The Final Synthesis

MICHAEL SCRIVEN

INTRODUCTION

This article concludes with some reflections on the state of the discipline of evaluation, reflections which synthesize a number of considerations. In this respect it is like many evaluations which close with an overview—a synthesis of preceding sub-conclusions—but it is unlike them in that its content consists of a discussion of the methodology of the last synthesis step in an evaluation—of whether and how to make it. In many evaluations, the lack of any explicit justification of the synthesis procedure is the Achilles heel of the whole effort. The data-gathering is impeccable, the analysis of documentation painstaking, the statistics flawless, most of the design state of the art—but the last step is simply a reaction to the rest, a high-inference judgment which is rather obviously idiosyncratic. Even in the best current product evaluation, usually a simpler matter than program evaluation, one quite often sees a conclusion which is a flying leap from the premises, essentially a non-sequitur. There—and throughout personnel evaluation—that last step is where the prejudices are most likely to show up, in the absence of rules governing the step, because prejudices feed on judgments. Sometimes, although much less often than is usually claimed, there is no way to avoid relying on judgment at this point. But, whether we call the last step clinical inference, intuition, professional judgment, connoisseurship, or impressionism, the solid body of the clinical vs. statistical research makes it clear that we rely on it at considerable peril. That research shows how a very simple rule, if it is empirically-based, can beat expert judgment—including the combined judgment of a panel of experts—in almost all cases.[1]

RULE-GOVERNED SYNTHESIS

Of course, the presence of a rule, even in the strict sense of an algorithm—a rule which decides every case without further exercise of judgment—doesn't automatically get rid of the problem. It's quite common to see a rule set up for the synthesis step that makes the

Michael Scriven, Evaluation & Development Group, 415 Drakes View Drive, Inverness, CA 94937; POB 69, Point Reyes, CA 94956, fax 415/663-1913, or email scriven@aol.com.

Evaluation Practice, Vol. 15, No. 3, 1994, pp. 367-382.
ISSN: 0886-1633

step algorithmic—but is invalid. We'll give an example in a moment. The same applies, with more force, to rubrics—guidelines for scoring particular performance—which are heuristics, still requiring judgment for their application. Many rubrics in composition evaluation are quite arbitrary and often provably invalid because of shared bias *in the rubric*. When that occurs, the raters agree about the rubric and apply it consistently, but the rubric is wrong. Perhaps the most common and serious case of this, although by no means the only one, occurs because composition instructors are not logicians, although some compositions—an important group—are arguments or critiques of arguments.[2] Other errors occur because of shared bias in the *interpretation* of the rubric. The same problem has been demonstrated in the rubrics for ETS-produced history tests, current NAEP reading tests, and in most places outside the LSAT where someone with logical training looks carefully at material calling for functional or logical performance.

To the traditional test experts, it's no surprise to find some sloppiness in dealing with constructed-response items, even if produced by respected testing companies. They would be inclined to say that people who play in the swamps (i.e., grading essays) should expect to get in over their heads. So it's perhaps more interesting to find the same kind of error enshrined in their own multiple-choice pantheon. One example is the usual—and usually unspoken—rule that one gets the final score in a multiple-choice test—the synthesis step—by adding up one point for each correct answer. This rule is obviously invalid, since (i) the items differ in difficulty (i.e., getting the answer right can represent different levels of achievement, so correct answers should be weighted differently), and (ii) the *incorrect* answers are different in their "insight content"—their nearness to the right answer, what we might call their accuracy. Hence they should also be weighted differently (instead of all being given zero points). The latter weighing is here called differential error-scoring (DES) to distinguish it from differential item-weighing (DIW). We just throw away all that information about the student's achievement in order to avoid the nasty business of justification of the differences in item-weights and error-scores. But of course *level* weighing is itself a judgment about relative weight and scores, and a massively incorrect one—on most, although not quite all, tests. So the usual synthesis rule is invalid; in fact, it represents as little as half of the information about the student's achievement that is present in an exam script.[3] That means of course that some students will be misgraded and misranked.

The usual counter-arguments here are that adding and validating differential weighing is:

i) complex;
ii) expensive;
iii) unreliable (interjudge agreement about the weights is likely to be low); and
iv) tests of the effects of differential weighting show that the extra complication does not lead to significant changes in rank order.

To which the answers are that:

i) it's not very complex if you restrict the weights to 2, 3, or 4;
ii) doing this makes the differential cost of the more sophisticated scoring system very slight—and you can cut the length of the test needed to get the same amount of information about candidates[4], which should compensate;

iii) informal trials suggest the agreement about weights *from this range* will be very good if the usual calibration procedures are used (although we certainly need more experimental work), but in any case the point is irrelevant to use by a single teacher;

iv) ranking is often not the issue, for example, in certification tests—they yield a classification as excellent, competent, or not, which will certainly be altered; and for good measure;

v) if ranking *is* the issue, for example, for college admissions or jobs, errors are extremely serious and not to be dismissed on the grounds that they are *often* minor. The recent development of the multiple-rating item, which incorporates both DES and DIW, as well as many other improvements, while retaining full machine-scoreability, shows just how much we have been paralyzed by the paradigm of the multiple-choice item.[5]

In this example, we've carried the war to the quantitative experts, to show that the qualitative foundations of some of their completely rule-based approaches are completely unsound. It's shooting fish in a barrel to detail the parallel argument that shows how often qualitative judgment is inaccurate: we all have our own repertoire of examples (e.g., of bad appointments based too heavily on good impressions at an interview); and flaws in some of the most highly-structured use of synthesis-by-judgment—that is, scoring rubrics for essay items—have been mentioned.

A case which particularly concerns me, since I've been working on a federal project to improve the evaluation of teaching for some years, is the well-known arbitrariness with which some school principals add up the available data to come to a conclusion about whether a particular teacher is doing a satisfactory, marginal, or unsatisfactory job. This result is hardly surprising if one asks where they could have got the rigorous and extensive training that it takes to develop valid judgment in such matters. At least in the English composition case, it's well accepted that such training is essential, and it is done: I know of no administrative training program that includes it, and no acceditation/certification process that requires it of every administrative training program: which is not to say there are none, but rather to suggest that they are not the across-the-board standard that is an automatic consequence of realizing the importance of the synthesis step.

It seems clear enough that while there's no silver bullet for the synthesis process we should try to get a valid rule in place whenever possible as long as we do so without distortion. Failing that, we should try for heuristics and rubrics; and failing that—as well as when we do that—we *must* do systematic and critical training of the judges in the remaining cases ("calibration"), whether we're dealing with proposal, personnel, or product evaluation.

Now, the hardest common case is probably the case of drawing conclusions at the end of a big program evaluation. The data have been summarized, the statistics analyzed, the context considered. How do we get to a final conclusion? Certainly, we should use rules, rubrics, heuristics, calibration of judges. But where do the rules come from; and what cases are so impeccable that we can use them for calibrating the judges? Not from our own preferences, or we are into the territory of the arbitrary. To go further, we need to look more carefully at the foundations of synthesis methodology, and we'll do that in two stages. First, we'll look at how the synthesis is possible at all, which gives us some leads on how to get rules and standards statements; then we'll look at the way we get the weights in particular cases.

THE BASIC LOGIC OF SYNTHESES

Almost all simple evaluations can be unpacked into two components: the performance claim and the standards claim—which are then combined. This means there are three distinguishable steps in the logic—testing, standards validation, and synthesis. But for the cognoscenti, given the performance data, the synthesis is instantaneous and unchallengeable; they have internalized the standards. A car in a road test does 0-100 kph (62 mph) in five seconds: that's its performance on the usual benchmark test for acceleration. This is in the top 0.001 percent of road car performance: and that's one type of standards claim, one which will support some conclusions about both ranking and grading. So, in this case the first and final synthesis leads us to conclusions such as: this car has outstanding acceleration (the evaluative claim, a ranking by merit). But if a motorcycle turns in the same figure; the conclusion becomes "Its acceleration is only run of the mill for a sporting bike", because the standards are quite different. For the expert, the performance claim alone speaks instantly in the language of merit, because they have already translated the standards into performance language.

Analogously, there are sub-evaluations in program evaluation—let's say it's an educational program we are looking at—we can evaluate on the basis of a quick look at the equipment in the program's office: it passes muster against reasonable standards. We look at the furniture in a classroom, the facilities in a lab, the bathrooms in a school, the logging of disciplinary reports, and many more items. In the course of a complex evaluation, dozens of these simple evaluative claims turn up along the way, usually (and correctly) referred to as facts on the ground that they are extremely reliable judgments made by experienced judges against simple valid standards. They are scrutinized, consciously or unconsciously, by a trained evaluator for errors of fact, standards, judgment or inference. There's nothing exciting about them; evaluators build on evaluations most of the time. These are not examples of the final synthesis that we're targeting in this discussion.

But even those "on-the-fly" evaluations can—and occasionally do—lead to a serious question, logically and hence practically speaking, when the standards are questioned. And when we come to the final synthesis, whatever standards we are using for that are much more open to question, because they are multi-dimensional rather than unidimensional, and because there are often no clear standards for programs of just the type being evaluated. Standards claims are the epistemological key to evaluation, because they are the vehicle for the evaluative element, and they are the most vulnerable point at which an overall evaluative conclusion can be attacked, whenever there are no pre-existing and impeccable standards to which we can appeal. How do we support them? Can we get evaluative standards from further facts? If not, how can they be defended? If so, the facts/value distinction was a fallacy, because this is exactly what it maintained could not be done.

Well, it *can* be done. The example about car acceleration, although simple, shows there's no difficulty of principle. But one needs to understand the underlying logic if one is to make any defense in complex cases. Here are just a couple of points from that logic.

First, while it can be done more easily for *ranking* purposes (i.e., if the conclusions you're after use the language of "better" and "best"), it's much less easy for *grading* purposes, where you want to be able to conclude something about true merit rather than comparative merit. In product evaluation, grading is still pretty easy, because standards

of true merit are *usually* pretty well matched to what has been done (or else they're unrealistic), so conclusions about ranking are helpful in getting conclusions about grading. But when we start evaluating leadership training programs, programs teaching the scientific method or moral excellence, or the best way to produce outstanding cardiologists, we can't assume that the existing instances run the full range of the standards we wish to apply, so ranking is not a good guide to grading.[6]

Second, in the harder cases, it must be realized that "getting values from facts" cannot be done by the traditionally recognized types of inference—deduction and induction. It cannot be done deductively, as Hume and Moore pointed out, and it cannot be done by standard induction of the kind that is partly formalized in statistical inference. On those grounds, it was thought that the gulf could not be bridged. However, it can be bridged by legitimate inference, and the kind of inference that can build the bridge is in fact the most important kind of inference in practical life and science. Various logicians tried to add to the two basic types of legitimate inference, for example, C.S. Pierce, who called his try at it "abduction" (but not Toulmin, who went in a different direction, although one that was enlightening for other reasons). It's not clear that any of them succeeded, but the job can certainly be done, or else every issue of *Consumer Reports* is full of fallacies or mere matters of taste.

I have used the term "probative inference" for my version of this crucial type of inference, partly because that reminds us it is the key form of inference in the law. It is inference that makes a *prima facie* case for a conclusion; the kind of inference which often needs to be qualified with the phase *ceteris paribus* (other things being equal), that is, inference which is heavily contextual. It is the "logic of considerations" not the logic of sets of premises that in some sense contain their conclusions (deduction) nor of premises whose relevance is established by general heuristics (Mill's methods) or quantitative methods (e.g., statistical correlation). Thus, we must first agree that up to a certain point at least, better acceleration is a consideration generally that counts towards making a car better, just like fuel economy, comfort, roominess, reliability, etc. Of course, conclusions based on high weighting acceleration near the state of the art for cars may not qualify as even prima facie true. If we are considering cars for our teen-ager, we may feel that "better" acceleration, beyond a very low minimum is a drawback. In buying historically significant cars, it may have essentially zero value. In general, however, it contributes towards the recognized function of a car considered as a means of transportation that it is better in the respects that decrease the time of travel, and make entry to freeways relatively non-intrusive and safe, for us and for those already on the freeway.

Thus we defend the relevance of criteria of merit by appealing to the concept to which they apply, and to the known facts. This is enough to get us to conclusions about comparative merit (in this dimension) and to some conclusions about "absolute" merit, for example, "It's peppy enough for all practical purposes." Several dimensional scores like this, and nothing heavy going the other way, gets us to conclusions about overall absolute merit, from nothing more than an understanding of the concept and some performance data. But when it comes to the evaluation of health policies, or sex education, or mainstreaming for the handicapped child, the inferences won't be quite so simple.

To tackle such cases, we need to be clearer about a feature of the example just discussed which becomes more important with respect to the big issues. This feature is that one type of premise often involved in the inference to evaluative conclusions is quasi-definitional ("Acceleration is part of a good car's abilities"). Now concerning getting

legitimate evaluative conclusions without going beyond the boundaries of science, the skeptics never denied that one could deduce evaluative conclusions from *evaluative* premises, including definitional ones. For example, they would have had no problems with the following inference: "(i) Christianity is by definition good; (ii) the proposition "Thou shalt worship and not work on Sundays" is a Christian commandment; so (iii) working on Sundays must be bad." They just thought that a definitional premise couldn't add any evidential support to the conclusion, since definitions are mere language rules, arbitrary in a way that facts are not. So deducing evaluative conclusions from facts plus evaluative definitions simply produced an arbitrary conclusion. One might equally well use a different evaluative definition ("Jewish doctrine is by definition good") and add the factual premise: "It says, "Thou shalt worship and not work on Saturdays" in order to draw a contrary evaluative conclusion.

Thus, the validity of the inference to an evaluative conclusion, and hence the truth of the conclusion, is totally dependent on the values you bring in via any arbitrary evaluative premise—hence worth little. How different from true science, they thought, where the definitions import nothing except linguistic shorthands—which can easily be eliminated simply by making the premises longer—and the factual content of the conclusion is thus directly guaranteed by the factual content of the premises.

Again, their fixation on the deductive paradigm, which is tied to the classical notion of definition as a statement of logically necessary and sufficient conditions, was their downfall. As Aristotle suggested, and Wittgenstein clarified with the notion of criteria[7], there are important types of statements which are partly definitional, partly factual, and they will support evaluative assertions while in turn being supported by facts about usage and the external world. Examples include statements about the functions of artifacts ("Cars are means of transportation", "Watches are time-keeping devices"), and the functions of certain jobs from which we get quasi-definitional statements about duties ("It is wrong for judges to take bribes") which are of course evaluative but not arbitrary (judges who were not thought to be wrong at *all* for taking bribes *in the absence of very special circumstances* would not be judges in our sense of the term. There are many others, for example, "Control group methodology provides the only satisfactory way to absolutely rule out irrelevant causes". Statements like these, with some expansion, will support *probative* inference to evaluative conclusions ("Good watches, *ceteris paribus*, keep good time", "Good judges, *ceteris paribus,* do not take bribes"), and in turn can be supported by factual and analytical evidence about usage and the way things work, whether the things are artifacts or social systems. Thus, some evaluative premises come from functional analysis, and functional analysis can be based solely on logic and evidence.

In case this sounds tricky, try this example: "It's desirable to have good weather for picnics". Fair enough to call this a fact? Happier if we add, "other things being equal" that is, in normal circumstances? Or perhaps we should add, "in general"? It doesn't matter what we add, it only modifies the statement, it doesn't make it any less an evaluative claim—and a matter of fact, and *part* of the notion of a good or successful picnic. Now think back to the car acceleration test above. The particular choice of "standards claim" is just a plain fact—it's just a fact that a five second 0-60 time identifies a car with very good acceleration, which in turn means that performance is a merit in a car—*ceteris paribus*.If you like, you can qualify the claim to refer only to cars for people interested in performance; the logical point survives. There are many cases where the standards

are matters of fact, that is, undeniable (even though evaluative), and that is the end of the matter—we have pulled the evaluative bunny out of a purely factual hat. Probatively, of course. So synthesis at this fundamental level, where we finish up talking of propositions as being *both* matters of fact and evaluative statements, is clearly possible and legitimate.

Searle and others have pointed out important cases in ethics where classical definitions will, in conjunction with factual premises, *deductively* support evaluative conclusions[8]; while little of evaluation is wholly within ethics, there are considerable slices of personnel evaluation that depend on the notion of duties, including the example of the judge's duties above. And these cases are another nail in the coffin of the argument that you can't get evaluative conclusions from factual (and definitional) premises.

While the general nature of evaluative inference is not the main issue here[9], the final synthesis step is indeed an example of evaluative inference, and it sometimes exploits just the model described, most obviously when all the sub-conclusions are factual (costs, intended and unintended side-effects, etc.) It will rely on a complex missing premise of roughly the form: "If the performance of X on criteria A, B, C, etc. is high (these being all the leading criteria), and there is no evidence of Q, R, or S (these being all of the serious threats to the inference, that is, most of the possibilities excluded by the *ceteris paribus* consideration), then one may conclude that the evaluand is a prima facie good [better, best, competent, etc.] wristwatch [judge, clinic, school, analgesic, etc.]." Notice that this kind of missing premise is nothing like a classical definition, and the inference will be nothing like classical deduction or induction. It is a probative inference, an inference to a conclusion which should be believed until disproved. So, to apply this in program evaluation, we need to work on getting a comprehensive list of the possible (not the actual) merits of the kind of program we're looking at; this typically means using both conceptual analysis and empirical needs assessment. Then we can look at how well the performance of the clinic (for example) scores on each of these scales. Now, that's not quite what we usually do, because we do not usually do the conceptual analysis seriously[10] and tie the needs assessment to it. It's time to change to this model.

To summarize, one important type of final syntheses is the kind we have just been discussing, where probative inference is involved. A second kind we've mentioned is the category of cases where deduction from classical definitions is possible, although these are quite rare in most fields of evaluation. The third type of case we have discussed is the type where the standards claim is fully factual ("Zero to 60mph in five seconds puts a car in the top 0.001 percent of cars on the road"), which will support claims of relative merit (in this dimension)—and give some support to claims in absolute merit—and in some situations the final synthesis involves nothing more than applying a set of these standards.

In many real cases, however, a more complex inference than this is required, where we have to show that the *particular configuration* of scores by E1 makes it better than E2 with its own configuration of scores. To do that we have to get at the weighing of the criteria by relative importance.

At this stage, however, it will be useful to look at some examples to clarify some of the steps. These will be arranged in a different order from the list just provided, to help lead us into the hardest cases.

REFINING AND EXTENDING THE RANGE OF SYNTHESIS INFERENCES

Category I

If we are evaluating a laptop computer, we begin by making a list of criteria of merit: light weight, computing power, readability of screen, keyboard touch/size/layout, durability, service quality/availability, etc. These come partly from understanding the nature and intended function of the artifact, and partly from understanding how it will normally be used, which may add some other functions. The totality is still part of understanding what a laptop computer is and does today (partly factual, partly conceptual/definitional). If the computers are being evaluated for our own use, we will be able to allot weights that match that use and our preferences. But, even though it is our own preferences we are dealing with here, this is not a trivial task, and an evaluator may be able to help with it, just as a vocational counselor may be able to help someone turn their strongly felt but imprecise preferences about jobs into a sensible choice of a particular career. When we're evaluating for the benefit of a wider audience, we will either fall back on the definitional and "common understanding" criteria, or we will do a needs assessment on that audience to get both the list and the weights, or, more likely, both. More in a moment as to how to express the weights in both these cases (personal and public). First, let's focus on the properties of the list of criteria, which now becomes very important.

We are going to rely on this list to justify the inference from performance data to evaluative conclusions. Developing these checklists is not an easy task because the list should meet certain non-trivial conditions:

1) it should be comprehensive (i.e., include all significant criteria of merit for something of type X);
2) the criteria should be non-overlapping (i.e., not include any criterion, or part of it, twice—to avoid double-weighting of the overlap area);
3) only criteria of merit should be on the list—that is, properties that are part of the meaning of the concept of a good X. This excludes *all* empirically-proven indicators, that is, statistical correlates of merit. This is the "tough relevance requirement"; for purposes of evaluation, not for purposes of, for example, prediction, empirical connections are not, in general, relevant—hence their use is unacceptable.) Are there exceptions? Yes, If we absolutely cannot directly test performance on the full or an adequate set of the true criteria, for example, because of lack of time through no fault of our own, then and only then can we fall back on the use of indicators. And we have to alter our conclusions to reflect this limitation. (In the case of personnel action, unless the *Titanic* is sinking, this normally means starting again rather than drawing conclusions using secondary (e.g., statistical) indicators).
4) the descriptive level in terms of which the criteria are phrased should be the same (this helps avoids subsumption, which leads to overlap, but also facilitates understanding, applying, and constructing the list);
5) the descriptions should be—or should be expandable into a form—such that one can determine the presence of the criteria (and, if relevant, their amount).

Now, if one can assemble such a list, as one often can, along with performance data about each candidate on each criterion, there are certain probative inferences one can immediately make, in many cases, even though we haven't tackled the problem of systematically weighing for relative importance at all. This is possible:

1) where the desired conclusion is about "good enough" or "satisfactory" (and certain other evaluative predicates);
2) in cases where one of the candidates has superior performance on all—or all the (clearly) most important—criteria;
3) in cases where only one candidate avoids running into trouble with minimum requirements on one dimension (dimensional cutoff) or all combined (overall cutoff). In these cases, the conclusions one can reach are usually conclusions about merit—but, with some extra data, they can be conclusions about worth and significance.
4) Even if one can only get *most* of such a list, or *most* of the relevant performance data, one can still make the inference, *if* one can give a reasonable argument that the missing elements are very unlikely to make a difference—and one can often do just that.

So this is a fundamental kind of evaluation, but still quite limited: it will only get us to some qualitative conclusions, and it won't make fine distinctions. We can often do better—and we often need to do better.

Category II

Here we bring in the notion of weighting the importance of the criteria. In the commonly used numerical weight and sum approach (NWS), the dimensions or criteria of merit are weighted on, say, a 1-5 or 1-10 scale and the performance scores are normalized to, say, a 1-10 or 1-100 scale. Then the scores are multiplied by the weights and summed for each candidate; the winner being the one with the largest score. This approach is most attractive for its simplicity (it provides an algorithm), but is invalid and often yields a *completely* wrong answer—that is, the errors are not minor. One logical flaw is the assumption that one can use a single numerical scale for weights, performances, and number of criteria. That assumption can be bypassed by using a qualitative weight and sum approach (QWS^{11}), in which there are five incommensurable qualitative levels of importance for the weights:

1. Essential.
2. Very Important.
3. Important.
4. Just Significant.
5. Not Significant.

Of these five, the first and last are dealt with separately before any summing occurs; the first is simply the price of admission for a candidate—so you drop all that lack any of these (dimensional cutoffs again), and the last is a guarantee of exclusion for a criterion. The performances are mapped onto whichever of the three remaining weights the criterion is allocated, which leaves the number of criteria as the only numerical scale.

Saying that the five weights—and in particular the three "working" weights—are incommensurable means that no number of points scored in the currency of lower-weighted criteria can overpower points picked up on a higher-weighted criterion. Thus, the total score for each candidate is a triplet (M/N/P), not a single number. This means there will in general only be a set of heuristics to work with, in the process of identifying the winner(s), but a number of rules can be defined that categorically settle numerous special cases, and the heuristics will settle all cases that should not be draws, although sometimes requiring three iterations. This approach greatly extends the range of evaluative conclusions that can be supported beyond Category I without getting into the illicit oversimplifications of NWS.[12]

Given that the range of allowable weights is thus simplified, the question of justifying the weights is greatly simplified: we only need to divide the criteria into three classes, apart from identifying the cutoffs and irrelevancies. That can often be done from criterial definitions of ideal types, but in many cases of program evaluation we will need to turn to a needs assessment done on the targeted recipients. Of course, needs assessment is not wants assessment, so (i) the survey has to be carefully designed by someone who understands the difference (that rules out virtually all of the dozens that I've seen); and (ii) surveys are not going to do the whole job—there will have to be sophisticated analysis of the results before and after the statistical analysis.

Category III

There is a second flaw in the NWS approach that is avoided in the QWS approach. In the technical language developed in personnel evaluation, NWS assumes that all criteria of merit are "compensatory criteria", that is, bad scores on any of them can be offset by good scores on others. In fact, as we've been mentioning at appropriate points, many criteria have cut-off levels, that is, levels such that a candidate fails for falling below any of them, regardless of their performance on other dimensions. (For example, to be an adequate teacher of mathematics, you have to know a good deal of mathematics to teach it competently to the 12th grade, no matter how brilliant your communications skills with 12th graders; and vice-versa.) These "standalone" criteria make the process of evaluation a "multiple-hurdle" or "multiple-cutoff" process. In the absence of—and often in addition to—any cutoffs on individual criteria, there are typically cutoffs on the total score.

The best way to portray this situation is perhaps the use of a histogram, cut across with a grid showing the relevant quality levels, for example, the grades A-F, and with a vertical column for each dimension of merit. This framework allows us to construct a profile of a candidate's performance, from which it's easy to see the strengths and weaknesses. On this chart one can put in a double horizontal line across each criterion at the cutoff level, if it has one.[13] Profiling is sometimes a way to paint a picture of several candidates without actually drawing evaluative comparisons. However, if some, or all, or all but one of them, miss on a cut-off, profiling may be enough to demonstrate the required type of evaluative conclusion—for example, identifying the best or the acceptable candidate(s)—without getting into weighing the criteria at all. If one is prepared to add very simple weighing—e.g., in faculty evaluation, by saying that teaching is to be weighted as much as research, and service only about half as much—a large proportion of cases can be decided (in the faculty case, essentially all cases). And some of the remaining cases should not be decided, a possibility to which we now turn.

CONSUMMATED AND UNCONSUMMATED EVALUATIONS

To synthesize or not to synthesize, that is the question. We need, if possible, to distinguish between the type of case where it's improper to push for a final synthesis and the type where it's improper *not* to—where the failure to do so results in what the **Evaluation Thesaurus** calls an unconsummated or fragmented evaluation. Profiling is an intermediate case between consummated and unconsummated evaluations, often used in order to escape consummation, but often correctly used when no consummation is appropriate, or formative evaluation is being done (or justification in court for a summative evaluation has to be presented or available).

The distinction between cases where final synthesis is proper and cases where it is improper depends on three factors—the evaluation needs of the client, the limitations on available data, and the configuration of the facts that are available. If the client needs to decide which waterjet powered runabout to buy for her or himself (which assistant to hire, which database system to install to run the nonprofit, etc.), then a display of profiles won't cut it; you must try to develop a plausible weighing scheme, based on your assessment of the client's relevant needs. If the available facts of the case are limited, your effort may still not yield a ranking, and profiles are all you can do; the client must then pick, taking the chances involved in a situation where the profiles are incomplete. If you have all the data you need, and that still won't yield a ranking, then you have a tie, which is an overall conclusion; the client can then pick randomly or by applying secondary or new or newly refined preferences.

In big program evaluations, however, the client's needs or wants only set some parameters on what has to be done. Assuming that the client's problem calls for a ranking, then you must turn to the needs and—to a lesser extent—the preferences of the targeted recipients to provide the weights. The needs assessment plus wants assessment should be designed just so as to yield the relevant criteria and their weights. Typically, the procedure would require you to try for a comprehensive list of criteria and tentative weights based on the concept, the experience of the service providers, and the literature, and take that to the consumers for additions to the list and refinements of the weights, of course keeping in mind and in the design the possibility that a radical change in list and weights will occasionally be necessary. The resulting improved list and better weights will normally make it possible to use QWS to get you to an overall conclusion (which may be a tie).[14]

Methodological Prejudices

Your decision whether to consummate an evaluation should not depend on some preconceived notion of the nature of evaluation which excludes "this kind of thing", for example, the view that evaluation should only provide rich description. It's clear that good evaluation in all fields can and often does provide useful rankings by merit, and gradings against useful standards. And both of these can be used with great effect in formative contexts, not just to support decision-making.

In particular, your decision should not be based on the old view that evaluation should just uncover the facts, leaving to the client the interpretation of the facts according to his or her values. This doesn't even let you do evaluative profiling, since that requires *evaluating* performance in each of several dimensions. In complex evaluations, the client

is typically very badly equipped to create and then integrate a large number of sub-evaluations or dimensions due to absence of experience and the presence of myopia (from being too close to the issues) and prejudice. Failing to consummate in such cases is letting the clients down at exactly the moment they need you most. The dimension scores will often need to be converted into sub-evaluations, and in either case synthesized, and the way to do it appears to be via QWS methodology, with the lists, weights, and cut-offs based on the needs of the people the program is supposed to serve.

SYNTHESIZING A CONCLUSION ABOUT THE FIELD OF EVALUATION

The above methodology can be applied in detail to the field of evaluation itself, as an exercise as well as for its own interest. But that would take more space than we have, so we'll just sketch in one or two of the moves. Let it be clear that we're talking about *evaluation* here—the whole discipline, not just program evaluation or performance evaluation (e.g., student assessment), or product evaluation.

What kind of conclusions would we like to get from this evaluation? (This is where we should do the client needs assessment—the client being a combination of the editor and the author, both at least partly representing the audience for the journal, that is, AEA membership.) It would be interesting to get an overall sense of the state of the discipline by comparison with (i) other, more traditional, disciplines, (ii) against its own prior states, and (iii) against a proper ideal of a discipline (i.e., three kinds of summative ranking and grading). It would also be useful if this evaluation could provide some suggestions for improvement (formative evaluation). You can see why we won't have space.

What kind of considerations bear on these conclusions—what's the data space? There are about 30 intellectually and socially important fields of evaluation with their own titles, ranging from literary criticism through technology assessment, the judging of athletic events, restaurant and movie reviewing, applied logic (e.g., argument analysis), to appellate court jurisprudence. This is a good case where one should avoid averaging because that would conceal vast differences in sophistication and utility. So at least for overview purposes, a set of profiles would be appropriate. The size of the set is again formidable.

Nevertheless, some conclusions can be quickly drawn in particular cases: literary criticism is probably in the worst shape of anything laying claim to be a discipline of evaluation, this side of astrology, and even product evaluation—with all its advantages of low reactive effects and the relative constancy of the candidates' physical and socio-economic properties during the period of testing—is still a long way from real competence.[15] (Here, we're developing the sub-evaluations we need for an overall synthesis). Other conclusions apply across the board, notably the virtually complete absence of theory in all these fields except program evaluation. Even there, the theories, while usually describing themselves as theories of evaluation, are in fact only theories of program evaluation, and geocentrism is a sure sign of immaturity. A "C-" for program evaluation on this dimension, and an "F" for the other fields.

The general absence of theory is presumably due to the operation of graduates from the Thought Police Academy, according to whom value claims have only the status of expressions of personal preference, that is, no objectivity and hence no place in science or any respectable discipline. (The latest faculty appointment there is of course Dr. Deconstruction who joined a foundation faculty all of whom had their Ph.D.s in Positivism).

The absence of theory is terribly costly; different fields have reinvented the wheel or, worse, failed to do so; fundamental errors exist in many common evaluation procedures because of the failure to look at their logical foundations; the field lacks a viable self-concept and hence does not interact with other fields on a peer basis; and so on.

On the other hand, the payoff from the practice of evaluation in all these fields, many of them long antedating science itself, is still enormous; a definite "A-" grade on this dimension. While the academic miasma of deconstructionism somewhat blurs our vision if we spend much time in the ivory tower these days, the practical fields spend little time with such concerns because it's clear that—academic outcast or not—evaluation can and often does save lives, families, species, quality of life and money—and can perhaps save the globe. Of course, it also refines weapons for destroying the same, but the balance of use is our responsibility, not the discipline's. Still, it doesn't pay for the practitioners to get too complacent since the foundations of most practical fields lacking any theory or metatheory nearly always contain some cracks, and we've pointed out some serious ones here. Those flaws keep practice from the "A" or "A+" grade for which it should aim. (You can see that we have set of profiles on the evaluation sub-fields developing which are very different between themselves, and also have very mixed ratings on their internal dimensions—a high mark for practice, a very low one for theory (except in program evaluation), etc.)

Overall, the self-concept of the field—this whole collection of fields making up evaluation—gets a very bad grade for failing to confront the myths about the impropriety and subjectivity of evaluation. Even in the domain where theory has had some play, of the seven waves of theories of program evaluation that have led the discussion over the last 30 years, five reject the idea that the principal function of evaluation is to produce evaluative conclusions, surely a sign of profound confusion[16]. Still, let's say the grade is only a "D-"; the pressures were incredibly strong, so the degree of difficulty of a "C" was very high. A "D-" is surely a compensable grade, so there's still hope.

The most secure basis for that "D-" is the total failure of evaluation to recognize intradisciplinary evaluation, let alone help it to achieve a self-concept and improved methods. The standard disciplines have themselves failed to recognize or develop their own pervasive internal use of evaluation which ran in parallel with their denial of (what they saw as) its external legitimacy. Not even in philosophy, not even in philosophy of science, was there any recognition of this bizarre incongruity, this tacit contradiction between their constant and essential *internal* practice of evaluation and their theoretical professing to deny its legitimacy, based on categorizing it as something *external.*

If we make a rough distinction between theory and methodology, then we can concede that methodology is in many areas, for example, product and performance evaluation, quite sophisticated. However, looking further at program evaluation, which in many respects illustrates sophisticated methodology in evaluation work, we note that probably the most intense debate of the last several years, up to and including the Spring, 1994, issue of our journal *New Directions in Program Evaluation* has been the qualitative vs. quantitative debate. That fact provides a strong basis for a "D" grade in itself, since this is an issue right up there with thinking that the great problem for the U.S. Public Service is whether it should hire men or women.[17] It's a non-dilemma: hire the best for the job, looking at system considerations as well as the day-to-day tasks to be performed, and you'll find it's sometimes one, sometimes the other, sometimes both. The gender of the hire should be almost entirely incidental to the merit for the job; the choice of qual or

quant methods should be entirely based on merit for the job. There are mighty problems in the methodology of evaluation that better deserve out time, for example, identifying an algorithm for apportionment; working out whether, when, and how evaluations can include, imply, or support recommendations—and answering the same question about the relation of evaluations to explanations of the operation of (for example) the program being evaluated (the "program theory" issue). The over-emphasis on the qualitative/quantitative issue may be largely a by-product of the underdevelopment of theory.

Evaluation is concerned with significance, not just merit and worth. In considering the significance of the field of evaluation in general, remember that no discipline can even establish that it *is* a discipline without serious evaluation—of its *own* methodologies, its instruments, its data, its inferences and theories, its publications, etc.—let alone continue to operate scientifically without using intradisciplinary evaluation all the time. Thus, even ignoring its vast practical pay-offs, one can argue that evaluation is by far the most important discipline of all—with the possible exception of logic. Yet it is never represented in the councils of the disciplines, something which needs to be changed, and will change only and as soon as the field acquires an appropriate self-concept and the self-esteem that goes with it and is required in order to fight for a place in the sun. The very idea that official evaluations commissioned by Congress are normally done by the National Academy of Sciences and the National Research Council (NRC), august bodies who don't recognize evaluation as a discipline, is bizarre. It has led to some appalling results, for example, in the NRC's recent evaluation of the "productivity paradox" claim—the claim that information technology has not added productivity to the service industries. (They dismissed it, using a massively biased panel, without addressing the serious arguments for it.)

Compared to its own early days, the field has progressed vastly in many of the essential logistical respects that distinguish a discipline. It has professional associations on three continents and two in North America alone. The two associations in North America are both large and powerful, the one in Australia moving in that direction, and the one in Europe just beginning. It has many journals of its own, and several conjoint ones—some of them excellent, some struggling to differentiate themselves from run of the mill applied social science—and it has opened the door to the publication of evaluation studies in many others. An "A" on this developmental dimension.

Looking at the overall situation, one must ask first whether evaluation has scored below a cut-off level on any dimension. The case for a cutoff seems only to be strong on the practice dimension, and that is, for most fields of evaluation (i.e., literary and paint-art critics and wine columnists apart), a high point on their profile. So we *can* synthesize these diverse considerations, although the fields score so variably that separate profiles should also be provided. In this case, where one is in effect synthesizing a set of profiles, it is hard to justify much in the way of weighing of the dimensions, so one can perhaps best give a report card which is itself a kind of profile rather than a holistic rating. In any case, doing so is more helpful for formative purposes, one of our aims in this exercise.

"E has exceptional practical skills and potentiality. Although developmentally-challenged in the theory dimension so far, E is showing some signs of improvement in that respect. While E has been very successful with many practical tasks, and in social linking, there is a definite need to put more time into reflective consideration of E's own discrepant achievements and schizophrenic self-concept, an activity which will surely assist future achievements and in turn self-esteem and influence."

NOTES

1. See Paul Meehl, *Clinical vs. Statistical Prediction*, University of Minnesota Press, 1957. There is a large body of follow-up literature which confirms the results reported in this groundbreaking metastudy.

2. See B. Davis, M. Scriven, and S. Thomas, *The Evaluation of Composition Instruction*, EdgePress (for the Carnegie Corporation), 1981. We scored 16,000 compositions, on two topics, one calling for the criticisms of an argument, one calling for expressive writing, using both trained composition evaluators and trained logicians. There were major differences between the results from the two groups of judges in scoring the arguments.

3. The following very simplified analysis will convey the sense of the point. Suppose half of 100 questions on a test are hard and half easy, and the hard ones are (roughly) twice as hard as the easy ones. Allocating four points for the hard questions, and two for the rest gives a total of 300 points. If you score all questions two points, for a maximum of 200, you'll be missing as much as 100 merit points (one third of the total) that would show up in a correct scoring system. Now suppose that the best wrong answer to all the questions represents half as much information or understanding as the right answer (i.e., should be scored two points (out of four) or one point (out of two), depending on the difficulty of the questions), whereas the worst wrong answer is worth nothing (shows zero relevant knowledge or skill). Then, by ignoring this difference, you throw away another 150 points of information about some evaluees (those who get next-best answers on each question will get 100 points on the hard questions and 50 on the easy ones, instead of zero). So the total amount of information obtainable from the test is 300 points on the right answers, and 150 on the wrong ones, for a total of 450. The usual synthesis rule only picks up 200 points of this, so there's a maximum loss of more that half the information points about merit actually present in the test. Of course, this analysis is oversimplified (e.g., because one student can't be eligible for all the DIW and all the DES points), and there are some tests that are not this far off. However, there are many which are worse (including, this time, the LSAT). This is just a plausibility argument using an informal notion of information content, but it's strong enough to show that serious errors about the level of competence of a testee are involved in the usual scoring process.

4. We use this term "candidate" here to cover whatever is being evaluated, program, product or person.

5. And faced with the false dilemma of using multiple-choice items or dubiously "authentic" assessment. There's a brief description of multiple-rating items under that heading in Evaluation Thesaurus (4th edition, Sage, 1991); a longer account is in Critical Thinking: Defining It and Assessing It (Sage, forthcoming).

6. Even in product evaluation, this possibility has to be kept in mind. For example, we still do not have on the market a toaster that will produce evenly toasted slices of bread from different loaves, after reinsertion and resetting, and from a succession of slices as the machine warms up. Clearly a really good or an excellent toaster should do this, although the best toaster on the market does not. (Based on use of the top-rated toasters from Consumers Union 7/94 report.)

7. Details are in Scriven, M., "The Logic of Criteria", Journal of Philosophy, Vol. 56, October 22, 1959, pp. 857-868.

8. Searle's case is "You promised to do X (factual premise). One should keep promises (standards claim, true by incontestable definition). So, you should do X (evaluative conclusion)."

9. It is discussed further in Evaluation Thesaurus (ibid.), as are most of the issues skipped over lightly in what follows.

10. What does that mean? It means, for example, brainstorming with experts and focus groups on "What would an ideal clinic (etc.) contain? How would it be run?", doing a literature search for lexical clues to criteria of merit, doing cost analysis on combinations of the resulting list to get

a concept of the "workable ideal clinic", etc. More details in the next section, for example, about what features the list of criteria must have.

11. Some may feel that the term "Qualitative Weight and Sum" is self-contradictory, but I think it just represents the ideal of a successful marriage between the qualitative and quantitative approaches.

12. Incidentally, NWS cannot be improved enough to save it in the way that multiple-choice items are improved by using a simple, limited, numerical, weighing scale, such as the 2,3,4 scale suggested earlier; the remaining errors are still often fatal.

13. Note that the cutoff is not necessarily the C grade; it may be a B or a D, as is often the case at different levels in college. The overall cutoff, that is, the minimum required average grade, can be shown as a double bar on an extra column to the left of the dimensional ones, with any extra requirements for passing provided in a note attached to that bar, for example, "Cannot be below C on more than one dimension, even if the dimension has no cutoff".

14. Worst case analysis suggests that you may occasionally need to go back to a short interview of a small sample for final resolution.

15. Details are in Scriven, M., "Product Evaluation—the State of the Art, "Evaluation Practice, Vol. 15, No. 1, February, 1994, pp. 45-62. It's notable that even Consumers Union, which does so much evaluation so well, goes wrong on the synthesis step.

16. Details are in Scriven, M., "Hard-Won Lessons in Program Evaluation" which makes up the Summer, 1993, issue of New Directions.

17. Note that essentially all contributors to this issue of the journal conclude that the two approaches are compatible. If there's anything left to the debate that's worth debating, it is specific case studies, to see if and how each approach can contribute or combine with the other in new ways.

REFERENCES

Davis, B., & Thomas, S. (1981). *The evaluation of composition instruction.* EdgePress.

Meehl, P. (1957). *Clinical vs. statistical prediction.* Minneapolis, MN: University of Minnesota Press.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Beverly Hills, CA: Sage.

Scriven, M. (1994). *Critical thinking: Defining it and assessing it.* Beverly Hills, CA: Sage.

Scriven, M. (1994). Product evaluation: The state of the art. *Evaluation Practice, 15*(1), 45-62.

Scriven, M. (1993). Hard-Won Lessons in Program Evaluation. *New Directions* (Summer).