DOCUMENT RESUME

ED 079 394 TM 002 994

AUTHOR Scriven, Michael
TITLE The Evaluation of Educational Goals, Instructional Procedures and Outcomes or The Iceman Cometh.
NOTE 33p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Cost Effectiveness; Educational Needs; Educational Objectives; *Evaluation; *Evaluation Methods; *Evaluation Techniques; Formative Evaluation; *Models; Summative Evaluation

ABSTRACT

A model checklist conceptualizing the evaluation process is presented and discussed. It is quite general and is intended to apply to the evaluation of educational products, procedures and most outcomes. The Pathway Comparison Model consists of the following: (1) characterization--how generally or specifically to describe the "treatment"; (2) clarification of conclusion with client--award of merit, best buy, etc.; (3) causation--Does it enter? How is it to be handled?; (4) comprehensive check of consequences; (5) conceptualization--compression typically using preceding data but may use some from steps 6-8; (6) costs--including disruption, etc., and the costs of the evaluation; (7) consumer characteristics--market and need analysis, covers consumers for the product and the evaluation; (8) critical competitors (real, ideal, etc.--repeat 1-7 for them); (9) credentialing--combining; and (10) conclusions and communications--data processing, design, writing, and dissemination. A detailed checklist for product evaluation is appended. (KM)

Michael Scriven

THE EVALUATION OF ~~ETCETERA~~ Educational Goals, Instructional Procedures + Outcomes

OR THE ICEMAN COMETH

0. Foreword

So far, you've had all the good news; here comes the bad. You've planned and charted, picked daintily from the delicious smorgasbord of spicy objects laid out by acronymic caterers, meditated about great goals, urgent needs and exotic philosophies--and eventually you may even have done something. A great trip, but the day of reckoning cannot be postponed forever. As they say, fly now and pay later. Enter Evaluation, the great deflator, the destroyer of dreams, the last trumpet---or perhaps, on a different view, the last strumpet, the whore of the establishment, the Great Seal of superficial inspection.

The crucial question is whether we have any real standards of objectivity in evaluation, or whether it's a mutual back-slapping--or back-biting--exercise. If we take a close look at the "interlocking directorates" situation in evaluation, we can become very nervous about objectivity. There are not very many evaluators carrying the responsibility for evaluating the big federal programs. And they are often called in to write proposals, to judge them, to judge the resulting projects for the project manager, to help project staff improve their work, to judge their product for the funding agency, &c. And they are often themselves producers and managers of competing products. This complex situation cannot avoid producing some conflicts of interest.

Again, there are problems about the stupefying constraints on the resources available for evaluation, resulting in necessarily superficial reports. In the light of these weaknesses, is there really anything left that's worth having?

The nice feature of evaluation is that, like hope, it springs eternal in the humane breast. What kind of question is the question whether evaluation is worth having?

It is of course a question which can only be answered competently by an evaluator; indeed, anyone who did answer it competently would by definition be an evaluator. Moreover, the answer must be affirmative since the question itself is of great importance and hence its answer (which, as we have seen, is itself an evaluation) is worth having. Less trickily, rationality and responsibility require that we always obtain the best answer we can get to questions of the form "Is X worth doing?" before we commit public resources to X. So, one can no more evade evaluation than one can evade philosophy--all one can do is avoid discussing it openly and critically. And since open and critical discussion is about the best way we know to decrease bias and increase the scientific status of a practice, such evasion would be a great mistake. Evaluation needs evaluation to keep it honest, it needs new methods to keep it flexible; but even if you don't like it , you can't leave it.

This paper pursues a course aimed squarely at the improvement of the objectivity of evaluation, a simple course but not the one usually followed by those with the same goal. The usual conception of improving objectivity involves a simplistic and long-outmoded idea of what science has to be like. Not that the conception is inappropriate for some sciences, say, mathematical physics; but it just isn't appropriate for much else. Messy sciences, and especially applied sciences (including applied physics), actually depend less on exact mathematical formulae--though they may use them--than they do on rough models, convenient approximations, checklists and trained judgment. Very often, in fact, one can extract from the trained judgments the cues to which the judge is responding, and these provide us with a checklist that can be used to make the implicit inferences explicit and thus take a significant step towards objectification. And it is this path--so characteristic of trouble-shooting procedures in electronics or medical diagnosis; in criminology and taxonomy--that I'm undertaking today.

But the checklist approach that follows is not the most practical kind of checklist one can give--it is one aimed at conceptualizing the evaluation process, not at the details of a particular kind of evaluation. I have worked up a detailed checklist of the

latter kind for product evaluation and published it elsewhere--I'll add it as an appendix to this paper. I have also almost completed one for evaluating teachers, and next year will work on one for evaluating student work more usefully than is commonly done. But the model presented here is what underlies these practical applications. It is not as simple an account to read as I thought it would be while setting it down; but it does convey a supposedly comprehensive coverage of the evaluation process that we all apply informally when we pass judgments of merit on education-related entities.

This general model of the evaluation process applies without special modification to the evaluation of educational products, procedures and most outcomes. I shall add to this (at the end) some further comments on the evaluation of goals.

1. The Pathway Model of Evaluation--the Basic Perspective

Conceive of evaluation as an information-processing activity. It begins with observations on data and it finishes with an evaluation, i.e., a judgment of merit. Typically this process involves a vast amount of condensation, and it is useful to see evaluation as a sequence of data-compression steps. The extreme case is grading a quiz or term-paper; we begin with the raw data of student responses, perhaps 6000 words. We conclude with a single letter grade. Along the way, since we are usually involved in a teaching activity and not just an evaluating one, we probably put down a good many words of advice and criticism. But our judgment or overall merit, i.e., our evaluation, is sometimes important, and sometimes very legitimately and usefully expressed by a single letter.

The process of inference intervening between our perception of the performance and our evaluation nearly always involves (or can be usefully reconstructed as involving) some intermediate steps. We may, for example, fragment the original performance, evaluate each part against discrete standards, and assemble the results thus:

Dimension	Percentage of maximum possible score(by this student)	Weighting of this dimension in total score	Weighted Score (by this student)	4. Maximum Possible Score
Originality	45%	2	90	200
Clarity	70%	1	70	100
Coverage	30%	2	60	200
			220	500

And we may (for other reasons) consider 220/500 to be about the minimum satisfactory passing level, i.e., a grade of C-. There are two distinct sub-processes here. The use of the marking schema conceptualizes the performance (this involves both devising an appropriate taxonomy and measuring the specific performance in terms of the dimensions of the taxonomy). Then the grading sub-process applies a value-label to the performance as conceptualized; this sub-process I call credentialing the performance. Sometimes, of course, grading is done off a "curve", in which case it appears at first sight to be only an example of a further conceptualizing step, since it leaves unanswered the question of real merit, telling us only about relative performance--a very different issue. It isn't particularly easy to justify an "absolute" A, but it's certainly worse to assume that the top mark in a badly taught and incompletely examined mickey-mouse course where cheating is common and the content trivial represents an educational achievement of high quality, which is the very least an A signifies.

We should by now have buried the arid positivism of "value-free social science", according to which grading on a curve was the only legitimate procedure. Curiously enough, such sceptics were never consistent enough to recognize that they were distinguishing the top 15% or 20% from the bottom segment, i.e., they were making a judgment of absolute merit within the curve system. If that judgment was defensible, ther there is certainly nothing qualitatively different about the judgment that this quarter's exam was rather more difficult than usual, that the TAs were confused about a critical issue, that the class definitely worked harder than usual, that the evidence of better talent selection by the college is overwhelming--in short, that more than 15% deserve As this time. One might say that absolute merit is just merit relative to all significant relevant comparison groups, not just to the handiest

or the one where quantitative scaling is possible. So--grading (even on the curve) goes beyond the value-free description of performance; it involves credentialing.

Taking a rather different example, we might be studying a remedial reading program and here we might colligate data, compare raw (mean) scores with another kind of intermediate criterion (to help us conceptualize the achievement) such as national average reading scores at a given grade level, and make a further (credentialing) step to conclusions of merit of the results, thus;

Data----------> Mean gain of 1.5 years against national norms ----------> Highly meritorious program

Even in the case of an instant "global" evaluation response, e.g., to a short essay answer, the evaluator will usually be able to give reasons for his judgment when pressed, and we can reconstruct the process of evaluation from these reasons as involving intermediate (conceptualizing) criteria.

A popular candidate for one of the intervening stepping stones in the "inference pathway" to our evaluative conclusion is the goals of the project;

Data----------> 95% success in goal achievement ----------> Meritorious project

The simplest pathway model thus involves two steps of data transformation, the first of which does not yield explicit judgments of merit, the second of which does. But the first is so chosen as to make the second possible, just as, when picking a pathway through scrub or across a stream, one selects the next step on grounds of its promise for reaching one's eventual goal as well as for its immediate accesibility. In designing or critiquing an evaluation it is quite useful--and relatively unusual--to keep the necessity for completing such a pathway in mind. The initial step(s) or conceptualizing steps, have the main function of enabling us to get a "grasp" of the data; but the final steps answer the questions that are important to us in evaluating educational performances (as opposed to doing "pure" research).

The developer or teacher has always got one conceptualization of the data in mind: if he (or she) feels he's been successful, he or she, naturally sees the data as "demonstrating success in achieving such and such goals", and hence (since those goals would not have been adopted they were felt to have merit) the project is judged meritorious. But there are many other ways to see most projects. It's just as important for the evaluator, as opposed to the developer, to retain an open mind about the legitimacy of radically different interpretations of data, as it is for the scientist reading a research paper in which the author proposes that certain experimental results support his theory. It may be best to start an evaluation without hearing about the goals. This methodology of "goal-free evaluation" is a procedure for preserving that openness of mind; and it could in fact be transferred to the more common scientific research context, though as far as I know it has not been attempted there. Looking at the project with the eye of experience, unbiased by a pre-formed goal-based conceptualization, one is more likely to notice important effects that were not intended, and form a quite different conception around these, with quite different potentiality for credentials. The goal-based evaluation will be inferior in this case because it has overlooked an important part of the picture ("side effects", from its point of view).

The steps which lead from the conceptualized (or criterion-referenced) intermediate conclusions to the eventual conclusions involving judgments of worth or merit or value are the ones I call credentialing steps. Because these are often not made explicit, if considered at all, the next section takes up one instance in modest detail.

2. The Credentialing Steps

In evaluating the impact of busing, for example, one is likely to appeal to parameters such as percentage racial mix on each campus as criteria. It is easy to transform/condense the primary data into these terms, so it's a workable first step on the pathway. But how do you show that that specific achievement is merit-

orious? Look ahead; what further stepping stone would get us nearer to an evaluative conclusion? "If the school population is integrated, the students will be more likely to..." what? We need to fill in the space with some behavior which is either obviously or demonstrably a desirable outcome. Usually the choice is that you can either go for the big money, on a weak research basis, or for a small prize on a better foundation.

Thus; there's a small chance that the students will be more likely to treat others as equals without regard to color and that would be meritorious, since it's both a constitutional and a moral obligation in many circumstances. There is a larger chance that the students will be involved in some kind of social interaction; but it's not so easy to show that that is a merit pay-off. Other things being equal, it may have some value--but then other things aren't equal because there were heavy costs involved, both in busing itself and in the break-up of ability-grouping, SES-bonds, &c., all of which--other things being equal--have their own merits. But not as much merit? An abstract description of the dimensions would suggest this, e.g., "social egalitarianism is better than academic achievement increments." And that's the way the point is likely to be put in the heat of argument. But the real question is whether we are in fact getting a substantial specific gain in democratic behavior that offsets the specific costs. Isn't any gain on such a crucial variable worth far more than this magnitude of costs? No--for two reasons. First, the gains may be real but sub-threshold for social action changes off-campus. For example, there may be significant affect changes showing up on projective tests, but absloutely no overall change in the ex-student's or off-campus student's choice of work, dates, emplyees, charities, political candidates, loan applicants, employers &c. In that case--on this evidence alone--busing is unlikely to be worth what it costs particularly because of the next point.

The second point is that costs include opportunity costs; the busing money could have been spent in many other ways aimed at the same goals, e.g.--to look at the broadest decision-space--for the administrative costs involved in getting and using federal housing funds to convert the school districts into integrated neighborhoods, or in

subsidizing social service enterprises by integrated student teams, or by alloting black teachers and principals to white schools and classrooms, or by setting up integrated tours, visits, games, expeditions, camps &c.

So the credentialing steps in the pathway model usually need some detailed support and often involve an application of some aspects of social, moral or political theory. And they essentially always involve a comparison of actual with possible pathways. For this reason, I usually refer to this approach as the "pathway comparison model."

The conceptualizing steps rest on, and if challenged require, substantiation in terms of statistical theory, or experimental design, or tests and measurement theory &c., on the psychological side; and on accounting/systems analysis procedures, on the cost side.

These conceptualizing steps often refer to objectives or norms or mean increments, which we can call criteria. The concluding steps are then the ones explicitly aimed at establishing merit, or credentialing (the criteria). In criterion-referenced testing we see a clear example of this; the criteria are so chosen as to admit of easy credentialing. Hence evaluation is often easier when results of such tests are available than when only norm-referenced instruments were used.

3. Flow-chart Loops

As one seeks a total evaluation pathway, both kinds of steps spin-off further questions or data needs; e.g., one sees that one could express the gains in terms of national norms, but the significance of that will be controlled by baseline data on gains by this grade in this school in previous years--do we have that data? If we do, a useful conceptualization may be possible, i.e., one that is nearer to representing the actual achievement of the new program. Or we may look at the conceptualization we have done and see that we can establish merit for a childcare center as long as there isn't a problem about increasing the amount or extent of conformist behavior. Do we have some data that will rule out serious effects in that dimension? The conceptualization

throws such deficiencies into relief. In designing evaluation, we arrange to get the answer; in deriving an evaluative conclusion, ex post facto, we look for that data in what we have, and in critiquing an evaluation (meta-evaluation) we check to see if the loophole has been spotted and filled.

It is important to keep in mind that evaluation (when the data is already in) is simply one kind of data-interpretation or data-transforming. There is indeed one kind of scientific evaluation, not the educational kind, where this is very clear, as in questions like "Evaluate this theory or hypothesis in the light of such and such data." A certain framework is being given, in terms of which the significance of the data is to be expressed. Educational evaluation is logically quite like this. It involves relating the data to a framework of needs, wants, and alternatives, and expressing the relationship in the appropriate language, which is that of merit.

The educational evaluator often has to discover much of that framework--it is implicit in a particular context. The scientist, on the other hand, ususally works in a very standardized context when evaluating theories and hypotheses &c. The use of means to represent data, for example, will often lead to a point in a pathway from which one cannot reach the most important evaluation conclusions (which may depend on differences in variability between two treatments). The credentialing step absolutely depends on the contextual framework of decision-spectra, needs, &c.; and the evaluation represents a succinct analysis of the relation between that framework and this data--the evaluator squeezes a trickle of good wine out of the mass of grapes using the skills of analysis and the framework of the context.

Evaluating theories--the scientist's task--also leads to judgments of merit; the kind of merit is different in the two cases, but just as the pure scientist is inescapably involved in judging the merit of theories, so the applied scientist is involved in judging devices, processes and products. These evaluations can be both judged, when that support is available, to be themselves factual claims. Evaluations are just as scientific as descriptions, explanations, and predictions, when properly done--and no more and

no less debatable. There is no need to argue here about the ultimate objectivity of morality--a very special type of value framework. We can regard ourselves as having completed the evaluation pathway if we can get to a firm footing on the Constitution, Bill of Rights, and the few matters of common moral agreement between the major moral systems. Much educational evaluation involves no debatable moral issues at all--but it still involves judgments of worth or merit, which require support, just as do those of the pure mathematician or the physician.

The pathway model, in broad outline, thus involves taking a series of pre-planned steps from data to criteria--the conceptualizing steps--and from criteria to evaluative conclusion--the credentialing steps. It is now time to look at some refinements that are often important.

4. The Conceptualizing Steps--First, Characterizing

What is it that is being evaluated? Whatever it is, it can be described at several levels of generality and the evaluation process is affected by the level selected. We might legitimately say, of a particular job, that it consists in evaluating:

a) CAI (computer-assisted instruction)

b) a particular instance of the use of CAI

c) a CAI math program

d) CAI for ninth grade algebra

e) Suppes' use of CAI for teaching algebra to NYC disadvantaged ninth-graders in 1969

f) this use of CAI by these teachers in these classrooms; and so on.

If you are evaluating what's happening as an instance of CAI (e.g., (b) above), then you'd better put some work into finding out the extent to which CAI produced the results observed, by contrast with teachers inspired by CAI, the curriculum content and sequence, &c. If you're down near the ostensive (highly specific) end of the scale (e.g., (f) above), then you can forget those contrasts and simply evaluate, for example, the performance of these students by comparison with comparable others whose class-

room contains no computer terminal(s). You no longer need to fractionate the effect. The line between evaluating the effects of x and discovering which parts of x produce which effects is never sharp, but there is certainly a complete difference between the extreme cases; the second question is simply a research question. It is confusing and costly for the client if the evaluator strays over into the research area when it is not necessary. The first moment at which an awareness of this point affects the evaluator is in the characterization of the problem--what is it that he is supposed to be evaluating? The whole "problem of implementation" comes in here, and of course it's all one part of the general problem of correctly describing the sample, the population, and hence the legitimate generalization.

Again, much of the confusion about the role of Hawthrone effect with respect to evaluation starts with the characterization point. If you're evaluating CAI, as such, via this installation (and presumably others), you need to discount for Hawthorne effect, because your implied comparison, in the evaluation, is with other methodologies of instruction. You're evaluating CAI, and its competitors are ETV, CCTV, PTs &c. If you're trying to decide whether good things happened in these classrooms, which are distinguished by the introduction of CAI, then the implied comparison is with other standard-type classrooms (or with these very classrooms, if no innovation had occurred). And in that case, it's very important to include the Hawthorne effect. Not to misdescribe it, as something unique to CAI, but as something which in fact came along with CAI. I recall a superintendent saying to me recently, "I don't care whether it's the Hawthorne effect or not; my program of innovations is bringing in significant gains year after year and I want this recognized as gains, not treated as if it was spurious or incidental."

The characterization step usually determines the immensely costly issue of classroom monitoring. If you need to know whether CAI caused the good results you get on achievement tests (or got a fair trial if there weren't any such results) you need to know (a) if the students really used the terminals, and for how long; (b) what else

was going on in these classrooms by contrast with comparable non-CAI classrooms. If your client only needs to know whether there were good results from introducing CAI, you need never cross the classroom threshold (except to look at the moral dimension of the process, and perhaps its pleasure giving tendencies). A drug evaluation is not the same as a program of research aimed at finding the beneficial ingredients--it comes first. But should drug evaluation guard against the placebo effect, which is the counterpart of the Hawthorne effect? Only if you want to compare it with other new drugs. If the comparison is with no-drug, it's proper to include the placebo effect because the most important question is whether the treatment has benefitted the patient.

One might say that a useful characterization of whatever it is that is to be evaluated would include some specification of the implied or important comparisons. It is partly because I think such comparisons are implicitly present in all characterizations, and hence crucial to the design of the evalution, that I view all evaluation (at least implicitly) as comparative. Not just for pragmatic but for fundamental logical reasons, of which we have now mentioned two. Further practical implications of this will be developed at various points below

The CIPP-PDK model of evaluation (see, e.g., Educational Evaluation for Decision Making, D. Stufflebeam at alia, 1971) comes to a strongly overlapping but not identical position by a very different route. Focussing on the practical use of evaluation, they urge early clarification of the choices that will have to be made by the decision-maker, choices which the evaluation should assist. This quickly introduces comparisons, often ones which the decision-maker had not previously recognized and which can prove most helpful. But evaluation is not essentially tied to future decisions, and historical evaluation (which is not so tied) is logically just the same kind of process as the more formative kind PDK is discussing. I am here suggesting ways to support the "essentially comparitive" thesis about evaluation that will apply even when future choices are not

known to be affected. And the comparisons I identify are not always the same ones that do occur later, when there are such later choices. But the emphasis of PDK seems to me extremely healthy in most educational contexts and serves to support the truly central role of comparisons in almost every phase of evaluation.

One way of putting some of the preceding discussion that has intuitive appeal consists in stressing the necessity for early identification of the exact type of evauative conclusion for which one is aiming. The different types call for different designs, of course, and often lead to or from different characterizations. Some of the principal types, with comments, are set out in the next section.

5. Types of Evaluative Conclusions

Note that this taxonomy bears on each of the three principal paper-and-pencil modes in which the evaluator works -- in designing an evaluation; deriving one given the data (whether self-collected or not); or critiquing one (meta-evaluation). For designing an evaluation should involve anticipatory role-playing of the other two modes, deriving one involves anticipating the critic, and critiquing involves role-playing the designer and deriver. (The internal reciprocity of these roles rests on the ultimate logical unity of the critical and creative skills in the cognitive domain -- you can't create anything of merit unless you can distinguish merit from masquerade -- the skill of the critic -- and you can only criticise well by inventing alternative legitimate but unanticipated interpretations of experience/data -- the skill of the creator.)

Evaluation Type	Usual Verbal Expression	Usually Adequate Premises
Pre-evaluative (Goal or criterion achievement)	"The treatment X had the effect Y on the population of students, S, in conditions C; and Y was the goal or shows that the goal was achieved"	1. X was the treatment. 2. X caused Y. 3. Y implies that the goal was achieved.
Minimal evaluative	"X had a good effect (on S in C)"	1. X caused Y. 2. S or non-Ss {desired* / enjoyed* / were benefitted by} Y

* These premises give prima facie support for evaluative conclusions, not deductive support, but evaluation--like science-- only needs prima facie inference.

Overall evaluative	"X had an overall good effect"	1. As for minimal, plus 2. Y had no harmful effects on Ss or non-Ss, or 3. Y had much less significant harmful effects on Ss
Commendatory	"X was worth doing"	1. As for overall, plus 2. The cost of X was manageable 3. Y was worth the cost.
(This is almost a sub-case of overall evaluation, if costs are taken as a harmful effect.)		
Laudatory	"X was the best choice"	1. As for commendatory, plus 2. No other treatment, on data which was available, appeared as cost-effective.
Ideal	"X was the best possible treatment"	1. As for commendatory, plus 2. No other treatment was in fact as cost-effective.
Best-Buy	"X was a Best Buy"	1. As for commendatory, plus 2. X is a member of a group which offers the best or almost the best performance for significantly less cost than their performance-peers.

Note: "Cost-effectiveness" is a concept essentially lacking in, though suggestive of precision. I use it here simply as a mnemonic term, so that "equally cost-effective" means something like, "Equally effective in a given cost-range, or (almost as)/(more) effective in a (lower)/(higher) cost-range if the (decrease)/(increase) in effectiveness is deemed to (be far outweighed by)/(far outweigh) the cost difference". In these terms, "Best Buy" means "maximally cost-effective", and of course involves strong assumptions about the marginal utility of a dollar, at this cost-level; assumptions which Consumer Research, Inc. -- by contrast with Consumers' Union -- reject as too limited in applicability to justify using the concept in their rating system.

6. The Causal Step

In almost all evaluation, X is evaluated by looking at its effects. Hence most evaluations involve, as one component, determining what the effects of X

were (with respect to a certain range of variables of interest) or, at least, determining whether certain effects are effects of X. (For this reason -- one of several -- it is naive to suppose that evaluation is somehow less than or wholly different from scientific research, despite its omission from the usual lists and publications.) The difficulties with causal investigation in the educational context are well-known. It is worth stressing here that even purely causal conclusions are almost impossible without comparative data, either from classical control-group methodology, or from quasi-experimental design or from highly theoretical speculation about what would have happened if X had not been present. So once more, the comparative dimension emerges. There is a considerable 'conventional' (i.e. contextual) element in what we select as the appropriate comparisons for causal research on X, an element that is related to the comparisons that turn out to be important in the very characterization of X. What are the effects of intensive pre-school language arts tutoring on K-12 performance? The question cannot be answered without more specification of the implied comparison. One might think the answer obvious; the ideal control group would be pre-schoolers without language-arts tutoring. That will indeed give one possible answer; but what is actually needed may imply a different control group, viz. intensive preschool supplementation of the linguistic environment by non-tutorial methods. Or intensive in-school tutoring, &c. If someone asks you, as a social scientist, to answer the question, "What are the effects of sex?", you would ask for further specification ("What kind of effects? What kind of sex?") before beginning a finite answer or research project. In fact, almost all causal inquiries are like that one to a greater extent than we realize, often until well into a project. The evaluator, like any applied scientist, must be especially aware of this since it is no excuse for him that all facts are equal in the sight of Truth. They are not equally useful to either the client or society. Clarifying the question the evaluator faces may involve extensive discussion of alternative characterizations of the treatment and of alternative bases for the causal claims that are likely to be involved in the evaluation.

Selection and implementation of a design will often depend on still further discussions of taboos, costs and ethics. But the plain fact is that the classical experimental study is always the "method of choice", to be abandoned only after earnest struggle. (See Tatsuoka for an excellent methodological defense of this point; also P.E. Meehl reference and recent--fugitive document--remarks by Mosteller). It is true that there are excellent alternatives, if we have to go to them; quasi-experimental designs, especially interrupted time-series (Glass reference) and a procedure I call "elimination analysis" (a formalization of the procedure of the detective and the historian, using (a) exhaustive lists of possible causes, (b) "presence checks" and then (c) modus operandi pattern matching); one should perhaps add what is called "pathway analysis",though its practical utility is not yet clear.

In beginning this section, I said that evaluation of X almost always involves looking for and then at X's effects. An apparent exception is that species of process evaluation where one is looking at the moral qualities of the treatment. Typically, however, even this requires that one ascertain whether the observed qualities of the process (e.g., the avoidance of unnecessary verbal cruelty in dealing with the students) are really part of or an effect of X. This question will sometimes be answered without causal inferences if one has a rather clean characterization of exactly what is to count as X (section 4 above); but it is easy to see that what looks cruel to the observer may not seem so to the student, and hence that our main concern may have to be with the real effects of the treatment, not--as we thought at first--with its "intrinsic nature." But one could also say that this is a case where the intrinsic nature is being evaluated but must be inferred, is not directly observable. The line between evaluating X per se and evaluating the effects of X is not a conceptually sharp one (cf. "evaluating" a painting). Of course, most process evaluation is secondary or mediated evaluation, i.e., it consists in observing factors which are supposed to be connected with merit via some (usually dubious) theory. That is, most "process evaluation"

is really conceptualizing, not credentialing; or else one must take it to be unsound evaluation.

A type of process evaluation that is often thought to be legitimate consists in evaluating the content of texts/ lectures, for evidence of contemporaneity, errors &c. This is sometimes justifiable, but often involves the error of confusing the medium with the message. The crucial question here is what the student learns and retains, i.e., the effects of the content. Most elementary physics texts are full of falsehoods, but what's learnt may still be more valuable than what would be learnt from a text with the oversimplifications replaced by a mass of detailed corrections.

7. Criteria as a Device for Conceptualizing

For the evaluator, criteria are the standards or sets of categories in terms of which he or she conceptualizes the raw data, selected both for their prospective efficacy in expressing/compressing the data and for their promise for (or guarantee of) credentialing, i.e., for demonstrable connection with an evaluative conclusion. The use of the term "criterion" in the phrases "criterion-referenced tests" or "criterion behavior" is consistent with the use just suggested. "Behavioral objectives" are also criteria in this sense and so are many other goal-descriptions. Sometimes these non-behavioral criteria have the (attempted) credentialing built in, e.g., when we talk about the goals of a program as "improving computational skills by bringing them up to grade level."

It is often necessary for the raw data to be fragmented ("dimensioned") and each portion simultaneously conceptualized. Independently there may be a need for several successive conceptualizing ("boiling-down") stages or steps.

Once more, the key perspective is that of contrasts; the evaluator must seek and consider competing conceptualizations, i.e., those which appear equally legitimate as inferences but yield incompatible representations (suggested portrayals) of the results. ("You can say you've made a mean gain of 1.5 grade-equivalents. But you could also say they've gained far less than any preceding or comparable class in this school.")

Of the inferential steps between raw data and evaluative conclusions, the earlier ones normally instantiate principles of educational psychology, statistics, and theories of management, the later those of value-theory. But conceptualizing vs. credentialing is not facts vs. values. Sometimes the data are themselves evaluations (e.g., grades on various tests); but we can still distinguish conceptualizing (e.g., calculating average grades in various subject groups) from credentialing (recommending admission to Harvard Law School).

8. Costs, Audits and Accountability

Taking "costs" for the moment, to exclude opportunity costs (see #6), some of the dimensions of the contrasts that are important include installation vs. depreciation vs. maintenance, total vs. immediate, direct vs. indirect, dollar vs. psychic, materials vs. salary, hardware vs. softward, man-hours vs. machine-hours, per-student vs. per-subject vs. per-school, externally fundable vs. internally fundable, original vs. replication, deductible vs. gross, development vs. marketing. Which of these, or other, breakdowns are important depends on the particular problems of the client or community.

Once more, the perspicuous analysis of costs is very much a matter of selecting the most useful contrasts. It is a favorite aphorism in the accountancy end of evaluation that "There is no such thing as the cost of anythings" which is enlighteningly related to the corresponding remarks about "the cause" or "the effects" or "the correct description." Each should be interpreted as symptomatizing the need for very detailed contextual specifications before precise answers are possible

As in the case of the goal-free approach to conceptualizing, it is desirable if the evaluator can set up his or her own cost-categories before seeing those of the project accountants: it increases the chance of spotting some previously overlooked category or perspective.

It is difficult to convey to the average evaluation client--or indeed to most of one's

colleagues--the extent of the subjectivity in costing. If one can persuade them to read a book, then Unaccountable Accounting by Abraham Briloff, Harper & Row, 1972, usually produces the equivalent of religious conversion. The book can be summed up as proving that "generally accepted accounting procedures" often allow the same situation to be expressed as immensely profitable or completely disastrous, depending entirely on the accountant's preference; and by "shopping for an accountant", management has exactly this option in describing their own performance or that of a subsidiary they wish to drop, or an acquisition they favor. Nor is this a matter of selecting a shady operator, as Briloff's story about the Big Eight illustrates. It is not accidentally related to the appearance of Briloff's book that we have just seen an interesting occurrence of the opposite kind. In a case which will go down in history as the Dunking Donuts case, Price Waterhouse (another of the Big Eight) refused to go along with the company accountants on their procedure for handling interest. Dunking Donuts finally agreed--and shortly afterwards, fired Price Waterhouse. It was an expensive stand on principle, and the indirect costs (of nervous executives not hiring a firm with principles) may far outweigh the loss of a five-figure account. But Henry Hill, the senior partner of Price Waterhouse in charge of the case, was so obviously right, and Dunking Donuts so obviously using a dubious procedure to inflate earnings,* that the implicit point of this story is still as cynical as Briloff's apocraphyal one--the "generally acceptable standards" are usually highly manipulable The business magazines give a big play to the exceptions.

Getting down to cases again we can note a fugitive document by F.P. Johnson Jr , the president and financial analyst of a computer company (amongst others) in which he discusses ways of costing what are interestingly enough called evaluation services for computer systems--crucial for CAI applications. (A better title would be load-

*In building new franchise outlets, Dunking Donuts would get a loan, usually a seven-year note. The interest paid on this is obviously greater in year one than in, say, year five when most of the capital has been repaid. But Dunking Donuts wanted to enter only one-seventh of the total interest in year one; which of course made them look much healthier than they were (by about 15% of earnings, as I recall).

tuning, i.e., adjusting procedures and software to use the hardware optimally under the usual job constraints ("load") for that installation.) Three equally plausible approaches are discussed (based on the "discounted cash flow", the "payout time", and the "internal rate of return" analyses)--and lead to radically different perspectives on the defensibility of the investment. Only if it is understood how the three perspectives are related can a company treasurer (or investor) make a sound decision. Each is "true", yet each alone gives a false picture. In costing PLATO IV, the huge CAI project at the University of Illinois, very similar problems arise and--since the merit of CAI has always been extremely dependent on cost considerations--are really critical.

The financial case-history of a stock catastrophe like National Student Marketing Corporation offers a good deal of wisdom for costing service enterprises like evaluation as well as for evaluators in costing the services of evaluees. NSM shares sold at $71.50 in late '69, $1.00 in spring '72, and that loss was shared by many of the most prestigious funds and money managers. A key to the collapse was the misleading methods of profit estimation used (and audited) in '69; the true situation was there, but in very fine print and quite at variance with the tone of the financial report (Briloff, pp. 116-120). In short, the quality of analysis by both auditors and money managers is, to say the least, shoddy. The lessons for evaluators from these studies are numerous and some are very plain. One of them is well put by an outsider, F.J. McDiarmid, Senior Vice-President of a large life insurance company that lost millions in the debacle over Mill Factors. Reflecting on the failure of the auditors to detect (or announce) the corruption in the company's affairs, he says;

> The kind of auditing required to do this is no doubt both laborious and expensive and requires highly skilled people. It may not be forthcoming until finance company auditors feel that their primary responsibility is to investors and not to company management. One may doubt whether this will be fully achieved until auditors are retained and paid by the investors themselves. (quoted in Briloff, p. 131)

The auditor--and often the evaluator--is hired by the company he or she is supposed

to judge independently. The auditor's report is public and is taken to be a guarantee of soundness by the public and by the rest of the financial community. The sloppiness of "generally accepted accounting procedures" is so great, and the old gimlet eye so cloudy, and the motivation so lacking, that the real situation is often far different. Can we deny this about evaluation? Do we not sometimes place the imprimatur on projects that are far from deserving--pleading shortage of time or funds, or the absence of any necessity to do what our readers think we have done? One may feel that the situation is commonly different for evaluators in that they are often under contract to a federal funding agency, truly independent of the evaluee. But the agency is co-responsible for the project evaluated; it is the father even if the developer is the mother. It is often very clear that an agency doesn't want Congress to hear that it has been wasting money (e.g., when it swallows the negative Title I evaluations). And the evaluator's future employment has to come either from the agencies or from the developers. In fact, the money market situation is slightly better because there are supposedly independent regulatory agencies who can hire their own auditors, and ICC, SEC, and GAO have actually turned up a few scandals. But it is well known how seriously they have been co-opted, how often reports from their field staff are quashed "higher up", and how fast the lone rangers who call foul to the press are shuffled off to posts in Afghanistan or onto welfare. In all the great financial scandals of the '60s, from Leasco to Lockheed, there are only one or two cases where any disciplinary or corrective action by the agencies has resulted; none where it was adequate. Where can we look for consumer protection? The press? Sometimes--but the financial press is too dependent on advertising revenue, the educational press too short-staffed. Nader? Stretched too thin. Where did Briloff come from? He is a tenured professor of accounting, as well as a practitioner. It helps to have that basis for independence. We could use some life appointments for evaluators, to use the trick the judiciary relies on. The NIH Life Research Professorships were an interesting idea, no longer awarded; NIE should consider trying for the same thing in evaluation, where the independence is both more necessary and socially more valuable than in most research fields. One of the lessons of Watergate is that co-optability knows few limits when a man's am-

bitions or fears for his job and future are involved and when a lot of cash is floating around. The big federal projects involve a lot of cash and we should try to tighten up procedures before the grounds for scandal occur. A Life Evaluator might be a good example.

If I had the space, I would go on to the special topic of the costs of evaluation itself--the 10% Rule, the 1% Rule, and the concept of cost-free evaluation (the label is due to Dan Stufflebeam). The idea behind cost-free evaluation is that evaluation should normally be designed to effect significant measurable savings, and typically these should offset (at least) any direct costs of the evaluation. Exceptions are politically or legally required or morally referenced summative evaluations. Evaluation is not normally productive in the sense of creating a saleable product--and that accounts for some of the hostility towards it. But it is capable--when well-managed--of being productive in the sense of being worthwhile, a good investment, paying off. If an evaluation recommends that a project be terminated, it saves the continuation funds; if it recommends continuation, it saves products whose cost-effectiveness it can demonstrate. Is the cost-free conception of evaluation (a) realistic, (b) appropriate in all cases besides those indicated as exceptions, (c) productive of undesirable side-effects? The only way we'll find out is by doing more careful studies of (i.e., evaluations of) evaluation. This field of "meta-evaluation", or "secondary evaluation" as Tom Cook calls it, has now a tiny literature (Sanders, Cook, Scriven &c.) and some useful results. Its existence is important for the credibility of evaluation, for we need data on inter-evaluator reliability, costs &c. It seems to me a prime professional obligation of an evaluator to attempt to set up duplication or other check of his/her investigations, whenever there is time or money to do it (which is nearly always). This can be regarded as in-house meta-evaluation, and can be sequestered to improve reliability/credibility, or integrated to improve formative power. The Russell Sage Foundation is much to be recommended for its funding of a series of ex post facto meta-evaluations of important evaluations--Tom Cook did the Sesame Street one, and it has really improved our perspective on the original evaluation (most notably by showing that it was a summative evaluation done by a formative team with attendant weaknesses). One role of the

meta-evaluator corresponds to that of Briloff with respect to the accounting profession--the conscience/historian/critic role. As it develops, we may hope to see the same high standards of cost analysis applied as I am recommending for primary evaluation, and then we may find some answers to the questions posed above about the cost-free evaluation thesis.

9. Critical Competitors

If the client is interested in Best Buy evaluation--and very few uses of evaluation in the public education domain can avoid the obligation to call for that--then the most important of all comparisons in the evaluation process requires a look at the alternative options that would use similar, or other manageable resources.

Very often a client feels that he or she has already evaluated the decision to use resources in a particular direction (possibly with the assistance of external consultants) and is averse to having the evaluator go over that ground again. This makes good sense with respect to the internal ("in-house") formative evalutor in early stages of a project; it is only marginally defensible for an external formative evaluator and of course essentially irrelevant for a summative evaluation which would normally and properly be external.

The reasons for having evaluators frequently reconsider the choice of direction, the decision to throw resources into the effort to attain a certain goal, include;

a) the options may have changed--new products are now on the market, and a switch to them may still be worthwhile.
b) the evidence available about performance of the existing options, including the one chosen, may have changed, making a change--or termination--advisable.
c) difficulties (e.g., political) may have arisen in implementation which would not apply to other options
d) the original decision may simply have been erroneous--due to poor data or poor logic--and since this is nearly always a significant possibility there can be no justification for insulating that decision from criticism.

So the evaluator should usually, so to speak, "start from scratch," unless the evaluation is a routine formative. (Formative evaluation should frequently re-assess the whole situation, for the reasons just given; but not every time.) And one of the most important single elements in the evaluation that distinguishes it from a research design is the selection of the critical competitors or crucial comparisons. Even for Consumer Reports this is often a hard choice and a worse source of error than most of the other elements in their designs. One of their most brilliant choices occurred when testing proprietary carpet cleaners; instead of just testing these against each other, they tossed a dilute solution of Tide into the race. It won in a canter, at less than 1/10 the price. Teachers have to be tested against texts (in their cognitive role); texts against television (when CTW efforts are relevant); live lectures against CCTV; CAI against programmed texts &c. And more imaginative comparisons are important, against created competitors. The first person to pull the program roll from a teaching machine and try that on a student took the step that destroyed the fledgling TM industry and created that of programmed texts. In looking at a fancy CAI math-teaching set-up, one's first thought is to do the analogous thing--use a print-out patch-up as a text competitor. It seems a shame to cut the color plates and the justified margins and the cloth cover off the grade school text--but are those frills worth more than a million a year in California alone (a guesstimate)? One must look at critical competitors for that money.

One of the most interesting examples of the imaginative and valuable identification of critical competitors is illustrated in the following story, which may of course have been slightly embellished by the time it reached me. A year or two ago the University of California put up an extremely ugly new building for the mathematics department, with heavy federal subsidy. After it had been put up, it was discovered that it had been extremely badly designed, as is the norm with educational buildings, especially in that the combined elevator and stair capacity was totally inadequate for the usual number of people inhabiting the offices and small classrooms. Moreover, the only indicator showing the whereabouts of the elevators was located in the base-

ment level which was not the point at which most users began their wait. So it was common for faculty and students aiming at the more remote upper floors to wait for very long periods in the main lobby without any knowledge of how much longer their wait would be. This led to a great deal of dissatisfaction, and eventually a committee was set up to look into the costs of extra elevators.

Well, the costs of an extra elevator turned out to be about the cost of a substantial new building--which illustrates one example of a surprising critical competitor. But, rather than abandon the upper floors in favor of a new building, and not having the wherewithal to meet that kind of bill anyway, the committee decided they should look at other remedies. At about this time they managed to discover an Elevator Expert. This was a semi-retired gentleman who had many years of experience with elevator installations. They turned to him for advice, thinking perhaps of the feasibility of a second staircase mounted concentrically with the present one. The Elevator Expert advised that the staircase was not feasible in terms of building costs and/or lost space. But, he continued, he thought he had something which might help, which, in his previous experience, had often helped. And it might persuade them that his interest was not in selling elevators with which industry he was no longer connected in any remunerative way. His suggestion was that they take very seriously the idea of installing elevator-location indicators in the main lobby. While the committee had of course realized that this was something people would like to see, it hadn't really occurred to them that it might be, in a sense, a genuine alternative to an extra elevator. That is, the net dissatisfaction level among users of this elevator system might be decreased, if an indicator was installed, by an amount comparable to the results of installing an extra elevator, or staircase. The Elevator Expert prepared a careful estimate of the costs of this, and to their amazement they found that the cost of post-construction installation of such an indicator was well over $100,000. There was some feeling on the committee that federal auditors would not be enthusiastic about this expenditure and a general mood of despair began to settle over the committee. At this point the Elevator Expert said that he believed he could take care of the problem, for a few

hundred dollars. But, he went on, he had preferred that they would allow him to go ahead and try this out, without any prior explanation. He felt that they really wouldn't believe that what he was going to suggest would work, and he wasn't really certain that it would--still, his previous experience led him to believe that this was an environment where it might. He proposed to forego his consulting fee if it didn't work, provided the committee would stand still for the relatively small costs involved in any case. And, he added, "You can be sure that what I install will have some utilitarian value even if it doesn't solve our problem." The committee was at this point happy to agree, and it was decided that the criterion of success would be evidence from a post-installation questionnaire that met the standards they had been hoping to achieve with the installation of an extra elevator (not that the standards assumed 100% user satisfaction, no installation known to man, let alone devised by him, has ever met that criterion).

A month later the committee reconvened with the expert who exhibited the entirely successful results of the survey. What had he done? He had made an installation at each floor level, something which was completely impossible with the indicators for economic reasons; and what he had installed was a full length mirror. It turns out that the narcissistic tendencies of the species academicus are enough so that the opportunity to reflect on the vision revealed by a mirror quite distracts their attention from the vicissitudes of inadequate service by elevators.

Of course, it rather depends upon whether you define the problem as reducing subjective irritation or loss of work time, whether you find the previous example of a critical competitor satisfying. But it well illustrates the possibilities. There's a very common feature of economic behavior that might be described as the tendency for institutions and individuals to be influenced into choosing a cost level for the services and products they purchase by factors other than quality. Examples of this are to be found in the prices charged by interior decorators and decorating services serving society customers, the price paid for management, legal, and feasibility studies by large public utilities, and the often staggering differences in profit level

that are effectuated by new management in a large corporation, obtained simply by reevaluating purchases in the light of merit rather than irrelevant considerations. I have frequently found that a push for what I've called a "cheapie version" of some very expensive product provides by far the best critical competitors to the original products. It also proves extremely unpalatable for the producers to work up such versions. But trimming off the gingerbread often cuts the price in half and rarely has much effect on teaching effectiveness. Good examples include the talking typewriter, CSMP, and the teaching machine-programmed text switch. There's no doubt that pushing for these things encourages the evaluator's reputation as a bean counter, nit-picker, or cost accountant-type. But then the value of an evaluator is not to be found in his image but in the educational gains he can facilitate.

Critical competitors may be pre-existing same-market entities; or pre-existing different-market entities (frost-free refrigerators compete with self-cleaning ovens for the consumer's marginal dollar); or special creations; or possible creations. Consumers Union doesn't evaluate ovens against refrigerators, and many evaluators get very nervous about such "peaches and pears" comparisons. But it is often with these that the skilled evaluator can be most enlightening.

Much more needs to be said on this point and on others that will appear in the list of the next section, but time and space prevents it; this section alone should justify the comparison element in the Pathway Comparison Model.

10. Conclusion

A general outline of the Pathway Comparison Model is given below--it will be seen that we have covered most of the main elements, but the details of needs assessment, the identification of side-effects, the media-design-dissemination issues about the report, and many others have been left aside.

In sum, the model stresses the idea of evaluation as a context-controlled data-

compression procedure; it identifies a number of considerations that require attention, not in a one-shot way, but in a repeated iteration of cycles that gradually tightens up an evaluation until it provides us with an objective but user-oriented assessment of merit.

The Pathway Comparison Model

1. Characterization (How generally or specifically to describe the "treatment.")
2. Clarification of Conclusion with Client (Award of Merit, Best Buy, &c.)
3. Causation (Does it enter? How is it (to be) handled?)
4. Comprehensive Check of Consequences
5. Conceptualization (Compression) (Typically using preceding data but may use some from steps 6-8.)
6. Costs (Including disruption &c. and the costs of the evaluation)
7. Consumer Characteristics (Market and Need Analysis; covers consumers for the product and for the evaluation)
8. Critical Competitors (Real, ideal &c...repeat 1-7 for each of them)
9. Credentialing (Combining)
10. Conclusions and Communications (Data-processing, Design, Writing, Dissemination)

11. Postscript--The Evaluation of Goals

A common task in evaluation consists in evaluating proposals, and a major component in doing that is--or so it appears--the evaluation of goals. Another important evaluation task involves the evaluation of the management role of personnel where some opportunity for initiative exists. Here again, one is interested in looking at the goals that are identified by the manager as desirable ways to utilize resources available to him or her. Again, there is a distinct task for the staff evaluator who comes on board relatively early in a project, of evaluating the project's goals vis-a-vis the actual practices of the project and what the evaluator may take to be the implicit values of the enterprise. Obviously--it seems--goal-free evaluation is not relevant here. Now the Pathway Comparison Model covers goal-based as well as goal-free approaches. But as a matter of interest to a large extent the goal-free approach can be employed. For example, in evaluating the choice of goals by a manager

or planner (for example, a teacher or superintendent), what one implicitly does is to evaluate the goals as management instruments for achieving certain products/outcomes. And one then evaluates the relative merit of those outcomes against other possible outcomes, for which different goals would have been required. That is, one converts goals into instruments for producing certain products and then does a goal-free evaluation of those products. In short, one treats goals not as means to further goals but as means to an end that can be evaluated by reference to needs which it may or may not be someone's goal to meet.

How does this apply to the evaluation of proposals? The procedure is very similar. One is really evaluating a proposal as a way to expend available resources in order to achieve a particular outcome; so what one does is to evaluate the probable outcome against other possible outcomes from the same resources. Notice that what one really does here is to short-circuit the discussion of goals, in exactly the way that goal-free evaluation recommends; if the goals are grandiose and unlikely to be achieved, one simply applies a "reality correction" to them. If the goals are rather too modestly stated, and one expects a somewhat more substantial outcome, then one applies a "modesty correction." If one sees side effects that the proposal does not mention, one takes them into account when evaluating the proposal &c. So in fact what one evaluates is probable rather than goals. The goal-free emphasis here is entirely appropriate. But suppose there are cases where no side effects appear probable, where the goals appear realistic, and two proposals are in front of you, each of them requiring the same expenditure of resources. Surely, then, one is going to be forced to evaluate goals, since only goals distinguish the proposals? The argument is still unsatisfactory, because the probable outcomes are still quite different, and it is exactly these that one is interested in evaluating.

Well, isn't there an earlier stage in the proposal game where evaluation of goals is crucial; the stage where one is drawing up a list of targets at which proposals are to be aimed; the target list for the RFPs (Requests For Proposals)? Isn't the list

of "Educational Priorities for 1973" which we often see amongst the papers that are supposed to guide us in a panel review of proposals, really a list of goals, and couldn't one perfectly well evaluate these? Indeed, isn't an evaluation of these done every year in order to decide on the ones for the following year?

There is certainly a back-handed sense in which one can here talk of evaluating goals. But the fact is that the relevant pragmatic activity is the identification of needs, and of the outcomes which we hope will result from setting these goals as priorities. We can certainly say that such and such a goal is a trivial one, or that such and such a goal is a more important or more valuable one than another; to take an extreme example we could say that serving mankind is a better goal than serving oneself. But the example is extreme just because it is in the abstract moral domain; when we come back to practical educational evaluation, the focus becomes more and more concentrated on probable outcomes rather than abstract goals. And for this we can simply apply the goal-free version of the model discussed previously.

Now applying that model certainly requires that one pay attention to needs, and the satisfaction of needs is one of the most important goals that men and women have. So, commonly enough, there is some coincidence between goals and the satisfaction of needs, and a needs-based evaluation will coincide with an evaluation in terms of the goals of somebody who has correctly identified the needs and adopted them as goals. But that is an accident and not a necessity in evaluation; and since there are so many errors in identifying needs, it is of course an obligation on the evaluator to work from the needs rather than the goals, thereby reducing the sources of error. That leaves the solitary candidate for "real" evaluation of goals, within the educational domain, in the hands of the staff evaluator endeavoring to assist project management. The problem is that most of what is involved here should really be called <u>description</u> of goals, or <u>reanalysis</u> of goals, rather than evaluation. For what the staff evalutor is doing is either pointing out discrepancies between the goals of different groups, or discrepancies between goals and achievement, or between the goals of the project

and the goals of the funding agency, or between the goals of a proposal and the goals of the later practice of the project &c. He or she is not really in the position of evaluating these goals; or at least not primarily so. Still, there might be a situation in which re-evaluation of goals occurs, meaning by this a reconsideration of the whole enterprise of the project, or its particular emphases. The question could be translated into the form, "What should we do?" or "What would be the best use of resources by us?" And once one makes that translation it is of course easy to convert it into a problem of evaluating different proposals, i.e. different probable outcomes.

So there are really no examples of the evaluation of goals within the educational domain, that can't be translated perfectly well into goal-free terminology. Indeed the translation is usually of considerable assistance in improving the procedures of evaluation. Within ethics, now, there is indeed a task of evaluating goals; the goals for mankind, the goals for anyone seeking the good life. And part of some evaluations involves considering the ethical dimension of the activity. But even there, one should not get very much into the evaluation of goals rather than acts or achievements or probable achievements; for even the problem of verification is so much more difficult with respect to goals than it is with respect to achievements that it is undesirable to let much rest upon goals, which is to say intentions.

Appendix

THE EVALUATION OF PRODUCTS

A Proposed Standard Checklist of Requirements for Good Evaluations

1. Good evidence that the product does or will fulfill a need and/or find a market.

2. Good evidence that this need/market is important (because of size or vacuum or urgency, &c.)

3. Performance data must refer to eventual setting (not to supervised trials or early version).

4. Performance data must refer to student--or other ultimate consumer--gains, if possible (not just teacher gains or administrator gains &c.)

5. Performance data should refer to comparative performance of competitive products, if possible (not just to no-treatment control); in the absence of obvious competitors, they should be created, e.g., by creating "cheapie" versions of the product.

6. Performance data should refer to durability of effect, if possible (not just terminal state).

7. Performance superiority must be statistically significant.

8. Performance data should give absolute size and/or nature of gains, if possible (not just statistical significance).

9. Performance gains must be assessed as valuable and relevant to the need/market by more than one independent or uncontaminated expert judges (to show educational significance as well as statistical significance and substantial size).

10. There must be a systematic search for and study of side effects.

11. There should be a check for impropriety, injustice &c. in the process (of using, and/or administering the use of the product).

12. Cost data must be--comprehensive (disruption and "weaning" costs, capital vs. cash flow, maintenance &c.)
--verified independently
--provided for artificial competitors.

13. It is desirable if there is a plan for post-marketing support and improvement, involving a system for implementing internal revisions based on user feedback, mod-

ifications to suit new use circumstances, provision of user training, cost-reducing format changes when appropriate &c. (see Komoski's elaboration of this idea of his in a separate paper in this volume).

14. Dissemination plan (where appropriate) should be --clear
--feasible in terms of available personnel &c.