

The Essentials of Coaching Program Evaluation:

Formative, Summative and the Four Ds

William Bergquist

[Note: the following essay is based on a chapter I wrote for a book on organizational improvement quite a few years ago. This chapter never made it in the final version of the book. I have updated and focused in this revised version on the evaluation of coaching programs. The references may be a bit old, but the wisdom offered by these program evaluation experts is still of great benefit -- and sadly is still being ignored with regard to many program evaluation initiatives.]

"The tragedies of science," according to Thomas Huxley, "are the slayings of beautiful hypotheses by ugly facts." The leaders and managers of coaching programs in many organizations face this prospect when confronted with the need to design or select ways of evaluating their efforts.

Program evaluation may indeed be threatening to their cherished notions about how human and organizational resources are developed and about how change and stabilization actually take place. More immediately, evaluation can be threatening to one's beliefs regarding how a particular coaching project is impacting a particular department or the entire organization.

In this essay, I will review a series of appreciative concepts and tools that can reduce this threat by making the evaluative process clearer and more supportive. Effective program evaluation is a process that can be uncomfortable, for all growth and change involve some pain. Program evaluation, however, can be constructive. Furthermore, if it is appreciative, this evaluation process can meet the needs of both those who are serving and those who are being served by the coaching program.

I offer a brief excursion through the history of program evaluation and, in particular, through the major issues regarding the purposes that program evaluation serves and the various forms that program evaluation takes in serving these purposes. Probably the most important fact to keep in mind is that program evaluation has not historically been commonly used in most sectors of society. Talk has been cheap when it comes to thoughtful and systematic program evaluation. There has been a fair amount of

conversation about this type of assessment work in an organization, but it has not often actually been enacted. In the past, much of the work done in this area was confined to educational programs and, in particular, to the evaluation of programs for funding purposes or for continuing accreditation or authorization. Many of the advances in evaluation were made by members of or consultants to major philanthropic foundations (such as the W. K. Kellogg Foundation and the Lilly Endowment) or the United States Federal government who were asked to determine the worth of a program that has been funded or may be funded by their institution. Other advances have been made by those given the task to determine if a school or college should be granted a specific accreditation status.

Program evaluation has also been widely used in the sciences, criminal justice, medicine and social welfare, once again often associated with the assessment of program worth by governmental funding agencies. Following Sputnik, increasing attention was given to the achievements of American research initiatives, while attention also increased regarding the success of heavily-funded social programs under the banners of "The Great Society" and "War on Poverty." In more recent years, program evaluation has become more common within corporations and nonprofit organizations and in health care delivery systems. In most cases, this growing interest is unrelated to outside funding sources; rather, it emerges from a growing concern about quality products and services, and the growing concern about assessing costs and benefits associated with specific program offerings. Return-on-investment is now a commonly used (though often misunderstood and misused) frame of reference for corporate program evaluation. Similarly, in health care, "evidence-based" medicine has emerged as a way to determine treatment strategies and funding priorities.

Accompanying this expansion in the size and scope of program evaluation initiatives is the maturation of the field. A clearer understanding of the differing functions played by specific evaluation strategies has been complimented by a clearer sense of those features that are common to all forms of program evaluation. The most important distinction that has been drawn for many years regarding the purpose of program evaluation concerns the use of evaluation processes to determine the worth of a program and the use of evaluation processes to assist in the improvement of this program. The terms used to identify these two functions are *formative* and *summative*.

Formative and Summative Program Evaluations

A noted educational researcher, Paul Dressel, differentiated several decades ago between summative evaluation that involves "judgment of the worth or impact of a program" and formative evaluation that

Dressel defines as "the process whereby that judgment is made." The evaluator who is usually identified as the author of this distinction, Michael Scriven, offers the following description of these two terms.

According to Scriven, formative evaluation:

. . . is typically conducted during the development or improvement of a program or product (or person, and so on) and it is conducted, often more than once, for the in-house staff of the program with the intent to improve. The reports normally remain in-house; but serious formative evaluation may be done by an internal or external evaluator or (preferably) a combination; of course, many program staff are, in an informal sense, constantly doing formative evaluation.

As described by Scriven, summative evaluation:

. . . is conducted after completion of the program (for ongoing programs, that means after stabilization) and for the benefit of some external audience or decision-maker (for example, funding agency, oversight office, historian, or future possible users), though it may be done by either internal or external evaluators or a mixture. The decisions it serves are most often decisions between these options: export (generalize), increase site support, continue site support, continue with conditions (probationary status), continue with modifications, discontinue. For reasons of credibility, summative evaluation is much more likely to involve external evaluators than is a formative evaluation.

Scriven borrows from Bob Stake in offering a less formal but perhaps more enlightening distinction between formative and summative: "When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative." From an appreciative perspective, formative evaluation can be said to be an exercise in fully understanding the complex dynamics and causal factors influencing the operation of a program and taking corrective action if needed. By contrast, a summative evaluation allows one to identify and build on the specific successes and strong features of a specific program unit. Both formative and summative evaluations can be appreciative, and the comprehensive appreciation of any program unit involves both formative and summative evaluation processes.

Program Planning and Evaluation

Ed Kelly, an experienced program evaluator who served on the faculty of Syracuse University for many years, further differentiated judgments concerning the extent to which the intentions of the program

were satisfied and judgments concerning whether or not the program was any good. Concern for judgment necessarily involves issues of values, criteria, goals, customers and audience. Concern for evaluative process necessarily involves issues of method, instrumentation and resources. Both approaches to evaluation require a clear definition of *clientship*, a precise sense of the role of evaluation and an explicit understanding of the way the judgment or process will be used by the program staff and others.

In essence, program evaluation involves the development of a process whereby program activities can be interrelated and compared to program expectations, goals and values. The nature of this interrelationship will vary considerably. In some instances, external assistance will be required to establish the process, while in other instances the external assistance will be used to provide the interrelationship judgments once the process has been defined. In yet other instances, the external assistant (evaluator) both identifies the process and provides the judgments.

Regardless of the process being used, an effective program evaluation effort will commence with the initial planning of the program. In planning for any program, or in deciding on the initiation of a proposed program, the processes of evaluation are inevitably engaged. Those who plan the program will be concerned with the validity of their assumptions about needs, strategies and resources. Those who review their proposal will ask questions about feasibility, attractiveness and probable success. Others will ask how program achievement is to be measured. Program evaluation is not a topic to be addressed at the end of a planning process; rather, program evaluation should be a vital and influential element that is discussed at the beginning and given serious consideration throughout the process--especially if the evaluation is being engaged in the assessment of a complex process such as professional coaching.

The Four Ds of Program Evaluation

There are four basic types of program evaluation: (1) description, (2) documentation, (3) determination of outcomes, and (4) diagnosis. I identify these as the "four Ds." An outcome determination evaluation is conducted primarily for the purpose of judging the degree to which a program achieved its intended goals and outcomes. This "summative" approach aids decision-making about the continuation of the program. Ongoing decision making concerning the nature, content and scope of a program are best addressed through use of diagnostic evaluation. This type of evaluation is "formative" in nature, since it is conducted while a program is in progress and is used to continually or intermittently refine and improve the program. Program evaluations often are of greatest value when they aid the dissemination

of program results. Descriptive and documentary approaches to program evaluation are most often employed when dissemination is critical. Descriptive evaluation tells other people about the nature and scope of a program. Documentary evaluation provides evidence for the existence of the program and its outcomes. It illustrates the nature of the program and its impact. Following is a more detailed description of each of these four types of program evaluation.

Description of Program

The first feature in any program evaluation, according to Scriven, is the identification of the program unit(s) being evaluated. He suggests that this identification should be based in a comprehensive description of the program being evaluated. Thus, program description is always the first element of a program evaluation. It is also one of the final elements, for any final evaluation report will typically contain a description of the program being evaluated. Consequently, there is little need to spend much time advocating the importance of or identifying procedures for the description of a program. Nevertheless, most program descriptions can be improved. Given the importance of dissemination, one must be certain not only that information about the program is accurate and complete, but also that other people understand the program description.

Scriven suggests that a successful program description is something more than just the labeling of program components. I would propose that an appreciative approach to program evaluation also requires something more than a cursory classification or labeling of a program. It requires that the distinctive and most salient features of the program be identified and carefully described. A program description often serves as a guidebook for successful program replication if it has been prepared in an appreciative manner. It also often probes into the true function and meaning of a specific program.

Edward Kelly takes description and appreciative evaluation a step further in suggesting that one of the most important purposes of an evaluation is the provision of accurate and compelling portrayals – a vivid depiction or reconstruction of complicated social realities. Those people who are not present when an event occurs should have a valid and useful understanding of what it must have been like to be there. Kelly notes:

A portrayal is, literally, an effort to compare a rendering of an object or set of circumstances Portrayal evaluation is the process of vividly capturing the complexity of social truth. Things change depending on the angle from which they are viewed: multiple renderings or multiple portrayals are intended to capture the complexity of what has occurred.

In order to prepare an accurate description of a program, it is necessary not only to trace the history and context of the program and describe its central activities and/or products, but also to provide a portrait of the program (brief descriptions, quotations, paraphrases, and observations). What was it like being a coaching client? What did a typical client do differently on a daily basis as a result of participating in this coaching program? What was it like to walk into the office where this program was being enacted? How has this coaching program affected the perspectives (and actions) of the C-Suite leaders in this corporation?

Rather than always focusing on specific program activities, it is often valuable to focus on a specific program participant. Pick a "typical" person who has received coaching services. In what coaching activities did she engage? What worked for her? What didn't she like? Why? One might even want to create a hypothetical participant who represents "normal" involvement in the coaching program. A case history can be written that describes this hypothetical participant in the program. This case history can be much more interesting and in some sense more "real" than dry statistics, though the case needs to be supported by statistics to ensure that this typical person is, in fact, typical.

Documentation of Program

The most straightforward type of evaluation is documentation. When someone asks what has happened in a program or whether a program has been successful, the program staff can present the inquirer with evidence of program activity and accomplishment. Program evaluations that do not include some documentation run the risk of appearing sterile or contrived. One reads descriptions of a program and one even reviews tables of statistics concerning program outcomes but never sees "real" evidence of the program's existence. An appreciative evaluation always provides this real evidence. It discovers the footprints left by a program unit and appreciates the meaning of these footprints.

Some program evaluators even suggest that we are eventually led in program documentation to a "goal-free" evaluation. The documents speak for themselves and there is little need for an often biasing and limiting set of goals by which and through which an evaluator observes a specific program. Program documents often reveal much more about a program than is identified in a set of goals. Through the documents, one sees how a program is actually "living," and what emanates from the program that may or may not conform to its pre-specified goals.

Often after a program has been developed, someone will collect all the documents that have been accumulating during the course of the program. This may include minutes from major meetings,

important emails and letters, reports, formal and informal communications about specific program activities or products, productions of the program, video recordings of specific program activities, and so forth. These documents are usually stored in some computer file (or in the Cloud) for some vaguely defined use in the future. Often one suspects that the documents are stored to avoid the arduous task of sifting through them and throwing away the old, useless ones. Unfortunately, archives frequently are not used at a later date. As a result, the collection and storage of documents is rarely a rewarding or justifiable procedure in program evaluation.

Several problems are inherent in typical documentation processes. First, the documents often are stored with no master code. One can retrieve a document only by combing through vast arrays of irrelevant material or identifying some vaguely appropriate search terms. Even more importantly, there is rarely a summary documentation report that highlights the richness and value of the stored documents. Nothing entices one to explore the documents. Third, the documentation is usually not linked directly to the purposes or expected outcomes of the program and remains isolated from other aspects of the total evaluation. Many of the problems usually associated with documentation can be avoided if a systematic and comprehensive documentation procedure is implemented.

Determination of Program Outcomes

The third type of program evaluation is both the most obvious and most difficult. It is the most obvious because the term "evaluation" immediately elicits for many of us the image of judgment and assignment of worth. Has this program done what it was intended to do? Has this program done something that is worthwhile? Outcome determination evaluation is difficult because the two questions just cited look quite similar on the surface, but are, in fact, quite different. To know whether a program has done what it was supposed to do is quite different from knowing whether what it has done is of any value. The old axiom to the effect that "something not worth doing is not worth doing well" certainly applies to this type of evaluation. The problem is further compounded when an appreciative approach is taken to program evaluation, for both questions are important when seeking to appreciate a program unit. In the summative appreciation of a program's distinctive characteristics and strengths, we must assess not only the outcomes of the program, but also the value to be assigned to each of these outcomes. We also might wish to relate the outcomes (benefits) to the investments made in this program. What is the return on investment?

The first of these two outcome determination questions (Has it done what was intended?) is "researchable." We usually can determine whether or not a specific set of outcomes have been achieved. The second question (Was it worth doing?) requires an imposition of values. Hence, it is not "researchable." We can't readily answer this question without substantial clarification of organizational intentions. Yet the issue of values and organizational intentions cannot be avoided in the determination of outcomes. In another Library of Professional Coaching document ([Intentional Analysis: A Comprehensive and Appreciative Model for the Evaluation of Organizational Coaching Programs](#)) I examine ways in which the second question regarding the value of a program can be handled. In this essay, I explore ways in which the first question regarding achievement of pre-specified outcomes can be addressed.

Determining the Achievement of Prespecified Outcomes: There are two levels at which a program can be evaluated regarding the achievement of predetermined outcomes. At the first level, one can determine whether the outcomes have been achieved, without any direct concern for the role of the program in achieving these outcomes. This type of outcome-determining evaluation requires only an end-of-program assessment of specific outcomes that have been identified as part of a program planning process. To the extent that minimally specified levels have been achieved, the program can be said to have been successful, though, of course, other factors may have contributed to or even been primarily responsible for the outcomes. If one needs to know specifically if the program contributed to the achievement of those outcomes, then a second set of procedures also must be used.

Determining a Program's Contribution to the Achievement of Pre-specified Outcomes: This type of assessment requires considerably more attention to issues of design and measurement than does an assessment devoted exclusively to the determination of outcomes. In order to show that a specific program contributed to the outcomes that were achieved, a program evaluator should be able to demonstrate a causal connection. For example, a coaching program evaluation should show that one or more comparable groups of potential coaching clients who did not receive these coaching services did not achieved the pre-specified outcomes to the extent achieved by coaching clients who did receive these services.

In order to engage this comparison between a group that has participated in a coaching program, called the "experimental" group, and a group that hasn't participated in this program, called the "control" group, several research design decisions must be made. Most evaluators try to employ a design in which people are assigned randomly to the experimental and control groups, and in which both groups are

given pre- and post-program evaluations that assess the achievement of specific outcomes. Typically, the control group is not exposed to any program. Alternatively, the control group is exposed to a similar program that has already been offered in or by the organization. In this situation ideally there should be at least two control groups for the study of coaching practices, one that receives no coaching services and the other that receives an alternative to the coaching program being evaluated (such as a leadership training program or an alternative form of coaching).

While this experimental design is classic in evaluation research, it is difficult to achieve in practice. First, people often can't be assigned randomly to alternative programs. Second, a control group may not provide an adequate comparison for an experimental group. If members of a control group know that they are "controls," this will influence their attitudes about and subsequently their participation in the program that serves as the control. Conversely, an experimental group is likely to put forth an extra effort if it knows its designation. This is what is often called "The Hawthorne Effect." It may be difficult to keep information about involvement in an experiment from participants in either the experimental or control group, particularly in small organizations. Some people even consider the withholding of this type of information to be unethical.

Third, test and retest procedures are often problematic. One cannot always be certain that the two assessment procedures actually are comparable in assessing a coaching client's performance, behavior, attitudes, knowledge or skills before and after a program. Furthermore, if there is no significant change in pre- and post-program outcome measurements, one can never confidently conclude that the program had no impact. The measuring instruments simply may be insensitive to changes that have occurred. On the other hand, the coaching clients already may be operating at a high level at the time when the pre-test is taken and hence there is little room for improvement in retest results. This is the so-called "Ceiling Effect."

A control group can solve some of these test/retest problems, because if the problems are methodological, they should show up in the assessment of both groups. However, one must realize that the pretest can itself influence the effectiveness of both the experimental and control group programs and thus influence the two groups in different ways. Fourth, several logistical problems often are encountered when a classic experimental design is employed. In all but the largest organizations there may not be a sufficient number of people for a control group. There also may not be enough time or money to conduct two assessments with both an experimental and control group.

Given these difficult problems with a classic experimental design, many program leaders and program evaluators may have to adopt alternative designs that are less elegant but more practical. In some cases, leaders and evaluators have restricted their assessment to outcome measures. They determine the level of performance achieved by a group of coaching clients and use this information to determine the relative success of the program being evaluated. This type of information is subject to many misinterpretations and abuses, though it is the most common evaluation design being used in contemporary organizations.

The information is flawed even when a comparison is drawn with coaching programs in other divisions of the organization or in other organizations. One doesn't know if differences in performance of the coaching clients can be attributed to the coaching program being reviewed or to the entering characteristics of the clients. Did clients in the alpha division or at the alpha organization do better than clients in the beta division or at the beta organization because alpha clients were already better trained or working at a higher level than beta clients before they even entered the coaching program?

This confounding effect is prevalent in many of the current evidence-based initiatives and even the ROI investigations that call for clients to perform at a certain level without any consideration being given to their level of performance upon entering the coaching program. In order to be fair in the assessment of a coaching program's effectiveness, one must at the very least perform a "value-added" assessment. This type of assessment requires that a coaching client's performance be measured when they first enter the coaching program and again when they "graduate" from the program to determine the "value" that has been added, or more specifically the improvement in performance that has been observed and recorded.

Fortunately, there are ways in which to assess program outcomes accurately and fairly, without having to engage a pure experimental design that may be neither feasible nor ethical. Two of the most widely respected authorities in the field of program evaluation, Donald Campbell and Julian Stanley, described a set of "quasi-experimental" designs that allow one to modify some of the conditions of a traditional experimental design without sacrificing the clarity of results obtained. Campbell and Stanley's brief monograph on experimental and quasi-experimental designs is a classic in the field. Any program evaluator who wishes to design an outcome determination evaluation should consult this monograph. Three of the most widely used of these quasi-experimental designs are "time series," "nonequivalent control group design" and "rotational/counterbalanced design."

Campbell and Stanley's "time-series" design requires that some standard measure be taken periodically throughout the life of the organization: for example, rates of executive turnover, average duration from product conception to delivery, percentage of product rejection in a production line – or more obvious measures such as profit and loss, sales volume or number of customers. If such a measurement relates directly to one of the anticipated outcomes of the coaching program being evaluated, then we are looking for a significant change in this measurement over time. Hopefully, this change will occur after the program has been in place for a given amount of time among those units of the organization that are participating in the program. With this design, a sufficient number of measures must be taken before and after the program is initiated in order to establish a comparative base. At least three measures should be taken before and two measures after program initiation.

The second quasi-experimental design, "nonequivalent control group design," is a bit more complex; however, it will in some cases help the evaluator partially overcome the Hawthorne effect among experimental group members and the sense of inferiority and "guinea pig" status among control group members. Rather than randomly selecting people into an experimental or control group, the evaluator can make use of two or more existing units (teams, departments or divisions). Two programs being offered by the HR Department, for instance, might be offered to several units in the organization. One or more of these programs would be those already provided by the HR Department, such as a management development program or online technology update seminars. The new coaching program would be the additional option. Client units would select one of the program offerings on the basis of time preference, convenience of location, specific need at the moment, etc. It is hoped that these reasons would function independently of the outcomes being studied in the evaluation. One of the units would be given the new coaching program (the experimental group), while the other unit(s) (the control group/s) receive the program(s) already provided by HR.

The clients may need to be informed of the differences between the experimental and control groups before signing up, based on an understandable concern for their welfare. If this is the case, then a subset of the clients from the experimental and control groups can be paired on the basis of specific characteristics (e.g., motivation, current performance level, or level of emotional intelligence) that might affect comparisons between the self-selected groups. The two subgroups that are paired thus become the focus of outcome determination evaluation, while the remaining participants in the two groups are excluded from this aspect of the overall program evaluation.

A "rotational/counterbalanced design" also can be used in place of a classic experimental design, especially if no control group can be obtained and if the evaluators are particularly interested in specific aspects or sequences of activities in the coaching program being evaluated. The rotational/counterbalanced design requires that the program be broken into three or four segments. One group of program participants would be presented with one sequence of these segments (e.g., segment 1, segment 3, segment 2), a second group of participants being presented with a second sequence (e.g., segment 3, segment 2, segment 1) and so forth. Ideally, each possible sequence of segments should be offered. Outcomes are assessed at the end of each segment.

In the case of a coaching program, this design is most appropriate if several different coaching strategies are being engaged in work with each client or client group (for example, life and career coaching, executive coaching, and team coaching). An evaluator who makes use of this design will obtain substantial information about program outcomes, as well as some indication about interaction between program activities. The rotational/counterbalanced design might be used successfully in the assessment of each coaching strategy. It would yield information not only about the overall success of the coaching program but also suggest which sequence of coaching strategies is most effective.

Campbell and Stanley describe a variety of other designs, indicating the strengths and weaknesses of each. They show that some designs are relatively more effective than others in certain circumstances, such as those involving limited resources and complex program outcomes. In addition, they suggest alternatives to the classic experimental design for situations in which that design may be obtrusive to the program being evaluated or otherwise not feasible.

Diagnosis

Program evaluations are often unsatisfactory, not because they fail to determine whether an outcome has been achieved or an impact observed, but rather because they tell us very little about why a particular outcome or impact occurred. At the end of a program we may be able to determine that it has been successful; however, if we do not know the reasons for this success (if we have not fully appreciated the complex dynamics operating within and upon this program) then we have little information that is of practical value. We have very few ideas about how to sustain or improve the program, or about how to implement a successful program somewhere else. All we can do is to continue doing what we already have done. This choice is fraught with problems, for conditions can change rapidly. Programs that were once successful may no longer be so.

Over the years, Michael Quinn Patton has been among the most influential evaluators in his emphasis on the pragmatic value inherent in a diagnostic focus. Coining the phrase "utilization-focused evaluation," Patton suggests that:

Unless one knows that a program is operating according to design, there may be little reason to expect it to produce the desired outcomes. . . . When outcomes are evaluated without knowledge of implementation, the results seldom provide a direction for action because the decision maker lacks information about what produced the observed outcomes (or lack of outcomes). Pure pre-post outcomes evaluation is the "black box" approach to evaluation.

A desire to know the causes of program success or failure may be of minimal importance if an evaluation is being performed only to determine success or failure or if there are no plans to continue or replicate the program in other settings. However, if the evaluation is to be conducted while the program is in progress, or if there are plans for repeating the program somewhere else, evaluation should include appreciative procedures for diagnosing the causes of success and failure.

What are the characteristics of a diagnostic evaluation that is appreciative in nature? First, this type of evaluation necessarily requires qualitative analysis. Whereas evaluation that focuses on outcomes or that is deficit-oriented usually requires some form of quantifiable measurement, diagnostic evaluation (particularly if appreciative) is more often qualitative or a mixture of qualitative and quantitative. Numbers in isolation rarely yield appreciative insights, nor do they tell us why something has or has not been successful. This does not mean that quantification is inappropriate to diagnostic evaluation. It only suggests that quantification is usually not sufficient. Second, the appreciative search for causes to such complex social issues as the success or failure of a coaching program requires a broad, systemic look at the program being evaluated in its social milieu. Program diagnosis must necessarily involve a description of the landscape and the program's social and historical context.

Third, an appreciative approach to diagnostic evaluation requires a process of progressive focusing. Successively more accurate analyses of causes and effects in the program are being engaged. Since a diagnostic evaluation is intended primarily for the internal use of the program's staff and advisors, it must be responsive to the specific questions these people have asked about the program. Typically, a chicken-and-egg dilemma is confronted: the questions to be asked often become clear only after some initial information is collected. Thus, a diagnostic evaluation is likely to be most effective if it is appreciative in focusing on a set of increasingly precise questions.

Malcolm Parlett, the developer of a diagnostically oriented procedure called "illuminative evaluation," describes appreciative focusing as a three-stage information collection process. During the first stage:

. . . the researcher is concerned to familiarize himself thoroughly with the day-to-day reality of the setting or settings he is studying. In this he is similar to social anthropologists or to natural historians. Like them he makes no attempt to manipulate, control or eliminate situational variables, but takes as given the complex scene he encounters. His chief task is to unravel it; isolate its significant features; delineate cycles of cause and effect; and comprehend relationships between beliefs and practices, and between organizational patterns and the responses of individuals.

The second stage involves the selection of specific aspects of the program for more sustained and intensive inquiry. The questioning process in the second stage of an illuminative evaluation becomes more focused and, in general, observations and inquiry become more directed, systematic and selective. During the third stage, general principles that underlie the organization and dynamics of the program are identified, described and, as a result, appreciated. Patterns of cause and effect are identified within the program, and individual findings are placed in a broader explanatory context.

The three stages of progressive focusing have been summarized by Parlett:

Obviously, the three stages overlap and functionally interrelate. The transition from stage to stage, as the investigation unfolds, occurs as problem areas become progressively clarified and re-defined. The course of the study cannot be charted in advance. Beginning with an extensive data base, the researchers systematically reduce the breadth of their inquiry to give more concentrated attention to the emerging issues. This progressive focusing permits unique and unpredicted phenomena to be given due weight. It reduces the problem of data overload and prevents the accumulation of a mass of unanalyzed material.

These three appreciative characteristics of diagnostic evaluation (qualitative analysis, systematic perspectives and progressive focusing) are often troublesome for both inexperienced and traditional evaluators. These characteristics appear to fly in the face of a contemporary emphasis on precision, measurement, objectivity and the discovery of deficits. Such is not the case, however, for these three characteristics can serve to enhance rather than take the place of a more traditional "scientific" evaluation.

In looking appreciatively at cause and effect relationships in a complex social setting and working with a complex set of activities (such as coaching), a whole variety of tools and concepts must be considered. In attempting to better understand the workings of a specific coaching program, the evaluator, like the cultural anthropologist, uses a variety of data collection methods, ranging from participant-observation and interviews to questionnaires and activity logs. Parlett suggests that the experienced evaluator also emulates the anthropologist in making use of various data analysis methods, ranging from narration and metaphor to multivariate statistics. This approach to evaluation was identified by Parlett several decades ago--yet it continues to challenge most approaches to evaluation and points to the "new" kinds of program evaluation that are needed if we truly intend to better understand how professional coaching operates when successful.

An Integrated Approach to the Evaluation of Coaching Programs

The four types of evaluation just described all contribute to the decision-making and dissemination processes that inevitably attend the ongoing planning and development of any coaching program. An optimally effective and appreciative program evaluation will draw all four types into a single, comprehensive design. This design brings together the valid, useful and appreciative information collected from documentary, descriptive, diagnostic and outcome determination evaluations. It combines this program information with a clear and consensus- based agreement concerning the purposes and desired outcomes of the coaching program, yielding results that translate readily into program decisions and dissemination.

Initially, the idea of incorporating all four types of evaluation into a single, comprehensive project may not seem feasible. However, with careful planning, all four types can be employed at relatively low cost. First, the same information sources and data gathering procedures can be used to collect several different kinds of information. Careful, integrated planning not only saves time for the evaluators, it also reduces the chance that program participants will feel "over-evaluated" and "under-appreciated." Second, the whole evaluation process can be spread out over a fairly long period of time, if planning begins early enough in the development of a new coaching program. The long-term planning of widely dispersed evaluative interventions makes a major evaluation project possible and reduces negative reactions to those interventions that are made. In general, an outcome determination evaluation will require extensive attention at the start and end of a program, whereas program description and diagnosis require greatest attention during the middle of the program. Program documentation requires some attention before the program begins and during the program. The most extensive documentation

work is required after the outcome determination evaluation is completed and before a final report is prepared.

Does it seem like too ambitious a plan? The initial response is usually "yes!" Do we really need to do all this work and initiate all of these different approaches to evaluation? The typical answer is "no!"

Unfortunately, these answers stop the process. Program evaluation is either not performed at all or done in a manner that yields little insight about successful coaching or provides few compelling reasons to support coaching services on a sustained basis for many clients. If professional coaching is truly of value to our organizations, then it deserves the kind of careful and appreciative approach being advocating in this essay. It requires an integrated approach that embraces both formative and summative purposes, and that interweaves the four Ds. We should evaluate that which we value . . . and this certainly should include the enterprise of professional coaching.