counterfactual outcomes, we suppose that the evaluator sets a threshold that the mean o A must meet to be deemed acceptable. Then, in Section 3.2 we examine how this threshol In particular, we describe the problem of counterfactual outcomes, and review some of t solutions to this problem.

3.1 The Problem of Auxiliary Outcomes

Abstracting from the problem of counterfactual outcomes, let c denote the thresho evaluator. Then the criterion for judging the performance of program A is this:

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

[5] In Section 4, we provide additional details on the auxiliary outcome measures collected by the BHE for the QUE program and the alternative.

The evaluator observes only the auxiliary outcomes $w(z_A)$ of the population and not outcomes-of-interest $y(z_A)$. The problem is to use the data on auxiliary outcomes to lea $E[y(z_A)]$. For convenience, suppose that the auxiliary outcome vector w can take S possi numbered s = 1,...,S. Use the law of iterated expectations to write

$$(3) \quad E[y(z_A)] = \sum_{s=1}^{S} E[y(z_A)*w(z_A) = s]\cdot P[w(z_A) = s].$$

Here $E[y(z_A)*w(z_A) = s]$ is the mean value of the outcome-of-interest among the people wh s of the auxiliary outcome, and $P[w(z_A) = s]$ is the fraction of the population who reali

The practical problem, of course, is that auxiliary outcome data alone do not rev conditional means $E[y(z_A)*w(z_A) = s]$, s = 1,...,S. Thus implementation of criterion (2) only if the evaluator can bring to bear other information that reveal $E[y(z_A)*w(z_A) = s]$

The possibilities explored here all assume the existence of some historical perio were collected on both the auxiliary outcomes w and the outcome-of-interest y. These h may pertain to an environment that is different in some respects from that of program A the BHE may have access to data collected under different economic conditions, differen regimes or even different countries. The data may nevertheless be used to inform the e

---

[6] 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

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

Here $f(\cdot)$ is the chosen function and d is the specified threshold. It follows from the law of iterated expectations that $E\{f[w(z_A)]\} = E[y(z_A)]$ if $f(\cdot)$ is chosen to be the function $f(s) = E[y(z_A)*w(z_A) = s]$. With this choice of $f(\cdot)$ and with d set equal to c, performance criteria (2) and (*) are equivalent. Application of criterion (*) with other choices of f(.) and d may lead to distorted conclusions about the acceptability of the program being evaluated.

program A, provided that the historical period for which (w, y) data are available shar features with the environment under program A. Sections 3.1.1 and 3.1.2 make this expl

3.1.1 The Equal-Conditional-Means Assumption

Let the observable historical distribution of (w, y) be denoted $P_H(w, y)$. Assume value of w realized by a positive fraction of the population under program A was also r positive fraction of the population in the historical period. Now assume that, for each $P[w(z_A) = s] > 0$, the conditional mean outcome $y(z_A)$ under program A equals the conditio historical outcome y. That is, the historical data yield an unbiased conditional forec

$$(4)\quad E[y(z_A)*w(z_A) = s] \;=\; E_H(y*w = s).$$

This *equal-conditional-means assumption* and the law of iterated expectations (3) yield

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

By assumption (4), the historical data on (w, y) reveal $E_H(y*w = s)$ whenever $P[w(z_A) = s$ auxiliary outcome data on program A reveal $P[w(z_A) = s]$ for all values of s. Hence the use the right side of equation (5) to learn $E[y(z_A)]$ and so judge the performance of pr

The credibility of the equal-conditional-means assumption must be assessed on a c basis. The identity of the measured auxiliary outcomes may be critical, the assumption for some specifications of the auxiliary outcomes but not for others. Often, planners

outcome measures which are arguably related to both the intervention and the outcome of There are, however, no general criteria for ensuring the credibility of the assumption.

For the QUE program, the BHE has agreed to collect data on at least seven auxilia many of which measure cognitive skills. The equal-conditional-means assumption states unobserved mean life-cycle earnings of persons who have measured cognitive skills s in observed historical mean life-cycle earnings among persons who had measured cognitive s a reasonable assumption? It is if one thinks that the QUE program influences earnings effect on measured cognitive skills, but not otherwise. The assumption is less reasona that the program may influence earnings through a process that does not entirely manife measured cognitive skills.

3.1.2 Bounded Conditional-Means Assumptions

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

A flexible way to weaken the equal-conditional-means assumption is to use knowled s) to bound $E[y(z_A)*w(z_A) = s]$. Supposing that y takes positive values, a particularly *conditional-means assumption* is

$$(6) \quad \alpha \cdot E_H(y*w = s) \;\#\; E[y(z_A)*w(z_A) = s] \;\#\; \beta \cdot E_H(y*w = s),$$

Here " and ß are constants such that $0 \# \alpha \# \beta \# 4$. These constants, specified by the

[7] Treatments and covariates can serve as auxiliary outcomes. To formalize treatment as an auxiliary outcome, we simply define $w(z_A) / z_A$. 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.

express the strength of the association that the evaluator feels comfortable asserting

s) and $E[y(z_A)*w(z_A) = s]$. If " = ß = 1, we have the equal-conditional-means assumption

ß = 4, measurement of $E_H(y*w = s)$ reveals nothing about $E[y(z_A)*w(z_A) = s]$.

Assumption (6) and the law of iterated expectations (3) imply this bound on E[y(z

$$(7)\quad \alpha \sum_{s=1}^{S} E_H(y*w = s)\cdot P[w(z_A) = s] \;\#\; E[y(z_A)] \;\#\; \beta \sum_{s=1}^{S} E_H(y*w = s)\cdot P[w(z_A) = s].$$

If the lower bound on $E[y(z_A)]$ meets the threshold c, the evaluator can conclude that th

program A is acceptable. If the upper bound on $E[y(z_A)]$ is less than c, he can conclude

program's performance is unacceptable.

Many values of the constants " and ß will lead to a definitive evaluation. Let

$$(8)\quad \gamma = c \,/\, \sum_{s=1}^{S} E_H(y*w = s)\cdot P[w(z_A) = s].$$

From (7) we see that if " $ (the program should be accepted and if ß # (the program's

unacceptable. An evaluator need only know that " or ß satisfy one of these inequalities

efficacy of the program. Otherwise, the status of program A is indeterminate given the

and prior information.[8]

---

[8] As we do here, the literature on sensitivity analysis (see, for example, Cornfield et. al., (1959), Rosenbaum and Rubin (1983) and Rosenbaum (1995, Chapter 4)) examines the implications of varying certain unknown constants or parameters within some class of models. This literature, however, does not address the evaluator's problem of making decisions when the findings are ambiguous. That is, if" < (< $, the performance of A is indeterminate.

Consider the QUE program. There are many reasons why an evaluator may not be wil equal-conditional-means assumption. It may be that schooling norms have changed betwee period and the present, with consequent changes in the association between schooling an earnings. Or it may be that the very act of evaluating the QUE program has incentive e the association between cognitive skills and earnings. Administrators of the program, measured cognitive skills will be used to evaluate program performance, may choose to e schooling that have measurable effects on cognitive skills rather than ones whose effec measurable later on. This is particularly true for manipulable indicators such as gra

Concerned with these and other possibilities, the evaluator may find a bounded-co assumption to be more credible. If, for instance, performance indicators might be infl Hawthorne effect, the evaluator may want to assume (6) with " = 0 and ß = 1. That is, to assume that the unobserved mean earnings among persons with cognitive status s in 20 than the historical mean earnings with cognitive status s. This assumption may suffice the QUE program is unacceptable.

3.2 The Problem of Counterfactual Outcomes

Discussions of performance standards often exhibit considerable lack of clarity o threshold separating acceptable from unacceptable performance should be set and what ac taken if performance is deemed unacceptable. Much of the difficulty that evaluators ha thresholds and actions stems from the problem of counterfactual outcomes. In principle should be set equal to a mean outcome level known to be achievable by an alternative fe and this alternative should replace the operational program if the threshold is not met outcomes that would occur under counterfactual alternatives are not observable. Hence, from the problem of auxiliary outcomes, evaluators inevitably find it hard to specify w acceptable program performance.

The rich econometric literature on the analysis of treatment effects teaches that unique resolution of the problem of counterfactual outcomes. The conclusions that can the outcomes of counterfactual programs depend critically on what historical data are a prior information the evaluator can credibly bring to bear.

The dominant concern of the econometric literature has been to predict the outcom treatment programs — ones giving the same treatment to all members of the population — available historical data pertain to an environment in which treatment varies across th this context, the problem of counterfactual outcomes is known as the *selection problem*. selection problem show that if historical data on the outcome of interest are combined strong assumptions, the counterfactual mean outcome $E[y(z_B)]$ may be identified, implying threshold for judging the performance of program A. In practice, the most common assum treatments are statistically independent of outcomes in the historical data, as they wo classical randomized experiment. An alternative route to identification is to assert a latent variable model jointly describing how treatments are selected and outcomes deter alternative is to assume that treatment effects are constant across the population and some covariate, termed an *instrumental variable*, that is independent of outcomes but no See Björklund and Moffitt (1987), Friedlander, Greenberg and Robins (1997), Heckman an Heckman and Hotz (1989), Heckman and Robb (1985), Maddala (1983), and Manski (1989, 19 of the literature.

Concern with the validity of the strong assumptions needed to identify treatment to the recent development of a literature imposing weak assumptions that yield bounds o

---

[9] Whereas a variable v was originally called an *instrumental variable* if v has zero covariance with a residual, from the response function, the modern usage of the term has broadened to embrace assumptions that specified functions of v and, are orthogonal. Hence it is now necessary to specify the type of IV assumption one has in mind. Mean independence, quantile independence, and statistical independence assumptions (or the orthogonality conditions that these assumptions yield) have all been prominent in the literature. See Manski (1988) pp. 25-26 and Section 6.1 for discussion of the history and exposition of the variety of modern IV assumptions.

counterfactual mean outcome $E[y(z_B)]$. The starting point is to ask what can be learned from the historical data if no assumptions at all are made about the process determinin selection and outcomes. The result is a "no-assumptions" bound on $E[y(z_B)]$. From this evaluator may impose weak assumptions that have identifying power in the sense that the bounds. One set of results illuminates the identifying power of instrumental variable imposed alone, treatment effects not being assumed to be constant across the population Pearl (1997), Hotz, Mullins and Sanders (1997), Manski (1990, 1994), Manski and Pepper (1989), and Robins and Greenland (1996). Another set of results shows the identifying assumptions about the treatment selection process when nothing is known about the proce outcomes. For example, one may assume that each member of the population was assigned yielding the better outcome for that person. See Manski (1994, 1995), and Manski and N another set of results shows the identifying power of assumptions about the process det when nothing is known about the treatment selection process. For example, one may assu response is monotone, in the sense that the outcome of one treatment is always at least outcome of the other. See Manski (1995, 1997a) and Pepper (2000).

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

(9) Program A is acceptable if $d_0 - c_1 \$ 0$ 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 s

but rather permits treatment to vary across the population (see Manski, 1997b and Peppe general point remains that application of a conventional performance standard with a si separate acceptable from unacceptable outcomes is appropriate only if the evaluator can sufficiently strong data and assumptions. In other settings, the performance of progra possible states: acceptable, unacceptable, or indeterminate.

Consider the QUE program. Suppose that the alternative is the non-competitive al method. How might the evaluator predict what the outcome (e.g., life-time earnings) w baseline alternative? The BHE might make the *fixed-effects assumption* that, in the abse grant, students in a department would experience the same outcomes as the students in t actually did experience in the pre-QUE period before 1997. This assumption of historic plausible if there have been no changes in the department's environment over time. Alte BHE might make the *comparison-group assumption* that, in the absence of the QUE grant, a students would experience the same outcomes as the students in similar departments that grants actually experience in the period 1997 - 2002. This assumption is plausible if credibly identify a comparison group — similar departments except that they do not have

It may be that the fixed-effects and comparison-group assumptions both have some as do certain other assumptions, but that no one assumption stands out as clearly corre situation, which we regard as likely in practice, the BHE should bring to bear all of t assumptions, thus yielding a bound on $E[y(z_B)]$. If there is concern about the credibil assumptions but not others, the BHE might bound disagreements about the evaluation by e expected counterfactual outcome under a sequence of progressively stronger assumptions. are added, the bound on $E[y(z_B)]$ may narrow but may also be less credible.

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

In the remainder of the paper, we examine some of the specific issues involved in QUE program. In this section, we describe the established features of the program and important unresolved questions. With this as background, Sections 5 and 6 examine the monitoring and evaluation problems associated with this program.

4.1. Basic Description of the QUE Program

With the assistance of the World Bank, the Government of Indonesia has embarked u to improve the quality of education through the greater use of incentives in budgetary decisions. The general approach is to allocate some fraction of the development budget competitively awarded performance based grants. Under the old regime the allocation de competitive.

The Quality of Undergraduate Education (QUE) program was recently initiated by th government's Board of Higher Education (BHE) as a component of this effort. All acade in public universities were invited to submit proposals for block grants to improve the undergraduate education they provide. The first round of the competition for these gran out in 1997. Pre-proposals were received from 317 departments, 45 of which were invite proposals. In August 1997, 16 five-year grants were awarded with funding levels averag dollars per year. The grants are meant to provide new funding to the recipient departme supplementing their regular budgets.

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

The outcome of interest to the BHE is, broadly speaking, the value to Indonesian

having high quality university graduates, both of the departments that receive QUE gran which do not. In practice, the BHE and the World Bank have agreed that the program wil monitored and then evaluated using data to be collected on at least these seven auxilia 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 departm en route to graduation.

w5. Time to Employment - average length of time that students take to secure employme 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.

4.2. Monitoring and Evaluation

The BHE and the World Bank have agreed to monitor the auxiliary outcomes experien current grantees during 1997 - 2002 and then to evaluate the QUE program in 2002 at the year grant period. Monitoring means that the BHE will assess the performance of grante auxiliary outcome targets agreed upon by the grantees and the BHE. If a department's pe meeting its targets is deemed to be inadequate, the BHE may take limited corrective act the particulars of the case. It may, for example, provide technical assistance to a de inexperienced personnel. It may also delay the release or reduce the size of a payment

presumption, however, is that barring an incident of gross negligence or fraud, the gra continue to receive its annual funding throughout the five-year grant period. See Sect discussion.

Although monitoring has some of the character of an evaluation, the BHE usefully distinction between monitoring and the evaluation of the QUE program that will take pla the BHE must decide whether to continue the QUE program or to replace it with an altern point, we need to confront the fact that the QUE program is a work in progress rather t articulated funding program. The BHE and World Bank have not yet stated what it would the QUE program after 2002.

In Section 6, we select one version of the QUE program and one alternative for fu In particular, we suppose that in 2002 the BHE will interpret the QUE program to use a *grant renewal design*, such that grantees would have their grants renewed for an additio period if their measured auxiliary outcomes are judged to be acceptable, but not renewe auxiliary outcomes are judged non-acceptable. Every five years a new grant competition re-allocate those QUE funds that become available when some grantees do not have their We suppose that the relevant alternative is the baseline non-competitive funding system

---

[10] There are numerous other reasonable possibilities. Here are two other schemes which also 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 d
- 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.

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 considered variable.

1997.[11] In the notation of Sections 2 and 3, the performance based renewal QUE is progra baseline alternative is program B.

Performance-based renewal is a particularly interesting interpretation of QUE bec encompasses indefinite funding and open re-competition as special cases. If the thresh renewal is set so low that all existing grants are renewed, performance-based renewal i indefinite funding for the sixteen departments awarded grants in 1997. If the threshol that no existing grants are renewed, performance-based renewal is equivalent to open re

5. Monitoring The QUE Grantees

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

Each of the sixteen QUE grants specified target changes in performance indicators to be achieved 2.5 years and five years after grant initiation. These midterm and fina vary across the departments receiving grants, were established by negotiation between t departments. These targets are conservative so that a non-positive report would indica some type of corrective action or additional supervision. The BHE has yet to determine the targets to assess departments' performance and the actions it will take if the targ

[11] There are numerous other alternatives. In fact, 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 performance-based renewal design would provide an alternative to QUE. Other alternatives might retain the competitive funding idea of QUE but alter the number of grants or the award per grantee.

We consider the monitoring question here, restricting attention to the midterm targets. evident in Section 6, evaluation of the QUE program at the end of five years involves d considerations.

Consider the situation of one QUE grantee, the Department of Civil Engineering a University of Indonesia. Table 1 displays the target standards (T) as well as the basel the performance indicators of this department. In 2000, 2.5 years after the grants wer evaluator will observe the realized auxiliary outcomes (R). Let w_{Tj} and w_{Rj} denote this midterm target and realized values of the performance indicators w1 through w5. The di Section 3 suggests that the BHE should view these performance indicators as auxiliary o be used to predict the outcome-of-interest, namely the value to Indonesian society of h quality university graduates.

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

(10) $y_j(A) \;/\; N_j(A)\cdot I_j(A) - C_j(A)$.

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-int conditional on the performance indicators taking the values w_{Tj} and w_{Rj} respectively. Th might use this criterion to monitor the midterm performance of department j:

(11) Midterm Performance is acceptable if $E[y(z_A)*w(z_A) = w_{Rj}] \;\$\; E[y(z_A)*w(z_A) = w_{Tj}]$.

To implement this criterion as stated requires that the BHE know the conditional $E[y(z_A)*w(z_A) = w_{Rj}]$ and $E[y(z_A)*w(z_A) = w_{Tj}]$. As discussed in Section 3.1, these quantit knowable if historical data on (w, y) are available and if the BHE is able to credibly conditional-means assumption. Under weaker bounded-conditional-means assumptions of th in Section 3.2, the BHE can conclude that midterm performance is acceptable if the lowe $E[y(z_A)*w(z_A) = w_{Rj}]$ is greater than or equal to the upper bound on $E[y(z_A)*w(z_A) = w_{Tj}]$. the upper bound on $E[y(z_A)*w(z_A) = w_{Rj}]$ lies below the lower bound on $E[y(z_A)*w(z_A) = w_{Tj}]$ conclude that midterm performance is unacceptable. If neither of these conditions hold performance is indeterminate.

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

6. Evaluation of QUE: Comparison of Performance-Based Renewal and Non-competitive Fundi

In this section we examine the BHE's decision problem in 2002, at the end of the period. In particular, we consider how the BHE might compare performance-based QUE gra (Program A) with the alternative of baseline non-competitive funding (Program B). Our intended to develop some important points, but not to cover all of the difficult issues need to consider. Hence we shall make some simplifying assumptions. These are

(A1) In 2002, the BHE is only concerned with the next round of five-year QUE grants. commit itself to the QUE program beyond 2007 nor otherwise consider how departmen funded beyond that date.

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

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

(A4) Continuation of the QUE program from 2002 to 2007 only affects the sixteen depart receive grants in 2002. Departments that do not receive grants at that time have funding and student outcomes under the performance-based QUE grant renewal and th of the baseline non-competitive funding program.

Assumptions (A1) through (A4) greatly simplify the BHE's evaluation problem. We however, that these assumptions should not be taken lightly. The BHE should, in princi the next round of grants and so (A1) may not hold. University administrations may seek funding to substitute for departmental baseline funding, thereby violating (A2). Moreo give QUE grants of different sizes to different departments, also violating (A2).[12] Ass

[12] The analysis can be redone with different assumptions about the degree of substitution and the size of the grant for each department. In particular, basic evaluation methodology applies as long as the

plausible if relevant aspects of the higher education environment — the characteristics students, the mix of departments applying for QUE grants, the BHE's decision process in the state of the Indonesian labor market, etc. — do not change between 1997 and 2002. in the environment may occur and make this assumption suspect. For example, the mix of 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 that do not receive grants. QUE funding may allow the departments that receive grants effectively for students, thus altering the student bodies at non-recipient departments may allow students in departments that receive grants to compete more effectively for a of jobs after graduation, thus altering the job prospects of the graduates of other dep Moreover, the process of writing proposals for QUE funding may lead departments to crit 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 and then develop the implications of Assumptions (A1) through (A4) in Sections 6.2 and 6.2, we abstract from the problems of auxiliary and counterfactual outcomes and conside should act if it were somehow to have complete outcome data. In Section 6.3, we consid should act given the outcome data that are likely to be available.

6.1. General Features of the Evaluation Problem

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

---

evaluator knows the net costs of the program for each department.

$$(12)\quad E[y_j(F)] \;/\; \frac{1}{*J*} \sum_{j\,0\,J} N_j(F)\cdot I_j(F) - C_j(F),$$

where *J* is the number of university departments. We shall interpret the BHE as wantin funding program that maximizes $E[y_j(F)]$.

By assumption, the feasible options are the performance-based renewal version of and the baseline noncompetitive funding mechanism. In the notation of Sections 2 and 3 A and the baseline alternative is program B. Applying equation (12), we suppose that th QUE to have acceptable outcomes if

$$(13)\quad \sum_{j\,0\,J} [N_j(A)\cdot I_j(A) - N_j(B)\cdot I_j(B)] - [C_j(A) - C_j(B)] \;\$\; 0.$$

6.2. Evaluation With Complete Outcome Data

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

$$(14)\quad *_1(A, B) \;/\; \frac{1}{16} \sum_{j\,0\,J_1} [N_j(A)\cdot I_j(A) - N_j(B)\cdot I_j(B) - G]$$

be the average difference between the outcomes that these departments realize and those have experienced if they had not received QUE grants. Recall that $G = C_j(A) - C_j(B)$ is t QUE grant, which is approximately 400,000 U.S. dollars per year.

Let J_2 denote a hypothetical set of sixteen departments that would receive grants

is continued. Some of these, denoted J_{21}, would be members of J_1 that have their grants remaining 16 - $*J_{21}*$ members of J_2 would be new grant recipients. Assumption (A1) throug that the average difference between the outcomes that the departments in J_2 would reali grants and those that they would experience in the absence of the grants is

$$(15)\quad *_2(A, B) \;/\; \frac{1}{16} \sum_{j\ 0\ J_2} [N_j(A)\cdot I_j(A) - N_j(B)\cdot I_j(B) - G]$$

$$= \frac{1}{16} \{ [16 - *J_{21}*]\cdot *_1(A, B) + \sum_{j\ 0\ J_{21}} [N_j(A)\cdot I_j(A) - N_j(B)\cdot I_j(B) - G]\}.$$

The term [16 - $*J_{21}*$]·$*_1$(A, B) on the right side of (18) reflects the second part of Assu which states that students in departments that receive new QUE grants in 2002 realize t 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 recei Hence, in 2002, the BHE should use a two-stage process to decide which department shoul grants renewed and whether the QUE program should be continued. First, the BHE should maximize $*_2$(A, B). This is accomplished by renewing the grants to departments whose out than the group average $*_1$(A, B). Second, the BHE should continue the QUE program if the of $*_2$(A, B) is greater than or equal to zero. Formally,

Decision Stage 1: Selection of J_{21}

Let j 0 J_1. Subject to continuation of QUE, renew the grant to department j if

(16) $N_j(A)\cdot I_j(A) - N_j(B)\cdot I_j(B) - G \$ *_1(A, B)$.

Decision Stage 2: Continuation of QUE

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

(17) $*_2(A, B) \$ 0$.

Given Assumptions (A1) through (A4) and the availability of complete outcome data stage decision process provides a complete prescription for BHE evaluation of the QUE p prescription employs performance standards at both macro and micro levels. At the macr in Stage 2, the BHE judges QUE to be acceptable if its outcomes are at least as good as be achieved under the alternative of baseline non-competitive funding. To determine wh this macro criterion, the BHE employs performance standards at the micro level expresse Here the BHE judges each current grant recipient, deciding that performance is acceptab grantee’s outcomes are at least as good as the average outcome realized by all departme receiving grants. Observe that this micro criterion differs from the one discussed in which each grantee’s performance is judged relative to its own target values of specifi indicators.

6.3. Evaluation With Incomplete Outcome Data

Implementation of the two-stage decision process developed in Section 6.2 require 2002, the BHE know the values of $N_j(A)$, $I_j(A)$, $N_j(B)$, and $I_j(B)$ for each of the departmer receiving a QUE grant in 1997. The only one of these quantities that is directly obser

the number of students who actually enroll in department j in the period 1997 - 2002.
for simplicity that $N_j(B)$, the counterfactual number of students who would enroll if dep
not to receive the QUE grant, equals $N_j(A)$. This done, we may focus attention on what s
central problems of incomplete outcome data faced by the BHE, namely that $I_j(A)$ and $I_j(B$
observable.

The absence of data on $I_j(A)$, the average life-cycle earnings of students who actu
department j during 1997 - 2002, is a problem of auxiliary outcomes. With the passage
value of $I_j(A)$ in principle becomes observable. In 2002, however, the BHE will only obs
outcomes w1 through w7 and, perhaps, other yet-to-be determined *performance indicators*
The absence of data on $I_j(B)$, the average earnings that students in department j would e
absence of the department's QUE grant, is a problem of counterfactual outcomes. Depart
the QUE grant so it is impossible to observe what would have happened otherwise.

If the BHE, by combining extensive auxiliary outcome data and historical data wit
assumptions, is able to infer the unobserved values of $I_j(A)$ and $I_j(B)$ for $j \in J_1$, then
decision process described in Section 6.2 can be implemented. It may well be, however,
available data and assumptions only suffice to bound the values of $I_j(A)$ and $I_j(B)$, $j \in$
described in Section 3, the BHE should retreat from the traditional idea of using a sin
separate acceptable outcomes from unacceptable ones.

Bounds on $I_j(A)$ and $I_j(B)$ for $j \in J_1$ imply bounds on the group average outcome dif
B). Taken together, the various bounds imply that Decision Stages 1 and 2 cannot be imp
simple manner of Section 6.2. Instead, each stage must allow the possibility that outco
acceptable, unacceptable, or indeterminate.

In the micro-evaluations of Stage 1, the performance of each department $j \in J_1$ mig
acceptable if its predicted outcomes meet a high acceptance threshold, determined by ap
bound on $I_j(A)$, the upper bound on $I_j(B)$, and the upper bound on $*_1(A, B)$. Similarly, d
performance might be judged unacceptable if its predicted outcomes fail to meet a low n

threshold, determined by applying the upper bound on $I_j(A)$, the lower bound on $I_j(B)$, an
bound on $*_1(A, B)$. If the predicted outcomes lie between the two thresholds, then the a
department j's outcomes is indeterminate and the BHE must use some auxiliary rule to de
department should have its QUE grant renewed.

Bounds on the performances of individual departments aggregate into bounds on the
the QUE program as a whole in the macro-evaluation of Stage 2. The mechanics of aggreg
level bounds may be somewhat complex but the underlying idea is simple enough. The pe
renewal version of the QUE program should be judged acceptable if the lower bound on it
outcomes is sufficiently high and unacceptable if the upper bound on its predicted outc
sufficiently low. Otherwise, the overall performance of the program is indeterminate.
definitive answer to the evaluation problem may be desired, we must emphasize that ther
from the ambiguity of the situation.

We must also point out that the discussion of Section 3 considered a simpler one-
evaluation problem than the two-stage problem faced by the BHE in comparing performance
renewal with baseline non-competitive funding. The discussion of Section 3 would apply
were comparing the indefinite funding version of QUE with baseline non-competitive fund
case, performance standards would need to be applied only at the macro level described
2 above. However the micro level evaluation called from in Decision Stage 1 requires k
average outcome, I(A) for each department, not of the average outcome across all depart
Using E[I(A)*w] in place of I(A) in equation (16) is correct to the extent that w is a g
I(A), in which case the problem of auxiliary outcomes is solved.

7. Conclusions: Should Indeterminacy be Tolerated or Resolved?

In 2002 the BHE will begin the difficult task of evaluating the QUE program. Whi

details of this evaluation remain uncertain, there are general lessons to be drawn. On has maintained a useful distinction between monitoring (Section 5) where outcomes under are compared to prespecified outcome targets, and evaluation (Section 6) where outcomes compared to the outcomes that would have occurred under an alternative funding scheme (Another is that evaluation of block grant programs like QUE requires integrated micro e individual grantees and macro evaluation of the funding mechanism.

Regardless of the specific evaluation criteria to be applied, planners must confr the outcomes of interest are not observed. The outcomes under program A -- mean life-c under the QUE program -- may not be observed until many years after the evaluation. The outcomes under program B -- mean life-cycle earnings under the baseline alternative -- observed. An evaluator, confronted with the auxiliary and counterfactual outcomes probl combine the available data with credible assumptions on treatments and outcomes. Given information, the performance of a program may be deemed acceptable, unacceptable or ind

Suppose that an evaluation yields an indeterminate finding about the program’s ac What then? There are potentially two ways to resolve the ambiguity. One can always im 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. collection can be costly and time-consuming, imposing assumptions requires only a leap problem, of course, is that strong assumptions may be inaccurate and yield flawed concl an evaluator personally considers an assumption to be plausible, he must be concerned a credibility of his findings to policymakers and the public. These may be a diverse gro members may not share the evaluator's beliefs about what are and are not plausible assu evaluator must keep in mind that the weaker the assumptions imposed, the more widely cr reported findings. Let us face the fact that imposing assumptions that are not credibl eliminate the ambiguity in the evaluation problem.

If stronger assumptions are not imposed, the only way to resolve an indeterminate collect richer outcome data. We have examined the evaluation problem given specified da saying anything about how these data came to be available. In practice, evaluators pla determining what outcome data should be collected. Evaluators may be able to influence historical data on auxiliary outcomes and outcomes of interest, thus enabling applicati developed in Sections 3.1. Evaluators may also be able to influence the collection of program A, thus reducing the distance between the available auxiliary outcomes and the interest. If it is feasible to collect richer outcome data, either historical data or o program A, then the evaluator must decide whether the benefits of new data collection e After all, new data cannot resolve the problem of counterfactual outcomes. Even if the program A are known with certainty, the findings may remain indeterminate; outcomes und lie within the bounded threshold of program B.

It is important to stress that an indeterminate finding does not imply that the p be unwilling or unable to make decisions. It only implies that the planner should not c decisions are optimal.

References

Balke, A. and J. Pearl (1997). “Bounds on Treatment Effects from Studies With Imperfect Journal of the American Statistical Association, 92, 1171-1177.

Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. Springer-Verlag,

Björklund, A. and R. Moffitt (1987). “Estimation of Wage Gains and Welfare Gains in Sel Models.” Review of Economics and Statistics, 69, 42-49.

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

Cornfield, J., W. Haenszel, E. Hammond, A. Lilienfeld, M. Shimken, and E. Wynder (1959) Lung Cancer: Recent Evidence and a Discussion of Some Questions,” Journal of the Natio Institute, 22, 173-203.

Friedlander, D., D.H. Greenberg and P.K. Robins (1997), “Evaluating Government Training Economically Disadvantaged,” The Journal of Economic Literature. XXV(4), 1809-1855.

Heckman, J. and B. Honore (1990), “The Empirical Content of the Roy Model,” Econometric

Heckman, J. and J. Hotz (1989), “Choosing Among Alternative Nonexperimental Methods for Impact of Social Programs: The Case of Manpower Training,” Journal of the American Stat Association, 84, 862-874.

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

Heckman, J., J. Smith, and N. Clements (1997), “Making the Most Out of Programme Evalua Experiments: Accounting for Heterogeneity in Programme Impacts,” Review of Economic Stu 536.

Hotz, V.J., C. Mullins and S. Sanders (1997). “ Bounding Causal Effects Using Data from Natural Experiment: Analyzing the Effects of Teenage Childbearing,” Review of Economic 604.

Keynes, J. (1921). A Treatise on Probability. MacMillian.

Knight, F. (1921). Risk, Uncertainty and Profit. Houghton-Mifflin, Boston.

Lazear, E.P. (1995). Personnel Economics. MIT Press. Cambridge, MA.

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

Manski, C.(1988), Analog Estimation Methods in Econometrics, London: Chapman & Hall.

Manski, C. (1989), "Anatomy of the Selection Problem", Journal of Human Resources, 24,

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

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

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

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

Manski, C. (1997b), "The Mixing Problem in Program Evaluation,"Review of Economic Studi

Manski, C. (1999), "Identification Problems and Decisions Under Ambiguity: Empirical An Treatment Response and Normative Analysis of Treatment Choice," Journal of Econometrics

Manski, C. and D. Nagin (1998), "Bounding Disagreements About Treatment Effects: A Case Sentencing and Recidivism," Sociological Methodology 1998, Vol. 28, 99-137.

Manski, C. and J. Pepper (2000), "Monotone Instrumental Variables with An Application t Schooling," Econometrica, forthcoming.

Pepper, J. (1999), "What Do Welfare-to-Work Demonstrations Reveal to Welfare Reformers? for Poverty Research Working Paper, #2, August.

Pepper, J. (2000), "The Intergenerational Transmission of Welfare Receipt: A Nonparamet Analysis," The Review of Economics and Statistics, forthcoming.

Robins, J. (1989), "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Approach to Causal Inference in Longitudinal Studies," in Sechrest, L., H. Freeman, and Health Service Research Methodology: A Focus on AIDS, NCHSR, U.S. Public Health Service

Robins, J. and S. Greenland (1996), "Comment on Angrist, Imbens, and Rubin's 'Identific Effects Using Instrumental Variables'," Journal of the American Statistical Association

Rosenbaum, P.R. (1995). Observational Studies. Springer Series in Statistics, Springer-

Rosenbaum, P.R. and D.B. Rubin (1983). Assessing the Sensitivity to an Unobserved Bina an Observational Study With a Binary Outcome. Journal of the Royal Statistical Society 212-218.

Spencer, B. (1985), "Optimal Data Quality," Journal of the American Statistical Associa 573.

Spencer, B. and L. Moses (1990), "Needed Data Expenditure for an Ambiguous Decision Pro the American Statistical Association, 85, 1099-1104.

Wald, A. (1950). Statistical Decision Functions. Wiley, New York.

Walley. P. (1991). Statistical Reasoning with Imprecise Probabilities. Chapman & Hall,

Table 1: Performance Indicators for the Department of Civil Engineering, University of Indonesia

Performance Indicators	**Baseline**	**Midterm**	**Final**
1. NEE Score	750	770	790
2. GPA	2.57	2.65	3.00
3. TOEFL Score	450	475	495
4. Time to Degree (years)	6.30	5.90	5.00
5. Time to Employment (mo)	1.5	1.2	1.0