Evaluation Bias and Its Control*

by

Michael Scriven
University of California, Berkeley

June, 1975

Paper #4
Occasional Paper Series

*A paper originally written in conjunction with the Evaluation Center's Grant # GW - 7903 from the National Science Foundation.

Introduction

In this paper I shall consider certain aspects of the problem of obtaining unbiased information about the merits of a program or product, whether for purposes of decision making or for accountability. The evaluation of personnel, as well as the evaluation of proposals and evaluations, generally involves a different set of problems than those which I will consider here. However, some points made here will apply. Throughout, efforts are made to consider both the credibility and the validity of an evaluation, the former being (roughly) the audience's estimate of the latter.

Since the audience’s estimate is sometimes affected by considerations that are, as it happens, irrelevant in a particular case; and since the function of an evaluation is sometimes in part to provide credibility and not just validity, evaluation design must sometimes involve considerations that go beyond validity. This must not be viewed as pandering to prejudice, but as of the essence of certification, of accountability, in a more general sense of the educational and social obligations of the evaluation. (“It is not enough that justice be done, it must also be the case that it be seen that justice is done.”)

Let us begin by looking at some typical important practical cases of bias in program evaluation.

Divided Loyalty and the Co-option of Staff Evaluation

The simplest instance of bias in program evaluation is the case of the evaluator who is part of the program staff and loses objectivity because of social and economic bonds to the development staff, compounded by the cumulative effect of repeated acceptance (or rejection) of evaluative suggestions. The resulting situation of quasi co-authorship (or frustrated co-authorship) naturally destroys the external credibility of the evaluation and often the validity of the evaluative judgments.

The remedy is to add external evaluators. Being short-term consultants, these do not or cannot replace the staff evaluators for day-to-day purposes. (If they are not short-termers, they rapidly become quasi-staff.) Faced with these visitors, the staff evaluator often exhibits considerable ambivalence. Professional bonds struggle with work-mate bonds, with rather erratic results.

Another approach is possible within fairly large organizations, such as states, most school districts, and R & D units. This involves the systematic rotation of evaluation staff from project to project so as to avoid the effects of excessive loyalty or hostility. This is sometimes complicated by the need for special expertise (e.g., in math curricula), but the excuse is produced more often than it deserves. Rotation is usually possible, and nearly always desirable for much the same reasons as in the diplomatic and armed services. It should be imposed by management as part of the discipline of the job, the requirements of good performance, in much the same way as inservice updating should be required of staff physicians in a clinic.

Divided Loyalty and Project Monitoring

The project monitor from the funding agency faces related problems but in a different context. While visiting the project, he or she is seen as an external evaluator, but back in the capitol, a switch in role is often required (or naturally adopted) to that of project advocate.

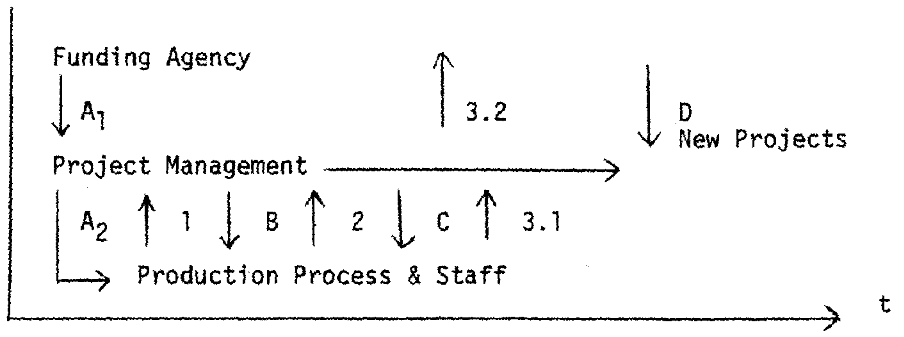
One recently espoused remedy is to segregate the monitoring function entirely, possibly through subcontracting the evaluation, to rechristen the liaison person with a title such as program officer or associate, and thus wholly legitimate the advocacy role at the agency. Another solution is to interchange the roles just described, that is, have the monitor carry out the evaluation and have the project appoint a resident advocate representative in Washington, or have someone on the staff who could go there at a moment’s notice. The big contractors, of course, adopt this alternative.

Now, using such cases and proto-solutions as a springboard, can we begin to see the outlines of some general approaches?

Organizational Bias Control

The first great step towards accountability (or just towards decent work) consisted in requiring that there *be* some evaluation of tax-funded or foundation-funded projects. At first, this meant no more than rechristening the final report. In any case it amounted to requesting Jones to be sure to tell the agency whether she or he had done a good job. This is obviously not likely to produce unbiased feedback, but it is less obvious that there are *two* sources of bias in the situation. The agency has made a grant, so it is in a parental role, and the success or failure of the grant is partly an evaluation of the agency itself. Organizationally, the situation can be represented in terms of a table or diagram as shown below.

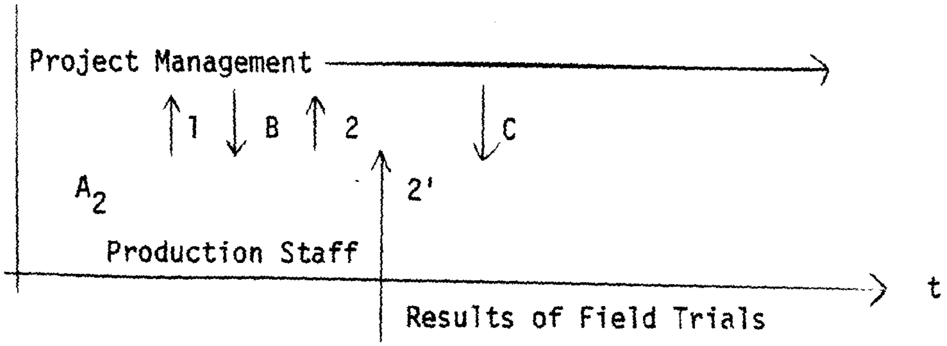
Figure I illustrates the situation we have just been discussing (letters represent actions, such as requests, payment, orders, and support; numbers represent evaluative feedback). We notice that there is a closed circuit in the evaluation of the project, beginning with Al, and concluding with 3.2, i.e., the money goes to the organization that evaluates its use. Similarly, project management sends money down A2 and gets evaluation feedback via 1, 2, and 3.1.



A = Initiation steps, staff appointments, materials orders, etc.
Al = Initiating actions by Funding Agency
A2 = Initiating actions by Project Management
B,C = Developmental decisions, staff replacement, etc.
D = Agency decisions about future funding
1 = Early formative evaluation feedback to service decisions B
2 = Later feedback to service decisions C
3.1 = Final summary of data to service 3.2
3.2 = Summative evaluation of *project*, can be regarded as formative by the *foundation* with respect to its ongoing activity, i.e., as serving D
t = Time

Figure I
Cycles of Decisions and Feedback: Internal Evaluation Only

These circles are risky, though not wholly avoidable. The trick is to provide some procedure that eliminates the complete dependence of management on a *single* feedback circuit. The feedback systems of Figure I are inherently biased positively, and we have to introduce a circuit with a balancing bias, or some reality constraint that cuts across a circle. This can be represented as in Figure II. The simple circle is broken by the intrusion of field trial results. Notice it is broken only if that data goes straight to project management. If it goes only to the production staff, it does not cut the circle (actually, an overlapping series of circles) of formative evaluation. And at best it only cuts the formative circle, not the summative one.



A2 = Initiating actions (the content of A) by Project Management

B,C = Developmental decisions, staff replacement, etc.

1 = Early formative evaluation feedback to service decisions B
2 = Later feedback to service decisions C
2'= Formative feedback from field trials provided to Project Management
t = Time

Figure II
Cycles of Decisions and Feedback with Internal Evaluation Plus Field Trials

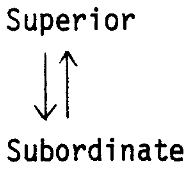
It is impossible to understand the persistence of the incestuous pattern of Figure I with its tendency towards optimistic bias unless one realizes that both parties involved have a motive for continuing it. The agency wants favorable feedback about its actions, and the project wants the agency to think well of it (and wants to think well of itself). So, of course, the stable situation is one of highly favorable evaluation. In the technological area, there is reality feedback to the agency later (breaking the potentially vicious summative circle) from marketing or medical data, which keeps the system honest. But in education and armaments, though such feedback is possible in principle, it is all to often transmitted through and hence open to (possible unconscious) corruption by the responsible agency: it does not break the circle.

Against this formidable alliance, the search for truth is a little short of soldiers. If the principal value of the funding agent is maximizing the social contribution of every dollar granted, as of course its rhetoric and in fact its situation requires, then there has to be an attempt to get evaluation of its projects from sources not quite so predisposed towards a favorable response.

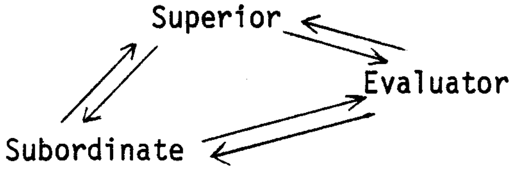
The circle we are now talking about trying to break is a third one, superimposed above the two already diagrammed. It begins with Congress, or the taxpayer via Congress, funding the agency and eventually receiving evaluation reports from the agency on its stewardship. Recent years have seen Congress increasingly sensitive to the managerial weaknesses of that system, bringing in OMB, GAO, and OCA as independent evaluators providing a feedback loop with at least less tendency to positive bias.

If only the reality data was readable by the amateur, one would just have to compare it to the reports for the evaluations to have a circle-breaking procedure. But that data, on big projects, needs computer processing, statistical reduction and expert interpretation before its significance is apparent. Each of these steps involves the possibility of distortion and the necessity of expertise. If the experts used are the same ones whose performance is being evaluated, the “reality” line does not cut the circle. Hence a more general solution involves using an independent gatherer/interpreter of reality data, the external evaluator.

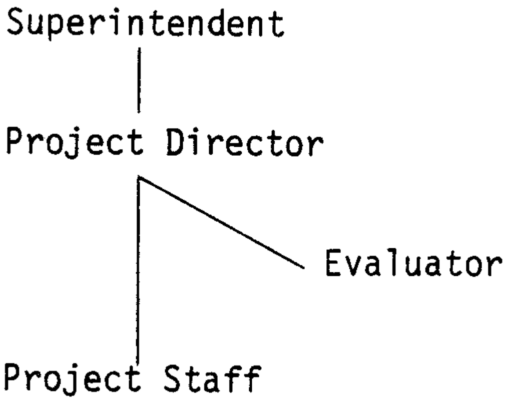
Perhaps we can infer preliminary forms of two general principles for minimizing bias from these considerations. They probably have only mnemonic status, not deep theoretical significance, but at least they are comprehensible. The First Principle is the Principle of Independent Feedback, which states that no unit should rely entirely on a given subunit for evaluative feedback about that same subunit Diagrammatically we need to replace:



with:



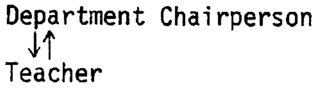
The fact that one has an independent feedback loop between *one* pair of levels does not satisfy the First Principle, it requires such an arrangement (not necessarily permanently installed) between *every* pair of adjacent levels. Here is a situation one frequently encounters:



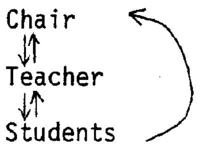
One young evaluator in this situation came to me not long ago with a sad but not unusual story. He had evaluated the project as required and submitted a report. All the critical comments in the evaluation were then excised, his name removed, and the result forwarded to the superintendent by the director as a "synthesized" evaluation of the project. Of course, the responsibility for getting corrupt feedback like that is partly the superintendent's, for violating the First Principle.

The cheap way to get independent feedback in these situations is to bypass the whole chain of command with the feedback loop and put a single evaluator in it, instead of duplicating the evaluator installed at the lower level; or one can use the evaluator already shown in a double role. A device I have introduced into Title VII evaluation arrangements simply requires that a duplicate copy of all communications from the evaluator to the project director goes up to the higher level (in the Title VII situation that means the project officer at USOE). This makes the director take the evaluation much more seriously in the formative stage, and makes cheating (as in the case described above) impossible in the summative situation. Of course, the director can supplement, annotate, or refute the evaluator's contentions--but must do so openly, not by excision. There are of course some costs in time and friendship, but there are no free lunches in evaluation.

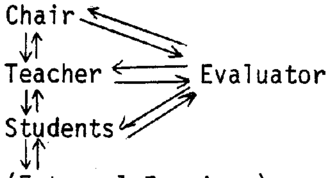
Consider the evaluation of teachers (or students) in the light of this principle, the same remarks apply to projects. Self-styled progressive schools and colleges sometimes go in for so-called selfevaluation--meaning a reflective but wholly self-generated report--as the key procedure. While it probably has a place in a decent system of evaluation, it cannot replace such a system. A better system uses feedback from the students directly to the department chairperson. A still better system (Swarthmore, Oxbridge, Australia) uses an external examiner to determine the students' achievement and hence the efficacy of the teacher, a better indicator than opinion. The feedback loops for the three systems differ as shown in the diagram below.



(Self-Evaluation)



(Student Evaluation)



(External Examiner)

It is extremely difficult to dismiss the arguments for external evaluation.

The most general issue in this area concerns the decision whether to segregate the in-house evaluation staff in their own unit, or have evaluators attached to and paid by each regular unit. In the case of an educational materials development institution (R & D center or publisher, for example), the options look like this:

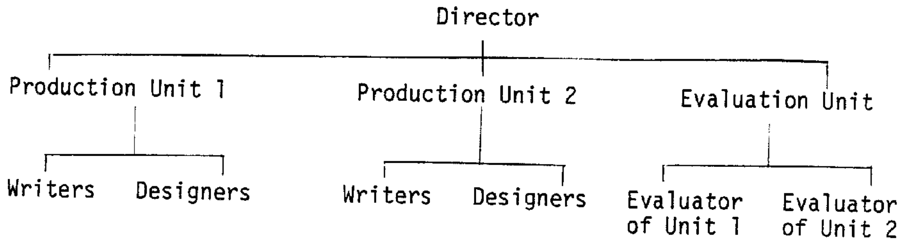


Figure III

Segregated Evaluation Model

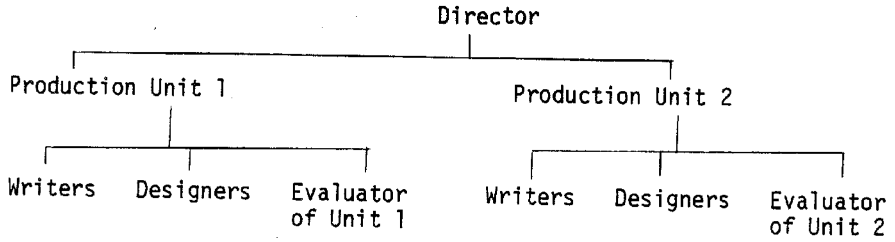


Figure IV

Integrated Evaluation Model

It is obvious that the “integrated” model (Figure IV) violates the First Principle at the first “step-down,” as far as the director’s feedback is concerned. And, provided that the about-to-be-discussed Second Principle is taken into account, it is my experience that the “segregated” plan does work better. Of course, the feedback from the evaluators must go to the unit managers as well as to the director of the whole shop, and there may <u>also</u> be a need for internal evaluation staff within the units of a large shop.

The trade-offs one must accept in using the segregated model sometimes involve loss of access to data or help with interpreting it, because the quasi-external evaluator in this model is seen as more alien, more of a threat than in the integrated model. A “hybrid” plan, similar to the one mentioned earlier, is also possible with the evaluators located in units but with double reporting duties. Sometimes this makes more sense if the outside reporting route is via a principal evaluation officer on the director’s staff who schedules regular meetings with the evaluators for purposes of development, sharing of problems, discussions with consultants, etc.

The Second Principle, the Principle of the Instability of Independence, reminds one that organization charts rarely reflect much of reality, and, in particular, the longer they have been true, the less true they are, particularly as a basis for evaluation feedback mapping. Independence, when it exists at all, is a fleeting state conspired against by almost all forces in a bureaucracy, and the Second Principle tells you that you have to have a definite program of systematic renewal and replacement or your evaluative feedback system will deteriorate severely, often without any sign that will be apparent internally. The simplest case of this is the staff evaluator who gets co-opted by the acceptance of his or her criticism or suggestions. Looking at the organization chart in Figure IV from the viewpoint of the unit manager in an integrated setup, it appears that the First Principle has been applied. And so it has, formally speaking. There is independent feedback if the criterion for independence is separate bodies. Obviously that is neither necessary nor sufficient. But it is a good start because anything less lacks credibility. It is not enough because friendship and enmity and ignorance and rigidification do not show up on charts. You must have something in the system that will identify deterioration of independence or objectivity and provide support or replacement where indicated.

The Second Principle requires that provision must be made to insure and continually reinsure the independence of the evaluators. The informal version of this principle is: Make sure the evaluators get evaluated. This suggests a worry about infinite regression, but no such necessity occurs in practice because of rapid convergence. To take a specific example, the Central Midwest Regional Lab used to have (and perhaps still does have) three levels of evaluators operating on a hybrid model with an annual external review by a National Advisory Board on Evaluation. The Lab’s director furthermore arranged for a steady influx of new blood to the National Board by rotating people off it. Thus at the “working level” there are staff evaluators, assigned and reporting to projects but selected with help from and monitored by the Lab’s evaluation officer (second level), who reports to the lab director and is directly overseen by the National Advisory Board (third level). The only organizational weakness that turned up in that scheme was what looked to me (as Chairman of the National Board) like a long-run (three year) reduction in sensitivity of the project evaluators, and the Second Principle would suggest using a rotation system to avoid this.

Another worry about the Second Principle is that adding hierarchies of evaluators looks like a costly business. It should involve no net cost at all, usually no net cost to the organization, and certainly no long-run net cost to the “consumer” (taxpayers and/or users), and it should be designed within that constraint. (See “The Doctrine of Cost Free Evaluation” [Scriven, 1974, pp. 85-93].) The Second Principle implies that independence requires regular verification and support and can be seen as the diachronic (through time) complement of the synchronic First Principle. The Second Principle may lead to recommending an oscillation between two organizational arrangements just because the aging of organizations leads to senility (after they achieve maturity). Moreover, a return to an arrangement that was initially inferior--on First Principle grounds--may provide considerable improvement in spite of what might appear to be decreases in the independence of the feedback. For example, after a period of heavy reliance on a particular external evaluator with a very definite “line” about evaluation, a project may benefit from a period of internal evaluation where the lessons learned from the outsider can be built into the ongoing work in ways that may be too subtle for the short-term external consultant to detect. But what has really happened here is not at variance with the Second Principle over a _short_ period. The staff--once-they have been sensitized by the external evaluator--are now in a position to produce evaluative suggestions that are independent of him or her and, at this stage, more useful to the project. I have seen this point corrupted into the idea that “from now on we don’t need external review.” Of course, it will only be a matter of months, at most a year, before the rigidifying effects of a constant social environment are likely to lead to oversights that a continued application of the Second Principle would identify, and so the introduction of some _new_ kind of independent feedback loop should be planned.

A milder treatment is to switch to another external evaluator. Such a move is sometimes called for by the Second Principle. It _Is_ called for if the suggestions from the original evaluator’s next visit are (a) completely predictable, (b) probably unfeasible or invalid, although (c) the situation or data or staff have changed very significantly. Here we have the not uncommon phenomenon of rigidification of the evaluator. That is one kind of loss of independence--the bias now being internal (to the evaluator) rather than external (e.g., due to economic advantages of a favorable judgment). On the other hand, there are occasions when the repeated evaluation is as true as ever and the advice given is as sound, and the reasons for rejecting it as unsound. Then a switch to another competent evaluator will predictably produce the same advice. To put it another way, there’s nothing wrong with the independence of the present evaluator, and the Second Principle cannot be invoked to justify a change. Loss of independence with time is a tendency and does not _necessarily_ occur in less than fifty years, though it’s _likely_ to occur in less than one.

A Closer Look at Bias

The quest for objectivity via the criterion of independence often leads to the use of “external” evaluators in both the formative and summative situations. Now of course, externality is always relative.

Using someone from another department or school may be external enough for one’s needs. But there are ties that bind across those little gaps--ties of family, friendship, political alliances, and even the sameness of professional commitment. It is nearly always possible to find important similarities and/or differences in the value-systems of any evaluator and any evaluee. That is too often taken to be a sign of disqualifying bias. It is not. It is only a _possible_ cause of such a bias, not proof of its presence.

There is a crucial ambiguity in the concept of bias. It is sometimes interpreted as a statistically likely _tendency_ to systematic error (against which nepotism rules are formulated) and sometimes as an actual and systematic increase in the frequency of errors. The former is crucial in credibility considerations and the latter (narrower) concept in validity considerations. We need to be clear that <u>only</u> the latter affects validity. The Second Principle does not guarantee increasing bias (in the second sense) only an increasing probability of it. In legal and moral, as well as scientific contexts, only the second sense is relevant (except when politics is part of the problem, which sometimes converts the issue into a credibility one). One way to put the point is to say that one can overcome bias in the second sense but not in the first. If one's spouse is put on one's staff, one <u>has become</u> biased in the first sense; one will have a <u>tendency towards</u> bias, in the second sense--but one may be able to transcend it. Bias, in the first sense, is a statistical tendency in a group of which you are a member; in the second sense it is a tendency which has in fact infected you. Only in that sense is bias fatal to objectivity.

We have been stressing considerations of independence here, because this represents a partial operationalization of the crucial concept of bias. Is the evaluator's opinion formed on the basis of the relevant available evidence, independent of the irrelevant considerations such as friendship? That is the key question, and it is not hard to find evaluators who are highly independent of their social ties in this sense. Of course, for credibility purposes one will have to avoid the extremes of nepotism, etc. And a refinement of the First Principle warns us to beware of regarding people who are physically separate as judgmentally independent when they are paid by the same hand and rewarded for the alleged success or actual continuance of the same project.

The Second Principle warns us to look at the diachronic dimension when checking independence, and it interacts with the First Principle in various ways. Suppose you do hire an outside firm for evaluating a project, a firm whose headquarters are in a distant state. This <u>looks</u> like real independence. But ask yourself what the reward system is for that firm. It isn't any more rewarding for them if your project is successful or not, <u>per se</u>--and that's why you value their opinion, why they appear independent. But look a little deeper, or <u>longer</u>. What is rewarding to them <u>over the years</u>? Success in their business, which of course requires a continued flow of contracts. Since such firms are very well aware of the power of the grapevine in getting further clients, they are often well aware that an evaluation which shows the client in a good light is much more conducive to later contracts than a critical evaluation. The reverse side of this coin was brought home to me when communicating with a network of evaluators on a USOE grant. I heard more than one sad tale of "blackballing" an evaluator who gave a deservedly critical evaluation. In short, the "independence" of an external evaluator can be seriously compromised by the constraints of business success. For a brilliant exposition of the same phenomenon in the world of CPA's, see Briloff (1973).

Think back to the example of student evaluation of teaching. The time sequence is crucial. If the effects of that evaluation on the teacher will occur before the teacher evaluates the students, they have an incentive to give false positive evaluations. If the teacher evaluates the students before the reverse occurs, they have a “getting even” motivation for false negative evaluation, and the teacher has a bribery motivation for false positive evaluation of the students. It is possible to handle these problems, but it is usually done badly because no one looks at the feedback loops.

An example of the way in which apparent independence is corrupted by professional ties can be seen in most accreditation reports by teams visiting, e.g., high schools. The team contains, e.g., specialists in driver education, who “site-visit” the driver education department and return with the judgment that driver education needs more support than it’s getting from the school administration.

Practical Implications

Four morals emerge of concern to us all, evaluators and evaluees alike. First, it is a serious management error to provide funds for external summative evaluation to a project, since if the project management contracts out evaluation of their work, the phenomenon just described will have the maximum effect, i.e., they will tend to pick “friendly” evaluators or fix the REP to eliminate some serious sources of negative evaluation (Sesame Street’s contact to ETS is an example). Second, where the funding agency contracts out the evaluation itself, thereby avoiding the preceding objection, one is still not entirely free of the problem since the agency’s own decision to fund the project is indirectly under evaluation and hence they too tend to want a favorable report, a fact which they quickly signal to the evaluator. Even where the project wasn’t much favored by them, but imposed by Congress, the agency is often incapable of avoiding ego-involvement in it. USOE’s suppression of a moderately critical Title I evaluation is a well-known example, and NSF has been involved in a similar case (so has every human institution, no doubt, the question is only whether serious efforts are made to minimize the frequency of such occurrences).

Now if an agency can’t ask its projects to get the summative evaluation done, and isn’t above suspicion even when it hires the evaluators itself, what’s left? Either a general-purpose evaluation office, like the general Accounting Office which currently serves this function as well as the fiscal one (albeit rather incompetently, since their staff has little training in the new role of general evaluation), or increased pressure from the ultimate loser (the taxpayer) via Congress to get the ego-protection of agencies rated lower rather than getting objective information to the public. Congress’ tendency to Monday-morning quarterbacking is a major cause of this trouble.

There is a “next-best” procedure, if neither of the preceding two suggestions can be immediately effectuated. It is quite natural for an agency that contracts independently for its evaluation project to use the same liaison officer for the evaluation contract as for the project contract. This is a fatal mistake. There must be at least separate individuals involved, even if not separate divisions of the agency. The reason is simple. The normal type of pressure on the liaison officer, discussed earlier, rapidly converts him into an advocate of the project back at the agency. Indeed, it is entirely appropriate that he should fulfill this role, since there’s usually nobody else to do it after the initial recommendation comes in for the review committee (which can be considered an advocate of the project in some remote sense). The problem is that if this project is also handling the evaluation project, the advocacy will lead to pressures on the evaluation contractors to soften their report, or is likely to lead to these pressures, in a way that simply reduces the independence of the feedback to the agency and the administrator. This has now happened too often for it to be ignored any longer.

One can see the sequence of sophistication in terms of the following steps in an imaginary history of evaluation arrangements. The first step consisted in asking the project to be sure to do an evaluation of itself. The second step consisted in asking the project to use an advisory committee of external experts to help it do an evaluation of itself. The next step consisted in requiring that it devote specific monies to evaluation; both of the previous steps, apart from bias, suffered from the fact that overruns were taken out of the hide of the evaluation. But this step still meant that the project--even if they appointed a subgroup of their staff to the summative evaluation task--was evaluating itself. The next step consisted in requiring that the project sub-contract the evaluation. This still left open the “control” of setting up the design part of the REP in such a way as to exclude appropriate criticism and selecting sub-contractors partly (and perhaps unconsciously) because of probable favorable tendencies. The next step was to have the agency sub-contract the evaluation. This is unsatisfactory for the reasons we have just described. The best arrangement is to have a separate agency in charge of evaluations, certainly cooperating with evaluation staff and liaison officers of both the project and the agency that’s funding the project: or at least a sub-agency.

Third moral: if projects cannot self-evaluate objectively and if the commercial evaluators are open to biases just mentioned and if the changes just mentioned have not yet occurred, it looks as if one will not be able to find good evaluators. There are two routes to go. The big shops like ETS, RAND, and SDC do have a degree of independence of any particular agency or officer and can afford to choose independence over back-scratching, at least part of the time, and they do have strong professional status needs as well as economic ones. The other route is exemplified by Briloff (1973) in the accounting field, i.e., by someone who has a permanent fulltime fall-back job which provides a perfectly acceptable alternative to contract work and one that is positively preferable to compromised contracting. It is not possible to conclude that the middle-size full-time shops are in fact less reliable, but it is harder for them to ignore illicit pressures. There are important trade-off advantages for them, however--efficiently manageable size, availability of university resources, flexibility of procedures, etc. Since the only real test of bias is error, and since some of these shops do run with a low error-rate, a consumer who is familiar with the track records might well pick a good midi-shop over the part-timer whose resources are limited or the big shop where there is considerable variability in staff quality. Nevertheless, we could do with some evaluators who are as beyond suspicion as organizational arrangements can make them. One might argue that Alan Post, the non-partisan Legislative Analyst for the state government in California, is one paradigm and the Supreme Court another. I have suggested to NIE that they should consider reviving a version of NIH’s Life Research Fellowship program for this purpose.

The fourth moral is that since the arguments under the third point bear closely on the present author's own role as an evaluator, they should be viewed with exceptional suspicion. Indeed, this is an essay on suspicion, since without it one cannot avoid serious contamination. But it is not an essay on the virtues of suspicion in itself. All suspicion can legitimately do is suggest possibilities against which one takes suitable but not absurd precautions, and the truth of which one subsequently investigates.

Negative Reactions to Bias Control Procedures

Given our cultural emphases, these systems of independent evaluation are likely to strike us as symptoms of distrust. Given a serious commitment to effective service, responsibility, or self-improvement, they would instead be seen as useful--or rather, essential--aids. Since it is a universal truth that self-evaluation is unlikely to be reliable, it is a necessary consequence of interest in truth that one supplement self-evaluation. Hence anyone interested in improving his or her own performance _must_ arrange for or endorse some kind of independent evaluation. Thinking about my own teaching or my own performance as an evaluator, I know that I need independent assessment of it, and I arrange it whenever funds can be obtained (which is essentially always, if one really tries). I use such feedback myself in the formative mode (when experimenting with alternative approaches) and expect it to be used by others summatively, that is, for judgment of my performance by my superiors or clients. It seems to me that a missing major goal in schools of education, and probably in all tertiary if not secondary education, is the affective goal of valuing justified criticism (which is not, of course, the same as enjoying it).

Long experience with lazy or corrupt supervisors in bureaucracies of all kinds makes it obvious that potentially effective systems of evaluation are open to all kinds of abuse and neglect. But the common labor-union (or professorial) response of refusal to participate in any such system is even less responsible since it rejects a legitimate demand instead of rejecting illegitimate abuses. A serious loss of credibility with the parent, voter, and/or taxpayer is a natural and appropriate result. Refusal to participate is, however, justifiable if _either_ of two considerations applies: first, that the proposed system is technically seriously inferior to another feasible and specifiable system with regard to which cooperation would be forthcoming (the inferiority to be judged by independent expert evaluators), or second, that a respectable system of independent mutual evaluation (of the administrative staff who commission or will conduct the evaluation) is simultaneously or earlier introduced. It should be noted that “technically inferior” is not contrasted with “morally inferior” (i.e., more likely to produce injustice) since it is a technical requirement that the system minimize injustice. The contrast is with “impressionistically inferior,” i.e., inferior in the view of unskilled personnel who react largely to perceptions of risks for them. A good evaluation system nearly always has to involve some moral elements, and its moral status requires it to weight the welfare of all people that it affects proportionately to their stake in the issue. That means it must weigh the rejection of outstanding job applicants in the balance against the retention of weak teachers, using the gains and losses for students and others affected (parents, employers) as additional currency. Morally speaking, too, it is outrageous that most educational systems which use administrators to evaluate teachers have nothing worthy of the name in the way of procedures for evaluating the administrators.

Efficiency, narrowly conceived, is not the only concern of evaluation systems. Indeed, it is entirely secondary to justice. And the cardinal principle of justice is that evaluators should be evaluated, a theme previously stressed but that deserves further explicit discussion in the following section. Its practical basis lies in that it follows directly from both principles already enunciated. The infrequency of its application is an illustration that evaluators are not much more attracted by tough self-evaluation than are their evaluees.

Metaevaluation

I have used this term to refer to the evaluation of evaluations or evaluators. Thomas Cook, in the most detailed study made of it so far, calls it--or a special case of it--secondary evaluation (Cook, 1974, pp. 155-222). Jim Sanders, in the only essay that I know of by another author on the topic, follows my usage (Sanders, 1973).[1] The term “secondary evaluation” suggests to me evaluation using secondary indicators, such as teaching style, instead of primary ones, such as learning gains. The term “metaevaluation” makes some sense to someone used to the academic terminology (metamathematics, metaphysics, metaphilosophy, meta-science, metapsychology, metaethics) but is for others an opaque neologism for which I apologize. In a sense this whole paper is a study in the methodology of metaevaluation. I will stress here a couple of particularly crucial points about what I would regard as standard operating procedures. The first arises from the requirement that evaluators should try to arrange that their own work be replicated, in whole or part, by other equally competent evaluators working independently. This is particularly appropriate where any non-standard methodology is involved or where particularly difficult synthesizing judgments of overall merit are

[1] (See also Stufflebeam [1975].)

involved. When this approach is used, it should not conclude with the submission: of the independent reports. Each evaluator or evaluation team should, after such submission, now critique the report of the other team and have the opportunity to submit a revised evaluation report involving such modifications as seem called for after reading the other report. In certain cases a combined report may be agreed upon, after a joint "convergence" meeting, a procedure Stufflebeam has encouraged.

A useful special case Of the preceding approach is the adversary arrangement, where one evaluator or team deliberately undertakes the task of making the very best possible case for the project, given the data, while another presents the case against. This was admirably done (on a micro-budget) in the TCITY evaluation by Stake and Denny (Stake and Gjerde, 1971, pp. 26-27;14). It caused trouble because defenders of the project felt it legitimated the negative comment. One would do better to discuss this mode of reporting with the evaluees and clients in advance to avoid unnecessary defensive reactions like this. Robert Wolf has recently extended this approach into the "legal model" of evaluation (Wolf, 1973).

The metaevaluations thus generated (as each team criticizes the other's evaluation) are very useful for the administrator-client. For they are the comments of two highly knowledgeable parties with a reputation on the line. Arranging a design that puts this kind of leverage on the evaluators is the moral equivalent of the pressure that the presence (or prospective presence) of an evaluator places on an evaluee, which has a certain natural justice, but it also provides, pragmatically speaking, a very substantial incentive for doing one's best. Goal-free evaluation, which I'll discuss in a moment, is a natural extension of this type of procedure.

The "double-teaming" procedures just described, besides their implicit recognition of the truth of the adage about sauce for the goose being sauce for the gander, are steps towards a scientific approach to evaluation in that they yield some data for calculating reliabilities. The approach applies equally well to the evaluation of proposals or personnel by panels/committees, indeed, it is a scandal that the big foundations, who dispense most of their funds through their peer review panel procedure, do not investigate the reliability of such panels, especially since there are a number of different ways in which panel reviews can be conducted, with the resulting probability of significantly different rankings.

The second suggestion I would stress is using the evaluees as metaevaluators. That is, the preliminary report from the evaluators should be made available to the evaluees for critical comment, and that comment-in raw form, or synthesized in a way acceptable to them--should go forward to the client along with the evaluator's original report and any modifications that the evaluator feels are appropriate in the light of this feedback. Guarantees that this uncensured response will be attached to the evaluation report will often have a favorable impact on openness to the evaluator at early stages.

The two preceding suggestions might also be taken as items for inclusion in a handbook of professional ethics. There are others besides the evaluees who might well be consulted as metaevaluators, for example, those whose resources are being used for the programs being evaluated. This proposal for “representation of the affected who are not involved” has a rather general application and essentially zero recognition. How many school board members are representatives of the childless community on whom the tax burden falls without any obvious returns? How many of the advisory panels for, say the National Park Service, include representatives of those who do not use the parks--but pay for them almost as heavily and might be interested in using them if their interests were provided for? Moreover, evaluators should look around carefully for people with special knowledge and interest in whatever is being evaluated, even if they do not qualify under the second suggestion above, that is, as evaluees.

Methodological Approaches to Bias Reduction

The reduction of bias in the sciences is normally achieved by the replacement of judgmental procedures by mensuration and calculation. To a considerable extent the same path can be followed in evaluation. In fact, the “calculations”--in this case, the statistics--are already pretty sophisticated, although their selection and interpretation still requires a good deal of judgment. Even there, the choice and significance of different statistics has been greatly standardized in recent years with increasing sophistication and advanced training. The problem is mainly with the qualitative framework of an evaluation, especially the elements in it that generate the value component of the conclusion. This means particularly the needs assessment, the comparative dimensions, and the costing.

I shall confine myself to a mention of four approaches that seem to me capable of having considerable effect in upgrading the objectivity of evaluation. First, there is the standardization or routinization of qualitative aspects of the procedures. A detailed study of scores of evaluations done during the last six years suggests that a great many of them (over 90%, at a guess) omit one or more considerations that are obviously relevant to the assessment of merit they are allegedly providing. The reasons for the omission are often ego-defensive or political. (For example, the failure to look at the comparative performance of critical competitors, essential if evaluation is to service purchase decisions and hard to avoid when responsible refunding is being considered.) But they are also often simple errors of oversight. Both kinds of omission can be reduced by using a standardized checklist approach, and I have been encouraged by the extent to which a suggested version of such a checklist was adopted in its first year after private circulation (Scriven, 1974, pp. 35-93). The orientation of that 13-point checklist and profile generator is towards pay-off evaluation. One developed by Maurice Eash and ERIE (Eash, 1969, pp. 18 24) is aimed more towards systematic product description and is naturally considerably more popular amongst producers. Both have legitimate uses, and both can no doubt be improved. ETS also has one with some special features (mine originated in some work with ETS on a product review contract). Some others have been proposed for special purposes, e.g., the CMAS (Curriculum Materials Analysis System) from SSEC, and the tremendously valuable checklist covering all the administrative aspects of an evaluation developed by Dan Stufflebeam (1974). The trend is there and, given support, can lead to very substantial upgrading of evaluation, especially of evaluations that should be fairly straightforward, but that often get bogged down in irrelevancies, or omit relevancies.

As an example of an irrelevancy, one sometimes hears the lament that we can't really evaluate educational products until we have an adequate theory of learning. This remark displays a total lack of understanding of the difference between evaluation and explanation. One needs great professional skill as a product evaluator to set up a valid assessment of color TV sets, but one needs to know nothing about electronics. On the other hand, to explain _why_ a particular set triumphed in the ratings _will_ require such knowledge--in fact, an extremely rare combination of theory, design, and production engineering skills. Theory may suggest breakthroughs in design, its contribution to _evaluation_ is at most that of supporting the use of certain secondary indicators as criteria for merit. Even there one needs only empirical correlations of those features with favorable evaluations. The checklist, like the trouble-shooting chart in the back of an appliance handbook, incorporates a massive amount of knowledge in a maximally task-oriented form; theories have the first, but not the second property.

But the improvement of evaluation is not the only pay-off from the checklist approach. I believe it has already produced significant improvements in products, for the producer is not only aware that the checklist may be--in some cases, will be--used in evaluating the product, and hence tries to meet the standards it expresses, but he or she is also (to a variable extent) interested in turning out a quality product and may find the arguments supporting the checklist persuasive in upgrading his or her conception of what that implies.

The second approach involves upgrading the _training_ procedures for evaluators, especially in the qualitative dimension. The simplest move would be to increase enormously the number of evaluations performed during the training period, perhaps to a hundred or more, with feedback in one form or another (such as tailored comment, programmed materials, or the issue of good and bad paradigm answers). Another procedure, which could be applied in modified form to the training of review panel members, involves a direct effort to achieve high inter-judge reliability without introducing correlated error, a procedure that I call _calibration_. This is an extension of the first procedure and involves using a basic set of cases, judging them independently, talking out differences as far as possible, testing on a new set, and so forth until reasonable convergence is obtained.

The third approach picks up where training leaves off, but focuses on the elimination of sources of bias external to the evaluator. We have already discussed some of these that arise from organizational and economic factors, the need for further contracts, for instance. We have also discussed interpersonal ties and argued for the use of external evaluators, at least in a supplementary role, for both formative and summative evaluation. Even when we had taken account of all the preceding suggestions, a type of biasing interaction occurs which has highly significant effects on the evaluator and needs to be dealt with. It has two dimensions which are, roughly speaking, affective and cognitive.

The affective influence occurs because of the generally submissive-obsequious hanging-on-every-word posture which it is difficult for an evaluee to avoid adopting towards the evaluator, especially if the latter is evaluating on behalf of the funding agency. This is somewhat too egogratifying for evaluators to suppose that it has no influence on them. “How can all these nice intelligent people who show their good taste by asking after my health and work so interestedly (and even, in formative situations, by selecting and paying me to do the evaluation), possibly not be doing something-truly worthwhile?” The best way to minimize this influence is by minimizing the social contact with the evaluee prior to submission of the preliminary version of the report. There is plenty of time for it later, during the interaction about the report, and then it is far less time-wasting for project staff. (Site-visit evaluations always have a disruption cost going against their utility.) In *reacting to* the draft evaluation, the evaluees have a focus for their activities and remarks, and the evaluator has a stake in the discussion so that a fruitful exchange can occur rather than a “show and tell” performance.

If one eliminates these prior social exchanges, how does the evaluator get briefed about the background, aims, and nature of the project? This question leads us to look at the cognitive biases that result from such a briefing. If one wants an unbiased view of what the project does, one would do better to talk to or, better, observe the *users,* not the producers. After all, whether formative or summative, a major function of evaluation is to look at the materials from the point of view of a prospective user. The user will not get a visiting fireman treatment. The user will not be concerned with *background* of the product or what it was *meant* to do, only with what it actually does. So the evaluator, in simulating the user’s viewpoint, does best to avoid all the “fringe benefits.”

Taking these considerations seriously leads one into doing goalfree evaluation (GFE). It is extremely important as a methodology for avoiding overfavorable evaluations and for detecting side-effects. Since one has not been told what the intended effects--goals--are, one works very hard to discover *any* effects, without the tunnel vision induced by a briefing about goals. If GFE sometimes errs in the direction of being too critical or missing a main effect, the cost of those errors is insignificant *because they can be picked up at the debriefing*. Putting it another way, the GFE mode is the best way to *begin* an evaluation because it is reversible without loss, whereas the GBE (goal-based) mode is not reversible and more likely to be biased.

One might describe GFE as a step beyond double-blind methodology. (Some of its critics would probably prefer to call it *totally* blind.) In double-blind drug studies, neither patient nor nurse and/or investigator knows which pill is the placebo and which is the experimental drug during the period of observation (which is when the bias would operate). The interest is to get the investigator to look just as carefully at all patients, without the kind of prejudice that might lead to projecting effects onto the group that got the experimental drug. And, of course, to ensure that the "treatment," which involves both a pill and its presentation, is equalized. The evidence about the effect of expectations on perceptions is so strong that an experimental design that does not blind the observer-investigator simply could not be taken seriously. In triple-blind, the investigator--who would now have to be different from the developer--would also not know what the intended effect was. He or she would have to discover what effect, if any, the administered substances had, from a study of patients' health, etc., through the period of drug administration, and thereafter.

Now what possible point could there be in such a procedure? Very simple: it will make the observer-evaluator struggle hard to find any and all effects, without prejudice, since his or her reputation is on the line, and the job has not been pre-defined. Reading a non-existent effect into the clinical picture, cued by inspiring messages from the research crew, is made less easy, missing a slight but crucial side-effect is made more difficult. Of course, the evaluator has access to the charts and medical history of each patient and it will often be easy to get an idea of the intended effect from these. But to make that idea precise, to describe the class of patients for which the effect appears to be such-and-such, especially given the absence of cues as to which received a placebo, will put the investigator on his or her mettle.

In the medical situation, the intended effects are relatively simple, the class of patients treated is a rather good indicator of the intended effect, and the consequences of reading non-existent effects into the data are considerably (but not entirely) mitigated by the double-blind situation. In education none of these considerations normally hold, the latter failing since double-blind studies are not generally possible. Consequently, the advantages of goal-based evaluation are particularly crucial there, whereas they may be only marginal in medical research. Apart from the methodological advantages of making the evaluator hunt for any effects and thereby reducing the chance of missing a side-effect, GFE provides yet another of the procedures for exerting accountability pressures on the evaluator in addition to those mentioned in the section on metaevaluation, and hence restricting the play of bias. There are more detailed discussions of GEE in House (1973) and Popham (1974).

Finally, it is well worth mentioning the advocate team approach for generating alternative plans, which can then be comparatively evaluated. This has been particularly carefully studied and developed by Dan Stufflebeam's staff, especially by Diane Reinhard (1973) who applied the emphasis on independence stressed earlier in talking about feedback channels to input. One notices a deficiency in this dimension of evaluation not only where complex plans are involved (the area where adversary methodology has been focused) but also in simple product evaluation where some ingenuity may be required to identify the appropriate alternatives. For example, the evaluation of CAT (computer assisted instruction) should normally involve comparison with programmed texts using the program content from the computer, since these can be produced for a minute fraction of the CAT costs, are portable, and simultaneously usable by many students.

Conclusions

An effort has been made to review a wide range of sources of bias in evaluation, and preventative measures for them. The resulting recommendations, taken in toto, provide a fairly comprehensive set of guidelines for setting up the broad outlines of an evaluation system.

Two normative principles were formulated, the first recommending independent feedback in evaluation, the second requiring regular review of the independence. A third principle is inherent in much of the later discussions of practical procedures, it asserts that the best guarantees of independence are ignorance and countervailing bias. There are no wholly unbiased evaluators but there are _arrangements_ which discourage them from bringing (some of their most damaging) biases to bear, _or_ where their biases are (at least partially) balanced off. The search for the pure in heart is more appropriate for mythology than methodology. We can _arrange_ for jurisprudence when we can’t _find_ it; it can be a property of a group of evaluators, even when it is a property of none of them. It’s a matter of balancing off, not perfect stability. We could call this the Principle of Independence as Dynamic Equilibrium, following our practice of grand titles for grim truths. When we want valid independent evaluation, we don’t use the driver-educator to evaluate the driver-educator, but we use one driver-eductor and one Latinist, or both in one, and that’s better even than an accountancy instructor (the implications for evaluating ethnic studies programs are obvious and possibly more exciting). To evaluate breeder reactors we use someone from the Sierra Club _and_ a member of Congress, not a retired judge of the Supreme Court. (When we’re concerned with credibility rather than validity, we pick the judge and require that the judge hear the others _and_ that a summary of their briefs be attached to the evaluation report.) Or, to evaluate a new drug, we use researchers who aren’t told what the drug is supposed to do. In short, fight fire with fire or with oxygen starvation, not by trying to make everything out of incombustible materials.

The Principles tell us that independence is essential, impermanent, and situational. Of course, one might say, we all knew that. But then why didn’t we value the knowledge enough to use it? Perhaps because we also knew, or thought we knew, the opposite; that independent advice is a luxury, or that it can be provided by a proper organizational arrangement of supervisors, or that it can only be obtained from really disinterested people. Knowing contradictory truisms about bias and its control is knowing nothing about it.

References

Apple, Michael W., Subkoviak, Michael J., and Lufler, Henry S., Jr. (Eds.). Education evaluation: Analysis and resDonsibilitv. Berkeley, California: McCutchan Publishing Co., 1974.

Briloff, Abraham. Unaccountable accounting. New York: Harper and Row, 1973.

Cook, Thomas. The potential and limitations of secondary evaluation. In Michael W. Apple, Michael J. Subkoviak, and Henry S. Lufler, Jr. (Eds.), Education evaluation: Analysis and responsibility. Berkeley, California: McCutchan Publishing Co., 1974.

Eash, Maurice J. Assessing curriculum materials: A preliminary instrument. Educational product report, February 1969, 2 (5), pp. 18-24.

House, Ernest R. (Ed.). School evaluation: the politics and process. Berkeley, California: McCutchan Publishing Co., 1973.

Popham, James W. (Ed.). Evaluation in education: Current application. Berkeley, California: McCutchan Publishing Co., 1974.

Reinhard, Diane L. Methodology development for input evaluation using advocate and design teams. Unpublished doctoral dissertation, Ohio State University, 1972.

Sanders, James. Untitled materials prepared for instruction at Indiana University, Bloomington, Ind., 1973.

Scriven, Michael. Evaluation perspectives and procedures. In James W. Popham (Ed.), Evaluation in education: Current application. Berkeley, California: McCutchan Publishing Co., 1974.

Stake, Robert and Gjerde, Craig. An evaluation of TCITY: The twin city institute for talented youth. University of Illinois: the Center for Instructional Research and Curriculum Evaluation, 1971.

Stufflebeam, Daniel L. Meta-evaluation. The Evaluation Center Occasional Paper Series, 3, Western Michigan University, May 1975.

Wolf, Robert L. The application of select legal concepts to education evaluation. Unpublished doctoral dissertation, University of Illinois, 1973.