Articles should deal with topics applicable to the broad field of program evaluation. Implications for practicing evaluators should be clearly identified. Examples of contributions include, but are not limited to, reviews of new developments in evaluation and descriptions of a current evaluation effort, problem, or technique. Manuscripts should include appropriate references and normally should not exceed 10 double-spaced typewritten pages in length; longer articles will occasionally be published, but only where their importance to EP readers is judged to be quite high.

The Evolving Syntheses of Program Value

ROBERT STAKE, CHRISTOPHER MIGOTSKY, RITA DAVIS, EDITH J. CISNEROS, GARY DEPAUL, CHRISTOPHER DUNBAR JR., RAQUEL FARMER, JOAN FELTOVICH, EDNA JOHNSON, BRENT WILLIAMS, MARTHA ZURITA, and IDUINA CHAVES

Robert Stake

ABSTRACT

In this paper, we object to Michael Scriven's claim that the basic logic of evaluation is criterial and standards-based. We note that valuing is an integral part of perception and that valuing within perception, repeatedly refined, is an even more basic logic of evaluation. We find unpersuasive his claim that making the final synthesis "governed" will diminish bias, noting that bias will find its way into the required statements of need, function, standards and weighting. We offer our alternative for disciplining the synthesis process, by urging more systematic and demanding critiques of emerging interpretations and values, and by more deliberately using competing conceptual organizers (e.g., goals, issues, decisions and elements of the rules Scriven advocates) as temporary and dialectical grounds for reconsidering the evolving meanings of the program, including its merit and shortcoming.

Robert Stake • University of Illinois, Urbana-Champaign, 190 Childrens Research Ctr, 51 Gerty Drive, Champaign, IL 61820.

Evaluation Practice, Vol. 18, No. 2, 1997, pp. 89–103.
ISSN: 0886-1633

INTRODUCTION

This article is a response to Scriven's (1994) paper, "The Final Synthesis," in which he urged professional evaluators to develop and pursue an explicated, rule-governed process for drawing conclusions about the merit of the evaluand. In this article, Scriven called existing summaries of value "the Achilles heel" of our work, inappropriately filled with "clinical inference, intuition, professional judgment, connoisseurship, and impressionism;" (p. 367) concluding that proper attention to criteria, weighting, and logic would discipline our work and make it more useful. We[1] applaud his effort to make final syntheses more carefully reasoned but urge most evaluators to beware the governed approach and to discipline and take pride in their judgment-governed, evolving synthesis processes.[2]

With Scriven we agree on much, so much that some who read earlier drafts of this paper had trouble discerning a difference. Yet, we have deep and fundamental differences with some of Scriven's central views, sending us in different directions at critical moments in an evaluation, causing us to allocate resources quite differently, causing us to rely differently on personal judgment. With Scriven, we agree that having formal plans, preordinate questions, explicit criteria, a sense of standards, intuitive awareness, conditional limits to values, and multiple perspectives will and should be a part of evaluation theory and practice. We differ on the preferred source of values, the role of explication, the best means of purging bias, and the need for retaining the dominion of professional judgment over governed syntheses.

RULE-GOVERNED RATIONALITY

At least since his "Methodology of Evaluation," Scriven (1967) has written about a common logic for all evaluation. He spelled it out under the title *The Logic of Evaluation* (Scriven, 1980) and later summarized it this way:

> Evaluation first involves *analyzing* a product into its principal dimensions of merit, then analyzing those indicators; then selecting benchmarks which will measure those indicators and running the benchmarks to get performance data. Then we work backwards, *synthesizing* the data into conclusions about each indicator; then synthesizing the conclusions about indicators into conclusions about dimensions of merit, and finally synthesizing those conclusions into an overall evaluation (1990, p. 15).

In "The Final Synthesis," he indicates that, for *program* evaluation, the analysis requires knowledge of the nature and functions of the program, the needs of the beneficiaries, and the program's resources, but that the logic is the same as for *product* and *personnel* evaluation.

It is easy to read Scriven as advocating a linear approach, a preordinate approach, particularly when describing the methodology of product evaluation (1992, p. 371). Thus, Fournier (1995) translated this logic as[3]:

1. *Establishing criteria of merit.* On what dimensions must the evaluand do well?
2. *Constructing standards.* How well should the evaluand perform?
3. *Measuring performance and comparing with standards.* How well did the evaluand perform?

4. *Synthesizing and integrating data into a judgment of merit or worth.* What is the merit or worth of the evaluand? (p. 16)

In spite of the strong ordering of the steps above (as well as linearly in the figure we present at the close of this paper), Scriven advocates an iterative, non-linear process of explicating needs, criteria, standards, and performances (1992, p. 371). But it is also clear that, not surprisingly, he sees the synthesis process as at least generally subsequent to the analysis and that, with repeated and detailed description of his qualitative weight and sum approach (QWS), he sees the ideal resolution of value as more a mathematical than a judgmental process (1992, p. 380; 1994, p. 365).

Probative Analysis

We read the logic in Scriven's prototypical syntheses not only in the rational steps listed above but equally in the use of probative inference:

> It is inference that makes a *prima facie* case for a conclusion; ... that is, inference that is heavily contextual. It is the "logic of considerations"... It will rely on a complex missing premise of roughly the form: "If the performance of X on criteria A, B, and C, etc. is high (these being the leading criteria), and there is no evidence of Q, R, or S (these being all of the serious threats to the inference),... then one may conclude that the evaluand is a prima facie good [better, best, competent, etc.] wristwatch [judge, clinic, school, analgesic, etc.] (1994, p. 371).

We interpret this to mean: define the required performances so sharply that comparing them to actual performances will make the merit of the evaluand obvious (and not a matter of judgment). Explication again is the watchword.

We can see this happening in product evaluation. Regarding program evaluation, we cannot see that our own practices, the practices of the evaluators we interviewed, or the practices presented at annual meetings and in journals fulfill, or even approximate, Scriven's probative inference conditions. Collectively, our ability to specify criteria, explicate needs, set cut-score standards, and measure performances is weak. We are greatly dependent on estimates, approximates, and indicators of questionable validity. The "measurement community" (e.g., psychometricians, sociometricians, econometricians) has been devoted to improving these measures for most of a century. It is not reasonable to expect that greatly better measurement will occur soon. Our evaluation theory needs to be based on a realistic view of our explication capabilities.

Scriven recognizes current inadequacies in our specifications and measurements but advocates this probative inference is the reasoning we should strive for. He implies that for improving our practice, more judgmental processes will not suffice, only leaving our practices both inaccurate and biased.

Universal Logic?

Although Scriven (1994, p. 370) implies and Fournier (1995) has found in her studies that the logic described above is common to all evaluation work[4], it is a logic that many of us in evaluation think we rarely rely on. After reflecting on our own behavior, we interviewed 13 theorists and practitioners[5] and were unable to find this logic in their words or embedded in

the actions they described. Our sample was small and perhaps unrepresentative but the responses reinforced our doubt that this is the logic of evaluation practice. What our interviewees said they did do was more a matter of seeking to understand what is going on and devising representations of production, performance, effectiveness, crisis management, staffing, or whatever is needed to help describe the evaluand. All along the way, they sought to validate the evolving representations.[6]

The rule-governed synthesis that Scriven advocates is not just a procedure to follow after data are gathered. It places early demands on design. Instead of strong and immediate emphasis on goals, claims, forthcoming decisions, issues, or productivity, Scriven sees the evaluator's early work as gathering data on needs, setting criteria of merit, etc. for probative inference.

Our interviewees reinforced the idea that criteria, as well as goals, issues, and needs, provide evaluators with frameworks or scaffolding. The evaluators spoke of criteria but not so much as succinct and critical particulars — as Scriven prescribed — but more as broad classes in which data can be placed and easily referred to. It occurred to us that criteria provide an explicit framework to do inexplicable processing of ideas about program function. Seldom are the criteria seen as direct criteria of merit but rather information categories from which interpretations of merit are made.[7] Standards are seldom explicitly identified. Personal judgment is common.

In discussion of our response-draft with him, Scriven acknowledged that evaluators often do not explicitly follow this logic. He added that the point of his paper was not to claim they do but that they should. And we asked, given how far we are from being able to do it, is this the proper target for our continuing professional development? In the paper, Scriven foresees explicit subscription to the rubric helping us to reduce reliance on evaluator judgment, especially that based on personal value.[8] We too would reduce bias but have little faith in rubrics for doing so.[9] We would rely more on critical review, i.e., exposure of conjectures and drafts to internal and external review, although less rational a process.[10] Our view is that *perceptual judgment* is the essential logic of evaluation more than is explicit comparison of performance to standards, and efforts to improve professional performance should be to help evaluators judge — not less — but better.[11]

ON KEEPING THINKING ECLECTIC

Scientists and poets, men and women, elders and children are alike in drawing upon both intuition and formal logic. They all do it. There are situations where one is appreciated more than the other but quality of thought sometimes will be raised by becoming more rational, sometimes by becoming more intuitive. In the final synthesis, there is need for both intuition and conscious reasoning. It would not be wise to put value resolution entirely into a rational mode. Rather, shifting back and forth between the formal and informal, the general and the particular, the hunch and the habit, increases the strength and comprehensiveness of the scrutiny.

At any one time, in any one place, interpretation may rely too much on one, not enough on the other. And the professional community, intentionally or not, may rely too much or too little on its rubrics. Scriven said that "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" (1994, p. 367). His remedy is to diminish the emphasis on judgment as found in "clinical inference, intuition, professional judgment,

connoisseurship or impressionism" (p. 367) and to rely on more explicit treatment of the considerations. Scriven acknowledges that explication and rules do not guarantee good logic but expresses confidence that they will diminish prejudice and caprice.

Explication and rules are good in that they can bring into sharper focus the more important considerations, e.g., certain needs, certain services, certain implications. They often do a disservice by sharpening the distinction between the stated and the unstated, by creating criteria even more clear than connoisseurs recognize (Petrie, 1995, p. 89). Parsimony has a recognized place in science. Its place in evaluation seems more a matter of personal style: Scriven and many others are essentialists, more comfortable with parsimonious explanations, satisfied with six or ten criteria for the comparison of cars or schools or public policies. Compared to realist Scriven, we are relativists, seeing all manner of local action and condition influencing events, and feeling uncomfortable with notions of simple scales and causal links.[12] Community and evaluation-team alike benefit from both essentialist and relativist thinking. Both thought processes can impede action. But not all considerations can be developed. Sooner or later, a finite report must be written.

Those who have long attended American Evaluation Association meetings know that Scriven is fond of citing the example of essay grading to illustrate how criteria enhance and objectify a traditionally subjective process, citing the work of Paul Diederich (1957) of the Educational Testing Service. He brought it up again in "The Final Synthesis" (p. 381). Dietrich did substantially increase the reliability of grading essays and the validity of predicting college grades but did not obtain evidence that, with readers shifting to criterial grading, the quality of the essays was better recognized.[13] Criterial treatment of any evaluand transforms experiential knowledge of it into a knowledge of selected characteristics.[14] When we know well the nature, functioning, and merit of something, we usually know more than its profile of scale values.

How Explicit to Be

Explication is not a choice. We cannot ponder, communicate, and exist collectively without explication and without rules built from explication. These were part of consciousness long before they had a name. Words themselves are simplifications of experience, misrepresenting but also enriching experience as we invent their meanings. Except for enumerations and appellations (e.g., ten students were present; the teacher was George), there are no veridical descriptors; no representations that mirror reality, none that draw us closer than experience to the real world. Explications stabilize meanings, facilitate discourse and often clarify. But many an effort to explicate does not do justice to meanings experienced or yearned for. Evaluators are met with anxiety wherever they go partly because people know that their accomplishments will not be fully represented. The competent evaluator strives to represent them validly without oversimplification. Explication needs strong counterforces to resist the tendencies to simplify.

Scriven has reflected upon the collective works of contemporary evaluators and found them wanting: too much judgment, too little discernible warrant in our syntheses. We do not challenge that assessment. He claimed that things would be better if the summation of grounds for value statements were rooted in more fully explicated criteria, standards, and decision rules. For support, he cited science, e.g., the prediction of eclipses and the diagnoses of disease. But we have heard Scriven's distinctions between science and evaluation too often (1991, p. 141, for example) to conclude that, in general, greater explication will improve value

resolution. We look forward, through metaevaluation, to the quality of value resolution in studies based on Scriven's approach to synthesis.

Relocation of Judgment

It requires intuition[15] and judgment to set criteria, weights, and standards and to codify needs, functions, and resources — just as it does to finalize a synthesis from informal data. Setting standards for evaluating the performance of administrators or agencies, as two examples, calls for thorough acquaintance with their functions and circumstances. It requires knowledge of ways in which their services are needed and what is available from others like them at what costs. Scriven's grounds, i.e., functions, needs, and resources, make sense to us and the subsequent criteria, weights, and standards can be useful. But whether we explicate these with precision or leave them impressional, we expect to exercise both judgment and rational thinking. Yes, we feel it important to be ready to answer an outsider's questions, "What were your criteria and how did you weight them?" even if we were not greatly aware of them before reaching our reported assessment of value.

We do not see it as important to decide which processes, intermediate or final, require the most judgment. It is important to ask, "Which processes generate the most satisfactory syntheses?" Or, "Given appropriate training, then which processes work better?" Right now, it is our estimate that we and our colleagues would do program evaluation studies poorly if based on formal explication of criteria and standards, worse than Scriven sees our profession doing now. But could we be trained to do a considerably more valid job by formalizing criteria and standards? We are not optimistic.

Scriven makes the important claim that by following the recommended inference operations, the subjectivity of evaluation will be reduced. As indicated above, he goes to considerable length to show that one need not rely on preference or personal value. We see his operations as moving the judgmental aspects from the final synthesis to the preliminary arrangements. Most inputs to a quantified QWS process require difficult interpretations: identification of needs, development of criteria, and assignment of weights. These call for the exercise of judgment. A similar operationalization could be applied to each of them to reduce the subjectivity there but it would relocate even earlier by setting those rubrics. We do not mean to say that the amount of total subjectivity is fixed. It can be reduced, but to place more emphasis on subjective operations in preliminary steps in order to reduce it in the final step does not seem to us a worthy plan for minimizing bias.

THE ESSENCE OF VALUING

Scriven (1994, p. 370) speaks of evaluation as having two essential parts, the standards claim and the performance claim, with synthesis being the subsequent combining of the two, and each of the two composed of many elements. From many ingredients comes a single value. In contrast, we see the act of evaluation as essentially a perception-judgment unity. Upon experiencing the evaluand, the evaluator feels, sometimes, not always, a recognition of quality.[16] Immediately on tasting the soup, the taster has a sensation of its goodness. Further inquiry or experience may modify the judgment but neither impulsive act nor deliberative inquiry necessarily includes recognition of the standards employed. Often, after the fact, one may tease out certain criteria, perhaps flavor, temperature, appearance, and hunger reduction, perhaps even

recognizing standards against which the taster experiences elation or rejection, but the satisfaction created by the experience does not demand such awareness.[17] It is the human, value-edged, perceptual response to stimulation, to the evaluand's being or doing, that is the essence.[18] The long work of evaluating a single program is adding to, and integrating and refining, the value-laden perceptions.

Whether or not formal recognition of criteria increases the validity of the evaluation is debatable. Granted, tasting soup, grading essays, and accrediting schools are likely to become more reliable with formal identification of criteria. But formally identified criteria only partially represent the complex criteria of experience. Often avowed criteria fail to identify poorly measured ingredients of experience. So we worry about the distortions of formal criteria. Still, preordinate criteria should help keep the evaluator from overlooking some important ingredients. Certainly, with them, the reliability of evaluation increases. So we welcome the guiding effects of criteria and resist their simplifications. In each new situation, we join Scriven in repeated challenging of preordinate criteria. We need to satisfy personal and collective judgment as to whether or not the criteria represent valid grounds for value resolution.

Standards Claims

Scriven says that standards claims are the epistemological key to evaluation.[19] According to common technical definition, a standard is a scale point (or multi-dimensional template or model). When performance exceeds the standard, the value category or consequential action changes. For example, below +.006 tolerance, the ball bearing is round enough; above, it is rejected. And, if its mean achievement score is above -2 sigma, the school is "approved;" below, it goes on the "Watch List." Ball bearings and schools are also evaluated non-numerically. Whenever an evaluation is made, one could say that an implicit standard was used. But to treat all evaluations as standards-based, as Scriven and Fournier have done, encourages unnecessary formalization.

Were evaluation fundamentally a rational process, standards could be considered the epistemological key to evaluation. As Scriven has pointed out, thermodynamics is the epistemological key to internal combustion engines but an auto mechanic needs little of that knowledge. His standards are the standards of practice, what has previously worked to effect repair. Working standards are more a by-product of practice, a rationalization, often useful, but, in evaluation work, seldom a known determinant of value. More than standards, judgment is the key to evaluation. Judgment is an intrinsic process, amenable to discipline and sophistication through standards, sometimes made to fixate on standards, but not fundamentally driven by them.

Scriven expressed concern about the evaluator's vulnerability in complex cases, saying that standards claims "are the most vulnerable point at which an overall evaluative conclusion can be attacked" (p. 370). It may be true. We do need to understand both internal and external vulnerabilities. We recognize that because, by Scriven's definition, constructing standards claims includes identifying functions, needs, criteria and weights as well as setting standards, it is not saying much to say that standards claims are a potential vulnerability. We will venture a guess that, for those of us in educational evaluation, measurement claims are as great a vulnerability. Both parts are vulnerable. Syntheses will benefit from better logic, better inference of standards claims, but generally, we think the way to better syntheses lies in another direction. We will look that way next.

OUR DIALECTIC ALTERNATIVE

As implied above, we would have most of the resources that can be put into professional development of program evaluators aimed at sharpening their judgmental powers and increasing external critique of their judgments. Some might go to making judgment more rational but most should go to providing better review mechanisms. Opportunities and expectations for self-reflection, and opportunities and obligations for professional critique, are what will do most for improving the value resolution process. We would submit perceptions of value to validity testing early as well as late. Critique, casual as well as formal, should be an ongoing process. Scriven mentioned this aim too, but considered it less a remedy than rule-based synthesis.

One of the ways the experienced evaluator regularly uses for gaining understanding and refining judgment is to study the evaluand dialectically, from different points of view, from different frames of reference (Habermas, 1979; Proppé, 1979; House, 1980; Guba & Lincoln, 1994).[20] Relevant here are the methods of triangulation discussed by Denzin (1984) and Flick (1992). Increasingly, they see triangulation as a process for assuring multiple viewpoints rather than reducing to a single viewpoint. Evaluators have long differentiated among conceptual frameworks for designing the study: goals, needs, issues, forthcoming decisions, discrepancies. Casually, at least, the competent evaluator shifts from one conceptual frame to another, often modifying the description, often refining notions of merit. It should not be a mark of competence to use but one model. Rather, evaluation practice is made more sophisticated by routinely drawing from a range of interpretive frameworks, iteratively trading one off against another, and provoking new interpretations. Thus, the complex merit of the evaluand is not likely to be lost in an index of value for each criterion. Rather, a contextualized discussion of multiple meanings of quality is envisioned.

It is our belief that our powers of evaluating are deeply rooted in experience. As a general contention, it seems reasonable to claim that human survival — in the face of socio-economic turmoil today as well as before the invention of language — can be attributed to our capacity for evaluating. With evolving language, communication and interdependence, successful living is more difficult, requiring constant evaluation, only some of which is conscious and deliberated. Successful program evaluation draws upon common-sense frames of reference as well as those developed by the learned disciplines. We are sensitive to highly complex phenomena and issues and we draw upon simple understandings. As did Chris's motorcycling father in Montana (Pirsig, 1974), we draw greatly on an ability to recognize "quality" even when we cannot satisfactorily define it.

We cannot interpret well all we see. Too much new information is encountered when visiting an evaluand. The human brain can only process so much information at the conscious level but handles more in other ways. Patterns cannot be discovered by formal analysis alone. According to Mintzberg (1994), we must use intuition as well as analysis to understand the larger picture. While analysis is breaking down to smaller pieces, intuition is pulling hidden patterns together so that we can make better judgments about the evaluand. Wertheimer, Kohler and Koffka called this *Gestalt:* understanding the big picture without total reliance on analysis (Angeles, 1981; Flew, 1984). Intuition, quality recognition, and judgment are resources not to be lightly set aside. If they are hurting us more than helping us, we need to upgrade them or replace them. We believe they often can be upgraded. We believe that an evolutionary regimen of dialectic critique is the best way to improve the final syntheses.

Unified Concept of Value

In "The Final Synthesis," Scriven speaks of program merit as ultimately singular, seldom situation specific.[21] He calls for a unified concept of value. Scriven acknowledges that different levels of merit may appear for major differences in program function but, with function specified, he urges culmination of the search for merit in a single synthesis. Speaking about the validity of evaluation in the *Evaluation Thesaurus,* Scriven wrote, "One should not talk about 'valid in this sense, but not in that,' only about 'validity (of the appropriate type).'" He added, "... validity, it has become clear... is a unitary concept;... " (1991, p. 372).[22] An effective final synthesis produces a valid concept of merit of the evaluand.

We prefer to respect the multiple perceptions of value in the situation.[23] Often the evaluator will be the one who has studied the situation with the greatest discipline, showing the greatest restraint upon personal preference and ideology. But never will the evaluator have all the best data, all the best reasoning, and all the best motive for perceiving value. Collecting the values of others, skeptically scrutinizing them, and reporting — without insistence on the superiority of the evaluator's determination of merit — is not just modesty; it respects the complexity of merit. And partly because we interact with clients and stakeholders, the occasions for synthesizing data and values are many. Synthesizing occurs long before the "final" synthesis and the syntheses continue to evolve long after.

We have no reason to believe that the value of a program is simple. Our reasoning persuades us that the merit of a program is complex and conditional. It is good for some things some of the time, better and poorer at other times. As personal, social, economic, political, and cultural conditions change, the worth of the program changes.[24] We want to describe these conditions. We do not want our representation of merit to be consistently simpler than the merit really is. We discourage the use of rubrics which assure a simple picture. We want to describe the complex "values manifold" to be part of the summary report.

Scriven defines merit in terms of performance. Even if we could measure performance well, we would not define merit only in terms of outcomes. Summative evaluation should portray also how well the program was constituted and operated. Following MacDonald (1977), House (1993) and others, evaluators have an ethical commitment to contribute to a just society. Programs designed to contribute to improvement in social well-being are to some degree meritorious just by existing. Following state of the art practices even in the absence of measured effects is to some extent meritorious. Scriven (1990, p. 17) thinks of these contingencies as candidates for performance weights but many readers will not. We find it troubling to define merit in terms of performance alone. Merit is not productivity.

Scriven works toward parsimony, unification, synthesis. We are working toward the perspectival, the conditional, the comprehensive. In a 1990 diagram of his pertaining to product evaluation (shown as Figure 1), he would have us first aggregate indicators of merit, analyzing performance, then begin a synthesis process, working through conclusions about indicators and dimensions to reach a unified resolution of value. Our process also starts small but ends wide, compound and complex, visualized as only the upper pyramid, but with both analysis and syntheses in that one pyramid together. We would capitalize on a concept of *unified* perception and value and encourage work toward multiple meanings of value. Our figure's endline would be wide-bottomed to show that, although we too report far less than we could, the summary of value is expansive. We urge that the representation of value not be simpler than the complex of judgments made. We resist any one best simplex of values. For different purposes and populations and contexts and times, the program has different values. We will con-

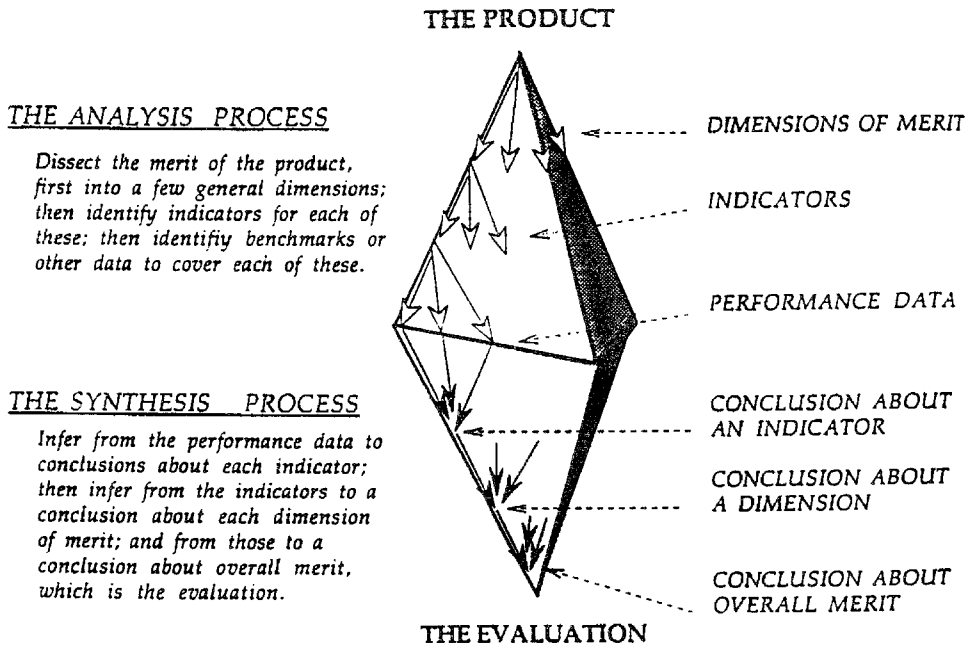

Figure 1. The Evaluation Double Pyramid (from Scriven, 1990).

Note: From Scriven (1990 p. 16). Reprinted with kind permission from Scriven and Elsevier Science Ltd (the copyright holder), The Boulevard, Langford Lane, Kidlington OX5 1GB, UK.

strue value, as Scriven does, in ways meaningful to a large and diverse community of readers but we will organize our data-gathering, analysis, and reporting in ways that (hopefully) help readers make up their own minds. Thus the final syntheses in many instances will not be those of the evaluators.

CONCLUSION

Scriven, like many others in the field, wants the practice of evaluation to change. In "The Final Synthesis," he attempts to guide us toward better evaluations by focusing on a particularly undisciplined practice — conceptualizing the evaluand's merit. In Scriven's words, "the lack of any explicit justification of the synthesis procedure is the Achilles heel of the whole [evaluation] effort" (p. 367). To reduce prejudice and bias in this process, he urges us to establish rules to govern our final determination of value. Scriven wants to move us away from a model of personal, intuitive judgment to one of analytic, objective heuristics.

The general evaluation methodology endorsed by Scriven revolves around better explication of what constitutes good performance. Most will read his writing as calling for a series of sequenced steps: developing and weighting criteria of merit, setting standards of excellence on each criteria, measuring performance of the evaluand on the criteria of merit, comparing the performance to the standards, and synthesizing the individual scores into a final judgment of quality. The process requires expertise in formal explication.

Will his methods cover our vulnerable heel? Can we produce less prejudiced evaluations through careful application of Scriven's general evaluation methodology? Will we produce better evaluations otherwise? In this paper, it has been our contention that his synthesis methods will not substantially reduce bias or produce better evaluations.

Scriven is correct in saying that most practicing evaluators do not pay much attention to a formal, explicit synthesis process. Many practitioners concentrate on elaborate designs, painstaking data collection, and careful data analysis. The final step is, as Scriven comments, often "simply a reaction to the rest, a high inference judgment (p. 367)." We have argued that this is the way it should be. Judgment is the natural avenue to determination of value. In fact, after interviewing a small group of experienced evaluators, it has become clearer to us that a formal, explicit and systematic synthesis process is alien to current practice. For many of us, here is how it really happens: we come to understand the evaluand and, directly in so doing, we come also to know its merit. Perceiving and valuing are an integrated act.

Seeking to reduce bias in the final synthesis, Scriven has left open the exercise of bias at earlier stages, particularly in the selection and weighting of criteria of merit. Even in designing a needs assessment involving diverse populations, we make difficult choices, more than a few estimates, each a node of potential bias. Thus, biases are possibly but not necessarily reduced, just shifted. The illusion of objectivity is strong, but bias is not controlled.

Scriven encourages evaluators to develop and progressively refine criteria and standards before critical measurements are taken, agreeing with Parlett and Hamilton (1976) that a strictly preordinate design is constraining. Evaluators come to understand a program, its issues and its merits and shortcomings progressively during continued contact. Criteria and weights change during the evaluation process. People, evaluators among them, do not naturally withhold judgment until systematic data collection can proceed. Upon first contact with an evaluand, meaning is constructed. The assessment of quality is part of that natural meaning-making process. We undoubtedly will refine our first impressions, and criteria and standards may appear along the road to determining merit, but they are only part of the dialectic process. Elsewhere (1981, p. 127) Scriven endorses this *gradual refinement of criteria* but did not make it clear in "The Final Synthesis."

Through extensive work in the field, the office, and the library, successful evaluators develop both explicit and implicit knowledge for recognizing program quality. Part of the inference process remains beyond awareness, but is not necessarily invalid. Criteria, standards, and weights shift from background to foreground as the determination of quality evolves from the study. The notion of a single final synthesis is inappropriate — the resolution of value has been continuous throughout the investigation. Final clarification and final critiques occur. Value recognition doesn't squeeze out of a final heuristic, but rather develops from the first contact with the evaluand — evolving, reshaped and dialectically formed, until the final words are written in the report.

Scriven has directed our attention to an evaluation's final synthesis. He tells us that expert judgment, intuition, or personal preference is not good enough when it comes time to decide an evaluand's merit or worth. Yet those are the natural human responses. We encounter a stimulus and immediately try to make sense of it — part of that sense-making involves an immediate assessment of quality. Evaluators need to embrace this natural tendency while continually challenging our personal interpretations of merit. We refine our value constructs by applying one conceptual structure after another. By downplaying formalization in general, we are not entitled to a free ride — disciplined judgment becomes all the more critical.

Evaluation practitioners owe Scriven an enormous appreciation. One cannot read his writings without thinking deeply about theory and practice. His strong convictions and forthright style force the reader to reconsider many aspects of practice that may have been taken for granted. Although we disagree with some of his persuasions, he has made us think seriously. Our conduct of evaluations will not be the same now that we have examined "The Final Synthesis." Hopefully they will be more reflective and rigorous, possibly better, but probably not more formal or more explicit.

NOTES

1. Authored by the 12 members of a seminar on Theories of Educational Evaluation, Spring, 1996, University of Illinois. Scriven read an intermediate draft and spent three hours with the authors discussing it.

2. Judgment is taken to mean the mental resolution of values by reasoning guided and tempered by human experience. A mathematical rubric for grading or selecting, however thoughtfully devised, is not considered here to be judgment. Judgment is subjective, necessarily so for the empathic understanding of human affairs.

3. When we asked him if he endorsed Fournier's interpretation, Scriven said he did.

4. Fournier said, "... numerous evaluation approaches developed by theorists vary from one another in many details, yet I find that they do share this common logic." (1995, p. 16).

5. They were Anthony Blair, Debra Bragg, Thomas Cook, Lee Cronbach, Norman Denzin, Lizanne DeStefano, Thomas Grayson, Michael Patton, Michael Ravitch, Kathryn Sloane, Robert Stake, Daniel Stufflebeam and Tim Wentling.

6. What ordinary practice involves is important as a practical consideration because if evaluators are already doing intuitively what Scriven says should be done rationally, then it should not be so difficult for the profession to improve. But if a different logic is being followed, then the logics should be compared, noting costs of upgrading practice.

7. This reinterpretation of the meaning and use of criteria is consistant with Schwandt's observations on the passing of "criteriology" (1996).

8. In 1975, Scriven's four steps for reducing evaluation bias were (a) replace judgment with mensuration, (b) do better training, (c) do goal-free evaluation and (d) use more design controls and comparisons (p. 27). We agree on training.

9. Bolman and Deal (1991, p. 25) have written of the problems of using a rational approach to study organizations.

10. Toumin (1983) wrote, "...we should ask scholars to pay more attention to the elements of interpretation — even of hermeneutics — that have nowadays become essential to both the natural and human sciences" (p. 101).

11. Schwandt (1995) also challenged the Fournier representation of evaluation logic as pertinent to a particular context, the modernist-academic-industrial context, but not pertinent to others.

12. Chandler (1973, p. 855) concluded that the cause of Napoleon's defeat at Moscow was the overextension of his supply lines. Tolstoy (1869/1957) suggested the soldier who burned his beans at breakfast was no less a determining force. Chandler was the essentialist, Tolstoy the relativist.

13. For a detailed analysis of the difference in mathematics achievement as seen by teachers and psychometricians, see Stake (1995). For additional discussion of writing skill assessment, see Breland (1996).

14. An example of an arbitrary criterion was published by Consumers Reports in describing the 1946 Hudson, saying that the windows were too small for a man with a hat on to stick his head out.

15. In 1966, Scriven said, "... intuition or judgment does not carry built-in guarantees any more than it is a condensed reasoning process. Once it has passed its qualifying examination, however, a new

judgment skill is just as respectable a part of the armory of reason as any well-tested instrument is part of the apparatus of science" (p 15). We would say that the intuitive powers of many program evaluators we know have passed the exam but further refinement is sought. Ways of refining intuition in evaluation were identified by Sadler (1981).

16. Scriven (1994, p. 370) says the expert has internalized many criteria. For expert and novice alike, we believe that many criteria have originated intuitively, some never becoming externalized.

17. In denying awareness of a standard as a prerequisite to judgment, we do not want to suggest that evaluators avoid implicit standards. For some of us writing here, McDonald's are an abiding standard for French fries.

18. Messick says that "value implications are not ancillary but, rather, integral to score meaning" (1995, p. 748).

19. Scriven (1994) goes to considerable length to demonstrate that many standards are matters of fact (e.g., good acceleration is a merit in a car) in an effort to include, in his definition of evaluation, the logical treatment of facts and performances (p. 372). In the circles in which program evaluators work, there is little demand for the exercise of that logic. Expert judgment is not only tolerated, it is admired.

20. Rettig (1990) said, "... people act in accordance with the world of reality they have construed... even though such constructions are different for different people (e.g., scientists, lawyers, truck drivers). Not only does the evidence originate in different sources, but is also differently organized. Hence, similar behaviors may have different meanings while very different behaviors may be alike in terms of the perceived evidence on which they are based. To understand these differences in social reality it is necessary to penetrate the world of others (p. 12)."

21. In his paper, Scriven makes a strong claim that it is not the work of evaluation to hang around after the reporting to help translate findings into action (p. 378). He faults those who are essentially remediationists or activists masquerading as professional evaluators. The twelve of us did not agree on all the positions taken in our response and particularly not on the question of whether the evaluator should merely report value resolution or help readers construct their own meanings of merit and shortcoming. For the most part, we join Scriven in refraining from making recommendations and otherwise participating in the utilization of study results but we deflect some of the resources for our final syntheses into making it easier for the readers to make theirs. Helping clients understand how the current evaluand fits into existing panoplies of value is a part of the evaluator's responsibility.

22. Scriven is strongly opposed to Messick's (1995) arguments for "consequential validity" (p. 742) because they treat evaluation as invalid if bad consequences can be attributed to it. We believe Scriven would say that messages can have integrity even if no one can understand them and that that integrity is the focus of metaevaluation. Further development of these arguments are beyond the scope of this paper.

23. Following Lévi-Strauss (1966, p. 7) and Denzin and Lincoln (1994, p. 2), we like the idea of researcher as bricoleur, a "jack of all trades" who adapts various ideas and tools to the solving of intellectual problems, resisting the idea of one best way of resolving value claims.

24. Worth and merit are not the same, but views of merit are influenced by worth perceived.

REFERENCES

Angeles, P. A. (1981). *Dictionary of philosophy.* New York: Barnes and Noble.

Bolman, L. G., & Deal, T. D. (1991). *Reframing organizations: Artistry, choice, and leadership.* San Francisco: Jossey-Bass.

Chandler, D. G. (1973). *The campaigns of Napoleon.* New York: MacMillan.

Denzin, N. (1984). *The research act.* Englewood Cliffs, NJ: Prentice Hall.

Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research.* Newbury Park, CA: Sage.

Diederich, P. B. (1957). The improvement of essay examinations. Research Memorandum RM-57-3. Princetion, NJ: Educational Testing Service.

Educational Testing Service. (1996, April). *Writing skill assessment: Problems and prospects* (Policy Issues Perspective series). Princeton, NJ: Breland, H. M.

Flew, A. (1984). *A dictionary of philosophy; Revised second edition.* New York: St. Martin's Press.

Flick, U. (1992). Triangulation revisited: Strategy of validation or alternative? *Journal for the Theory of Social Behaviour, 22*(2), 175–198.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. In D. M. Fournier, (Ed.), *Reasoning in evaluation: Inferential links and leaps. New Directions for Evaluation, 68.* (pp. 15–32). San Francisco: Jossey-Bass.

Guba, E. G., & Lincoln, Y. S. (1994) Competing paradigms in qualitative research. In N. K. Denzin and Y. S. Lincoln, (Eds.), *Handbook of qualitative research.* (pp. 105–117). Newbury Park, CA: Sage.

Habermas, J. (1979). *Communication and the evolution of society.* Boston: Beacon Press.

House, E. R. (1980). The logic of evaluative argument. In House, E. R. (Ed.), *Evaluating with validity.* (pp. 67–96). Beverly Hills, CA: Sage.

House, E. R. (1993). *Professional evaluation.* Newbury Park, CA: Sage.

Lévi–Strauss, C. (1966). *The savage mind (2nd ed.).* Chicago: University of Chicago Press.

MacDonald, B. (1977). A political classification of evaluation studies. In D. Hamilton (Ed.), *Beyond the numbers game.* (pp. 224–227). London: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Mintzberg, H. (1994). *The rise and fall of strategic planning: Reconceiving roles for planning, plans, planners.* New York: The Free Press.

Parlett, M., & Hamilton, D. (1976). Evaluation as illumination: A new approach to the study of innovatory programmes. In G. V. Glass, (Ed.), *Evaluation Studies Review Annual, 1,* (pp. 140–157). Beverly Hills, CA: Sage.

Pirsig, R. M. (1974). *Zen and the art of motorcycle maintenance.* New York: William Morrow.

Petrie, H. (1995). Purpose, context, and synthesis: Can we avoid relativism? In D. M. Fournier, (Ed.), *Reasoning in evaluation: Inferential links and leaps. New Directions for Evaluation,* 68. (pp. 81–91). San Francisco: Jossey-Bass.

Proppé, Ó. (1979). *Dialectical evaluation.* Urbana, IL: Center for Instructional Research and Curriculum Evaluation (mimeo).

Rettig, S. (1990). *The discursive social psychology of evidence: Symbolic construction of reality.* New York: Plenum Press.

Sadler, D. R. (1981). Intuitive data processing as a potential source of bias in evaluation. *Educational Evaluation and Policy Analysis, 3*(4), 25–31.

Schwandt, T. A. (1995). *Indexing the practice of reasoning in evaluation.* Unpublished manuscript, College of Education, Indiana University.

Schwandt, T. A. (1996). Farewell to criteriology. *Qualitative Inquiry, 2*(1), 58–72.

Scriven, M. (1966). *Primary philosophy.* New York: McGraw-Hill.

Scriven, M. (1967). *The methodology of evaluation.* AERA Monograph Series on Curriculum Evaluation, 1. Chicago: Rand, McNally.

Scriven, M. (1975). *Evaluation bias and its control* (Occasional paper #4). Kalamazoo, MI: Evaluation Center, Western Michigan University.

Scriven, M. (1980). *The logic of evaluation.* Inverness, CA: Edgepress.

Scriven, M. (1981). Product evaluation. In N. L. Smith, (Ed.), *New Techniques for Evaluation* (pp. 121–166). Beverly Hills, CA: Sage.

Scriven, M. (1990). The evaluation of hardware and software. *Studies in Educational Evaluation, 16,* 3–40.

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

Scriven, M. (1992). Evaluation and critical reasoning: Logic's last frontier? In R. A. Talaska, (Ed.), *Critical reasoning in contemporary culture* (pp. 353–406). Albany: SUNY Press.

Scriven, M. (1994). The final synthesis. *Evaluation Practice, 15,* 367–382.

Scriven, M. (1995). The logic of evaluation and evaluation practice. In D. M. Fournier, (Ed.), *Reasoning in evaluation: Inferential links and leaps. New Directions for Evaluation, 68.* (pp. 49–70). San Francisco: Jossey-Bass.

Stake, R. E. (1995). The invalidity of standardized testing for measuring mathematics achievement. In T. A. Romberg, (Ed.), *Reform in school mathematics and authentic assessment.* Albany, NY: State University of New York Press.

Tolstoy, L. (1957). *War and peace.* R. Edmonds (Trans.) Harmondsworth, England: Penguin. (Original work published 1869)

Toulmin, S. (1983). The construal of reality: Criticism in modern and postmodern science. In W. J. T. Mitchell, (Ed.), *The politics of interpretation* (pp. 99–117). Chicago: University of Chicago Press.