



DEPARTMENT OF POLITICAL SCIENCE
UNIVERSITY OF MASSACHUSETTS, BOSTON

WORKING PAPER SERIES

INTRODUCTION TO PROGRAM EVALUATION

By

CHARLES CNUDDE

Professor
Department of Political Science
University of Massachusetts, Boston
Email: charles.cnudde@umb.edu

December 2005

Note: This paper represents a work in progress. Please contact the author before citing any information from this paper.

Introduction to Program Evaluation

This paper introduces program evaluation, not by spending a lot of time defining it, but by providing two cases of actual program evaluations of certain type. The assumption behind this decision is that students want to get right into evaluations rather than hearing about them in an academic way. Nevertheless there is some terminology that is appropriate to consider because these terms pervade the field.

Introduction

The terminology at hand includes the ideas of formative and summative evaluations. Whereas program evaluation refers to the assessment of the impact of programs of agencies in the public sector, these other terms refer to process and out-come evaluations respectively. Many writers refer only to out-come evaluations because those measure the impact of the agency involved. On the other hand it is also important to access the nature of the processes that presumably lead to those impacts in order to test whether they are in fact their explanations. We will also need to consider the differences between quantitative and qualitative measures and the reasons for the controversy behind these differences.

The examples that we will consider are evaluations in the educational arena, at both the higher education and the public school levels. One case is an internal evaluation of an academic program. It presents a self-study required as a step in the evaluation process. The other is a critical evaluation of a state (and now, Federally,) mandated program of testing. Both are summative evaluations in that they are “out-come” oriented.

Background

The first case is a draft of a portion of a self-study prepared by the author for a larger evaluation which includes a formative evaluation as well. The larger report concerned curricula and faculty fields of expertise. The formative portion will not concern us at this time.

In this case example is similar to that of many other program evaluations in that it assesses the success of the unit in meeting its goals. Its goals, as set by the Board of Trustees of the UMass System, are instruction, research, and public service. “Success” in meeting the goals depends upon the measurable effectiveness and efficiency of the program’s activities. Observers can make their own judgments about this success or lack thereof through examining the measures of effectiveness and efficiency.

AQUAD As Program Evaluation

The Board of Trustees, the Constitutional governing body of the University of Massachusetts, requires a periodic review of every degree granting unit in the UMass System, (usually occurring every seven to five years.) The requirement is based upon a decision reached at the end of the 1990’s:

In December of 1997, the University of Massachusetts Board of Trustees Approved Academic Quality Assessment and Development (AQUAD), a component of the University Performance Measurement System, to be initiated at the campus level this academic year. (Provost Memorandum, September 11, 2003.)

It is clear from the Provost’s report that the Board had in mind what Carol Weiss’ terms (in her path-breaking study), “a way to increase the rationality of policy making”. (Weiss, 1972, pg. 2.)

Furthermore, although AQUAD assumes empirical measurement, it is not a “pure” experiment in Campbell and Stanley’s sense. Instead we should look at program evaluation as theoretically based upon the underlying ideas of experimental research, with the pure experimental design serving as a “model” rather than the reality of what is involved. (Campbell and Stanley. 1966, pg. 25.)

That is, experiment 6, the superior design in Campbell and Stanley’s classic analysis, provides the introduction of a treatment to an experimental group and a comparison of the results between this group and a control group. Both groups are composed of randomly selected individuals or units. In symbolic terms Campbell and Stanley give us the information listed below in Table 1.

Table 1. The Ideal Experimental Design

R	X	O ₁
R		O ₂

Where R indicates random selection, X indicates the treatment, and O₁ and O₂ indicate observations or measurements of the effect of X. Since O₁ receives X, it is the treatment group in the experiment, and O₂, does not as it is the control group.

(Source: Campbell and Stanley, 1966, pg. 25.)

How does this design apply to the reality of program evaluation? In this format, X refers to the operation of the program. The program is the treatment. The O’s refer to attempts to observe or measure the program’s effects. O₁ refers to observations of individuals or other units that experienced the program and O₂ to those that did not.

The comparison between O_1 and O_2 then is the evaluation. The hypothesis is that the individuals or units in O_1 will perform in a way specified by the program and O_2 will not, otherwise the program was not successful and the program failed.

We will leave the very important discussion of what (R) or randomization is to a later discussion. For now we may conclude that random selection removes any other systematic difference between the “treatment” group (the group of individuals or units that receive the treatment, X,) and the “control” group (those that do not receive the treatment) so that any difference between O_1 and O_2 have to be due to the effects of the program.

This model, then, is the basic logic of program evaluation. It is a model in that usually program evaluation does not include randomization and therefore control of which individuals or units receive the treatment and those who do not.

There are practical and ethical issues involved in decisions to withhold treatment from a control group. For example we do not withhold community policing from portions of the city if we believe that treatment will save lives. Instead we usually look at the “natural world” of which units experience the program and compare them with their experience in the past before community policing started. Thus in Campbell and Stanley’s terms, this design would constitute a quasi-experiment (pp. 34-63) because there may be other factors that differed across the time period that could “jeopardize” the validity of the results. Other terms for this style of research would be non-experimental (Blalock, 1964) or natural experiments, where such experimental control is absent.

Yet the evaluation proposes to investigate whether the program is successful. If so it is based upon the inference that X will bring a change in O in the predicted direction and the evaluation study, by inference, will verify the success of the program.

As Campbell and Stanley (1966, Preface) have noted, many examples of research of this kind takes place in an educational setting. The AQUAD review is such an example. We will have the opportunity to observe others as well.

Let's look at the AQUAD review as a case of program evaluation. As indicated above it is in several parts. The first is a "self study" in which the program under review describes its main features and a second is a report based upon a site visit by qualified academics, mostly from other institutions, chosen to assess the quality of the self study and the program. Last year the department of political science at UMass-Boston underwent such an AQUAD review.

Its self study, like most in the AQUAD system, is primarily a description of the elements of the program. However, the concluding portion of the self study goes beyond description. It attempts to document the effectiveness and efficiency of the department's performance.

As it is the department's own evaluation, it is an internal program evaluation. Nevertheless it is similar to other evaluations in the public sector in that it assesses the success of the organization in meeting the goals set through its accountability structure. The following is the effectiveness and effectiveness section of the Department's self study. (Department of Political Science, 2004.)

Effectiveness and Efficiency

The Department of Political Science has demonstrated over the years that it is effective in meeting the instructional needs of our student body and it is efficient in doing so. In this section of the report we will address the different meanings of the terms, efficiency and effectiveness, while necessarily settling upon particular operational definitions. What ever definition one prefers, our experience and much anecdotal evidence support the assertion that the department is both. The preceding pages of this report document it as well.

Nevertheless it is useful to restate that evidence. Those of us who have taught elsewhere know that our students, despite carrying heavy employment workloads, gain knowledge in our courses on a par with students at the top institutions in the nation. We know that many of our students win prizes for their academic and service achievements. After graduation our students go on to the better and professional and programs. Furthermore evidence from the institution's Office of Institutional Research and Policy Studies (OIRP) and from the department's own data confirms these observations with systematic information.

Effectiveness

As noted, scholars differ on the definition of effectiveness in the public sector and in higher education. To an extent, there is confusion over where effectiveness leaves off and where efficiency begins. Both terms require an examination of performance. This section will look at comparative performance. The next will examine performance per faculty member.

The Department's faculty is effective in making itself available to our students. In the one-year period ending in autumn 2002 (the latest data from OIRP) the department's undergraduate enrollment increased by almost fifteen per cent. This change can not be explained away as a one-time statistical "blip". It is a continuation of a trend. Over the five-year period of recent data, the department's increase has been almost eight per cent.

Nor was this a result of "piggy-backing" on a secular increase in enrollment for the entire institution. That is, if the institution gained enrollment, the average department should gain as well. However, the one-year change for the College was an enrollment loss of almost six per cent. For the Campus as a whole there was a similar loss over the same period amounting to about five percent. In short, there was about a twenty percent difference in the change in enrollments between the department and both the College and Campus over the year.

Over the five year period the enrollment growth rate for the department, noted above at nearly eight percent, compares favorably with the College data. The latter enrollment grew at less than three percent. The enrollment situation for the Campus as a whole was at roughly a steady-state. This means that the department's growth rate over the period was between two and three times that of the college and almost twenty-five times that of the Campus. Table EE1., below, provides the evidence supporting these conclusions.

Table EE 1. Trends in Undergraduate Enrollment

Unit	One-Year Change	Five-Year Change
Political Science	14.5%	7.6%
CAS	--5.6	2.6
Campus	--4.5	0.3

Source: OIRD, (2003) Statistical Portrait: Fall 2002.

Table EE 1, shows remarkable trends in undergraduate enrollment. Whether one examines the rate of change over one year or five years, the department is several magnitudes ahead of the College and the Campus. The effectiveness of the department's faculty out-performs relevant institutional comparisons in both recent and long-term changes in enrollments. When we measure effectiveness based on performance, the department has been doing its share, and more.

Another way to look at effectiveness is to examine the sheer number of students. Percent or rate increases over a very small number of students could be high, but the utility for the student body would be nil because of the small number being served. Consequently it is necessary to briefly look at the number of undergraduate majors in political science.

OIRD data for autumn 2003 show political science enrollments at 269 students. Although more recent department data show that number to be an undercount, nevertheless it is a reasonable size for statistical comparisons. If it were a random sample of the student body, for example, the changes discussed above would be statistically significant at acceptable social-science levels.

Yet another way to look at effectiveness is to ask whether these students, once enrolled, go on to complete their degrees? It is not an adequate measure of effectiveness to document high enrollments if those students simply recycle through the system and do not graduate.

Once again, OIRD data are instructive. In the year of the report there were seventy-seven undergraduate degrees in political science. The ratio of that number to the nearly two-hundred-seventy undergraduate majors is almost twenty-nine percent. This ratio compares favorably to those for the College and the Campus, which were about eighteen and twenty per cent respectively. Table EE 2., gives the data for this comparison.

Table EE 2. Fall 2002 Undergraduate Enrollment

Unit	Undergraduate Enrollment	Undergraduate Degrees	Degrees/ Enrollment
Political Science	269	77	28.6%
CAS	5863	1052	17.9
Campus	8113	1586	19.6

Source: OIRD (2003) Statistical Portrait: Fall 2002.

Table EE 2. gives the comparisons between the department and the College and Campus for enrollments, degrees, and their ratios. The comparison indicates that the department is more than thirty-seven percent more effective than the College in undergraduate degree attainment.

The similar comparison for the campus as a whole is a little over thirty-one percent. Thus the department not only deals with a large number of undergraduate students but it effectively brings them to graduation and at a higher rate than either the Campus or the College.

Efficiency

In the public sector one conceptualization of the difference between effectiveness and efficiency is that while the former may deal with performance per se, the latter compares performance with resource investments. It is “out-put” given “in-put”. ($E = O/I$). In the vernacular of Washington, D.C., it is, how much “bang for the buck.” For higher education the relevant issue is how much performance does the institution achieve per faculty member, noting the latter as the primary resource investment in this sector. An unfortunate bi-product of the use of the term efficiency is a set of related terms all of which have industrial connotations. Thus the discussion revolves about “investments” and “productivity”, the use of these terms should not color the discussion so as to ignore the question of educational quality.

Yet size matters. A department with more faculty members should, everything else held constant, produce more students and more graduates than a department with fewer faculty members. Consequently to gain an adequate measure of efficiency, it is necessary to compute performance as a function of the number of faculty in the units. As in the above discussions of effectiveness, the measures here will compare the department with the College and the Campus on the basis of the number of tenure track faculty. Table EE 3., provides the data for the comparisons.

Table EE 3. Undergraduate Enrollment and Degrees Per Tenure Track Faculty (FTE)

Unit	Undergraduate Enrollment	Degrees/Faculty
Political Science	25.6	7.3
CAS	27.0	4.9
Campus	24.5	4.8

Sources: OIRD (2003). Statistical Portrait: Fall 2002; Departmental Records.

Table EE.3. shows a kind of efficiency ratio for undergraduate enrollment and degrees successfully completed for the department, the College and the Campus. The Department is about as efficient as the College and a little more efficient than the campus as a whole in terms of undergraduate enrollment. The differences are about 1.6 in favor of the College and 0.9 against the Campus. Over all there appear to be negligible differences between the relevant units on this measure of efficiency. The department seems to operate on this dimension at about the same level as the average department on the campus.

The table does show important differences when it comes to efficiency in producing degrees. While the College and the Campus score about equally well on this dimension, 4.9 and 4.8 respectively, the department scores head and shoulders higher. At a ratio of 7.3 degrees per faculty member, the department scores about one-third higher than does the College and (even somewhat higher than) the Campus as a whole.

This difference means that the efficiency of the department provides a net savings to the Campus when it comes to investing in faculty to produce undergraduate degrees: for each department faculty member the Campus saves about one-third faculty member over the average department at UMB. Put another way, the department would be a reasonable place to invest additional faculty positions if the goal is to achieve more graduates from the institution.

Moreover, the data show that the increase in degree production by the department is not at a cost to undergraduate enrollments. Since there is a limit to the amount of time that faculty have available, one could foresee a situation in which faculty lead more students to their degrees but at a cost of dealing with students in regular course work and vice-versa. In such situations degree production would come at a cost in enrollments or additional enrollments at a cost in degree production.

The evidence that this situation is not the case for the departments is that the enrollment efficiency for political science is about the same as that of the College and the Campus as a whole. Thus degree production is much higher than average for the Campus while the enrollment production is about the same as for the Campus as a whole. The department holds its own in enrollments while achieving one-third more in degree production. Furthermore since the department prides itself on the quality of its instruction, this degree performance is not at the cost of quality in that regard either.

Final measures of efficiency concerns whether such a productive department in instruction lags in other areas of faculty productivity, such as in scholarship and service. Faculty members at most highly ranked institutions believe that instruction, service, and scholarship go hand-in-hand.

If faculty performance on one of these areas lags, faculty members raise the question of whether other faculty members as a group are doing the job. In asking this question they implicitly assume that the “job” requires the joint activities of instruction, scholarship, and service.

Service and Scholarship

Service activities have some definitional issues. Faculty members perform service to their departments, to their campuses, to the State, to the Nation, and to the discipline to which they belong. Although there is controversy concerning these definitions of service, we can approach an integration of these definitions by asking whether they scale in a statistical way.

In terms of service, departmental records indicate the number of such activities in which each member participates. Department records, and in particular, the departmental annual faculty report, provides information on the service activities of each member of the faculty. Ordinarily one takes these data at face value as measures of the faculty service contribution. Scholarship, on the other hand is another matter. Major institutions depend upon other sources of data, as will be argued below.

Answers to both of these questions, concerning service and scholarship, depend upon departmental records and nationally available records that provide information on these items. Table EE 4. gives the data on these dimensions.

Table EE 4. Service and Citation Experience of the Department of Political Science

Item	Total	Per Faculty
Citations	652	62.5
Service Activities	31	3.1

Sources: Departmental Records and ISI Web of Science (2003) Thomson ISI

In the area of service an analysis of the department records indicates that service activities among the faculty members are scalable, according to Guttman criteria. That is, if one ranks service in order from the department, to the campus to the State, to the nation, to the profession of political science, and in that order, the service activities of the members of the department are easily associated following that progression.

That means that although there are thirty-one documented service activities in the department, the more “difficult” activities, those at the level of the national profession, are fully represented among the “easier” activities that lead to those national service contributions. Although there are only five contributions to the national profession, those five are supported by activities at the levels below, statistically speaking.

As is the case with the definitions of service, as well as of, efficiency and effectiveness, there are controversies involved in the definition of scholarship. Some institutions simply count the number of articles and books published. No one seriously believes such counts adequately measure scholarship because it usually has a quality dimension in the way that most academic view it.

A different approach utilized by many superior institutions focuses upon the number of citations in the literature by others. The idea underlying idea of this focus is that the number of citations by others tends to measure the impact of a faculty member’s

scholarship on his or her discipline. For purposes of the self-study the department has obtained the number of citations of members of the tenure track faculty in political science faculty.

The data show that the average number of citations is over 65. (Source: ISI Web of Science (2003) Thomson ISI.) While the department does not have similar data for comparison with other units, it appears that this is a very favorable number, so say the least, for a department carrying such a heavy undergraduate load. It would be interesting to know what other units have achieved in this regard.

The number of citations is also a measure of the quality of instruction. The department is addressing the needs of the University's undergraduate students with faculty whose scholarship has been judged as rating a high enough quality to qualify for citation by other scholars publishing in the field. The faculty members in the department have an impact on their discipline at what looks like a high level of frequency. For instance, that these citations occur at an average rate of over 65 times in the recent source means that others cite their work at a high level.

Conclusions: Effectiveness and Efficiency

The Department of Political Science, composed of a relatively small number of faculty members (about ten full-time in the tenure track before the sad death of our department chair) has achieved a high level of effectiveness and efficiency. By one measure it is about one-third more productive than the average department on the campus. By another

measure it is thirty per cent more effective than average. Finally its faculty members have been active in service and have been recently cited in the literature by other scholars, attesting to the high quality of the talent the department places before the University's students.

MCAS as Program Evaluation

Before turning to a second case in evaluation it will be instructive to look at some observations as part of the logic in preparing for it. The previous evaluation provides answers to questions that the program managers believed were important to the Board of Trustees, the authorities the managers faced in the accountability structure. For the next program evaluation, that of the Massachusetts Comprehensive Achievement System (MCAS), the role of the accountability structure is even more important. From the managers perspective the previous evaluation documented the success of the program in meeting the three important goals of instruction, research, and public service. Consequently one could assume that the evaluation was a success. The MCAS evaluation raises questions of doubt about that easy conclusion.

MCAS was a part of legislation passed in 1993 in the Educational Reform Law passed that year by the Commonwealth Legislature. It establishes a testing and evaluation procedure for the schools in the state. In the case of MCAS there is less agreement on the success of the program. In the previous case the evaluation did not explicitly examine the relationship between the formative and summative evaluations. That is, the question

remains whether the processes the program manages resulted in the outcomes measured. That question is one of causality and it is one that when raised explicitly plagues the MCAS program.

In any program evaluation one can look at the issues as questions of causality. Is the effectiveness of the program in reaching its goals a result of the processes of the program? In Blalock's terms, one is attempting to make inferences about the causal effects of the program. (Blalock, 1964) The causal structure would look like this:

Formative Evaluation ----- Summative Evaluation

Of course there are other factors that influence program goals. These other factors need to be taken into account in a properly designed program evaluation. For example, success in student job placement in jobs might not be a result of the educational program, but instead of the selection process the institution uses to recruit individuals who can be predicted to be successful, such as upper class students. Similarly, diversity in the student population might be a happenstance of the location of the institution rather than any program of student recruitment.

The important point is that in non-experimental program evaluations, these other factors can not be controlled through random selection. Instead, we introduce statistical controls as surrogates for controlled experiments. In the example above, one would control for the class status of the students in evaluating their placement.

Consequently the use of modern statistics has come to dominate many program evaluations. For this reason sophisticated program evaluations need to look at the advantages and disadvantages of quantitative analysis.

Moreover, in any example of program evaluation observers can critique the adequacy of the measurement. In the case of educational evaluations there are many factors at play in program success. As the examination of the second evaluation will indicate, as one becomes sophisticated about data measurement, one can raise questions about the quality of the observations themselves.

A final observation about the AQUAD evaluation as presented here, is that it emphasizes quantitative measures. Are quantitative data the only way to access success?

A Critique of MCAS as Program Evaluation

We can begin to answer that question by looking at the short example of program evaluation in the other sector of the educational arena that we will consider in this introduction, MCAS. Although there have been many questions about MCAS the critique below from an educational leader whose school performed well under the procedure is especially telling.

In the legislation the state mandated MCAS to:

1. Test all public school students across the Commonwealth, including students with disabilities and students with limited English proficiency.
2. Administer such tests in selected grades.
3. Measure performance based on the learning standards in the “Massachusetts Curriculum Frameworks.”

4. Report on the performance of individual students, schools, and districts.
5. Serve as one basis of accountability for students, schools, and districts (for example, grade 10 students must pass the MCAS tests as one condition of eligibility for earning a high school diploma.)

In addition to meeting the requirements of the Education Reform Law, the MCAS Tests also fulfill the requirements of the federal No Child Left Behind (NCLB) law. NCLB requires annual assessments in reading and mathematics for students in Grades 3-8 and high school. Students also must be tested annually in science in an elementary school grade, a middle school grade, and a high school (10-12) grade. This requirement is fulfilled in Massachusetts by testing students in grades 5, 8, and high school.

Source: Massachusetts Department of Education. (2005) "Massachusetts Comprehensive Assessment System" (Mass.Gov. website.)

The Department of Education test items and scoring guidelines make it clear that MCAS emphasis is on quantitative measures that can be used to hold schools and teachers accountable for student performance. Under the new program graduation depends not only on passing courses but passing the test as well. If students fail the test their diploma would be withheld. Schools also can be graded on the percent of students who pass. The Massachusetts reform predated the Federal legislation. It was an attempt to establish "bench marks" for performance in an attempt to improve the quality of education in the state. It included increased funding of the public schools and special increases for schools in financially poor communities.

Now the state program has been co-opted by the newer Federal act, the testing program has reached a new level of importance. Failing schools will be identified so parents can choose whether to continue sending their children to them or, if alternatives are available, to send them elsewhere. The Market theory underlying the Federal plan holds that schools with failing scores would have an incentive to improve to prevent enrollment and related budgetary losses.

Thus MCAS is an accountability tool that reaches the students, the teachers, and the schools. MCAS, like the Federal effort, measures student success or lack thereof.

Summary measures of that performance then can be used to rank the teachers and the schools. Not only will students fail, but schools will as well.

MCAS as it has been implemented has drawn criticism. A school head master reports that MCAS as program evaluation showed her school to be a success. “The MCAS results for the class of 2003 at (school) have just been reported, and I am faced with a paradox. Our students did well. Like many other schools across the state, we had a dramatic reduction in the number of failing students” Nathan, 2005, pg. 4.)

Nevertheless she presents an informative critique of the program.:

Few Educators imagined that the MCAS would become the sole method for Accessing and validating student achievement. The Massachusetts Education Reform Act of 1993 promised schools that would fully develop the talents and skills of every child in the commonwealth. This legislation was a response to serious inequities in funding and performance in school districts across the state. The sponsors of the act promised the public “accountability” and clear benchmarks for student achievement. No longer would students be passed through school systems without acquiring academic skills. . . .
But this way of thinking about teaching and learning, for all its superficial logic is fundamentally flawed. First, research shows that high-stakes tests discourage and demoralize at least as many students and teachers as they motivate to work harder. The notion that threats of punishment (e.g., withholding a diploma) will create a positive learning environment and radically transform our most beleaguered educational institutions is not only discredited by research but is also grotesquely wrong-headed and cruel. High-stakes environments push dropout rates up, particularly for the most vulnerable students. The state’s current policy will hurt most grievously the very students who are supposed to benefit most from it. . . .

A one-size-fits-all test that determines every student’s future takes the most important decisions about teaching and learning away from those closest to students: their teachers and families.

That students . . . might actually do well in high school and perform adequately in reading, writing, math, and science but still get low scores on standardized test in

inconceivable to the high-stakes enthusiasts. Common sense and common experience prove that their idea of school reform is a fantasy—a war game in which young people are the expendable pawns.—Nathan, Linda. 2002. “The Human Face of the High-Stakes Testing Story” Phi Delta Kappan. 83, no. 8, (pp. 595-600.) Quoted by permission.

The point of view represented in this passage is of the “front-line” teacher with intimate knowledge of student performance. For her, it would seem, the abstract score derived from simple paper-and pencil-test is a superficial measure of a student’s knowledge. In this sense her reaction is similar to the general charge that quantitative measures are inferior to qualitative means in evaluation. That is, often what we are trying to evaluate is more complex than any simplistic measurement instrument could capture. The field of education is a good one to illustrate this charge, because what could be more complex than learning.

These issues raise questions concerning the causal structure assumed in program evaluation. Not only do evaluations need to include the process factors (formative evaluation) that presumably cause the outcomes (summative evaluation) at issue, but the measurement of outcomes themselves are at issue. That is, the measured outcomes may be very incomplete indicators of the goals at issue, such as learning. In causal terms measured learning, such as MCAS scores, could leave out much of the complexity of what we really mean by learning. Thus:

	Measured Score
Summative Evaluation	
	Learning

Not only have we not controlled other factors that affect learning but in our quantitative measure we have not adequately accounted for everything involved in learning as well. If this causal structure is what we mean when we say the quantitative measures are inadequate, then the logic of the relation of formative evaluation to summative evaluation needs to be reexamined. In statistical terms this causal structure means that we have error in our measurement. If such error occurs in the summative evaluation, it would be impossible to rule it out in the formative evaluation as well. This means that all of our statistical associations in the evaluation have their validity jeopardized, not only because of the lack of random controls but because of measurement error. The sad fact of all this is that qualitative assessment will not solve the problem either. If we can show causally what the problem is in a quantitative sense, a move away from the quantitative format merely obscures the problem.

Conclusions: the Politics of Program Evaluation

Yet if we can show that there is a “common-sense” problem here, why is there pressure on evaluation in the quantitative direction? The policy makers who push MCAS and AQUAD are in a logical bind. They want to achieve accountability for excellence, yet they typically preside over such a large number of units that their only way to make comparisons is through simplified measures. It is logical for them then to push for quantifiable measures to simplify their problem of controlling the agency. Yet the charge from those most intimately informed of the actual performance is that the simplification itself may elude excellence.

These examples suggest that the conflict between quantitative and qualitative procedures is not merely a conflict over methods. It is at base, political and it depends upon where in the accountability structure the actors in the conflict sit. Those near the bottom are concerned with excellence in the agency's performance. Through their face-to-face contact at the operating level they know what this means. They prefer qualitative assessment. Those at the top are concerned with adequate performance while allocating resources across a number of agencies. They prefer quantitative measures. Those in between prefer one or the other, depending on how close to the bottom or top of the hierarchy they find themselves. If this is correct, we can conclude that both sets of proponents are correct. What kind of evaluation one would conduct, depends upon what kind of question is asked. It is one concerning accountability and resource allocation or is it one concerning strict excellence at the operating level.

In addition to questions about method, we can also raise questions about the philosophical underpinnings of program evaluation in general. That is, we know that information is "socially constructed". That philosophical tradition is in contrast to the "neo-positivist" view which derives from the work of the English philosopher, A. J. Ayer. (1936.). In practical terms, try telling the Board of Trustees, or the State Department of Education about the philosophical assumptions they make in requiring program evaluation.

Partly as a result of the emphasis of the Federal government in evaluating its grant-in-aid programs, program evaluation has become pervasive in the public sector. Yet it is surprising to see the extent of program evaluation at large in modern society take place without serious challenges. The serious student of public affairs needs to know the

positive and negative features of evaluation. This knowledge presumes knowledge of the process and techniques of this area of knowledge. We who are about to evaluate salute you.

References

Ayer, A. J. (1936) Language, Truth, and Logic.

Blalock, Hubert. (1964) Causal Inferences In Non-Experimental Research. (Chapel Hill, NC: University of North Carolina Press)

Department of Political Science (2004) “Self-study, Effectiveness and Efficiency”. (Boston, MA: University of Massachusetts)

Department of Political Science (2004) “Annual Faculty Reports”. (Boston, MA: University of Massachusetts)

Massachusetts Department of Education. (2005) “Massachusetts Comprehensive Assessment System” (Mass.Gov. Website)

Nathan, Linda (2005) “The Human Face of the High-Stakes Testing Story”. Phi Delta Kappan. 83, no. 8, pp. 595-600.

OIRD (2003), Statistical Portrait, Fall 2002

Provost’s Memorandum (1999), “Academic Quality Assessment and Development”, (Boston, MA: University of Massachusetts)

Weiss, Carol. (1972) Evaluation Research, (Englewood Cliffs, NJ: Prentice-Hall)