Theory E²: Working with Entrepreneurs in Closely-Held Enterprises: XIII. Assessment in the Enterprise Cycle (Part Two)

William Bergquist, Ph.D.

[Note: This essay and most of the previous essay in this series are based on a document I prepared in the early 1980s for the evaluation of a major higher education program that had received substantial funding from a large American foundation. These essays were later revised so that they might relate to a broader range of programs. I had intended to update these essays – but found that they held up very well over the many years. I decide to publish these essays as part of the Closely-Held Enterprise series as they were originally written. I hope you find them to be of relevance in your own work with entrepreneurs.]

There are four basic types of program evaluation: (1) description, (2) documentation, (3) determination of outcomes, and (4) diagnosis. 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, and illustrates the nature of the program and its impact. Following is a more detailed description of each of these four types of program evaluation.

Program Description

The first feature in any program evaluation, according to Michael Scriven (1991. p. 205), 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 an assessment. 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 (1991, pp. 121-122) suggests that a successful program description is something more than just the labeling of program components. We would propose that an appreciative approach to program assessment 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 (1977) takes description and appreciative evaluation a step further in suggesting that one of the most important purposes of an evaluation is the provision of sufficient 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 (Ost, 1977, pp. 1011):

> 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 student in this course? What did a typical employee do on a daily basis as a result of this personnel policy change? What was it like to walk into the office where this program was being administered? How has this program affected the life of a specific manager 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" program participant. In what activities did she engage? What did she miss? What didn't she like? Why? One might even want to create a hypothetical participant who represents "normal" involvement in the 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 employees can present the inquirer with evidence of program activity and accomplishment. Program assessments that do not include some documentation run the risk of appearing sterile or contrived. One reads descriptions of a program and one even reviews the 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 (Patton, 1990, pp. 115-117; Scriven, 1991, pp. 1980-182; Worthen et al, 1997, pp. 94-95). 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 a program. This may include minutes from major meetings, important memos and letters, reports, formal and informal communications about specific program activities or products, productions of the program, audio or video recordings of specific program activities, and so forth. These documents are usually stored in some file cabinet 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. 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 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.

The first of these two outcome determination questions is "researchable." We usually can determine whether a specific set of outcomes have been achieved. The second question requires an imposition of values. Hence, it is not "researchable." We cannot readily answer this question without substantial

clarification of organizational intentions—such as I present in Chapter Six of *Creating the Appreciative Organization* where I describe the process of Organizational Chartering (Bergquist, 2003). Yet, the issue of values and organizational intentions cannot be avoided in the determination of outcomes. *[In a later essay in this series I will examine ways in which the second question regarding the value of a program can be handled through use of a tool called Intentional Analysis.]* At this point we will explore ways in which the first question regarding achievement of pre-specified outcomes can be addressed.

Determining the Achievement of Pre-specified 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 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, the evaluation should show that one or more comparable group of customers, production lines or competitive organizations that were not exposed to the program did not achieved the pre-specified outcomes to the extent achieved by one or more groups that were exposed to the program.

In order to achieve this comparison between a group that has participated in a 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 there should be ideally at least two control groups, one that receives no program services and the other that receives an alternative to the program being evaluated.

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 customers in 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. In assessing a customer's or worker's attitudes, knowledge or skills before and after a program, one cannot always be certain that the two assessment procedures actually are comparable. Furthermore, if there is no significant change in pre- and post-program outcome measurements, one can never be certain that the program had no impact. The measuring instruments simply may be insensitive to changes that have occurred. On the other hand, the customers or workers 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 entrepreneurs may have to adopt alternative designs that are less elegant but more practical. In some cases, entrepreneurs have restricted their assessment to outcome measures. They determine the level of performance achieved by a group of outpatients in a mental health clinic 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 program units in other clinics. One doesn't know if differences in performance of students or recovery rates for mental health patience can be attributed to the program being reviewed or to the entering characteristics of the students or patients. Did the students at the alpha charter school do better than students at the beta charter school Was it because alpha students were already better educated or working at a higher level than beta students before they even entered the classroom?

This confounding effect is prevalent in many of the current initiatives that call for students to perform at a certain level on standardized tests without any consideration being given to their level of performance upon entering the school. In order to be fair in the assessment of a school's effectiveness, one must at the very least perform a "value-added" assessment (Astin, 1990; Bergquist, 1995). This type of assessment requires that a student's performance be measured when they first enter a school and again when they graduate from the school to determine the "value" that has been added, or more specifically the improvement in performance that has been recorded, with regard to their test scores

A similar case can be made for the assessment of a mental health clinic's performance. The patients at Clinic Gamma may be rated higher at a mental health status review than patients from Clinic Delta. However, we don't know if this can be attributed to differences in the severity of mental health problems being treated by the two clinics or to other extenuating circumstances. There may be differing socio-economic levels among the patients, differing levels of funding or staff support for the two

clinics, or differing criteria among those who rate mental health status at the two clinics. Simple outcome measures are rarely either accurate or fair. They certainly should be avoided in making decisions regarding program continuation, expansion, or modification.

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. Donald Campbell and Julian Stanley (1966) have described a set of "quasi-experimental" designs that allow one to modify some of the conditions of the classic 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 (see also Isaac, 1979). 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 attrition in a college, average length of stay in a hospital, percentage of product rejection in a production line. If such a measurement relates directly to one of the anticipated outcomes of the program being evaluated, there may be a significant change in this measurement. 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, the "nonequivalent control group design" will in some cases help the evaluator to 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 groups. Two therapeutic programs, for instance, that offer the same type of services might be identified. Clients would select one or the other program on the basis of time preference, convenience of location, etc. It is hoped that these reasons would function independently of the

outcomes being studied in the evaluation. One of the therapeutic programs would be given the new program, while the other (the control group) receives the services already provided by the agency.

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, socio-economic status, 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 program being evaluated. The rotational/counterbalanced design requires that the program be broken into three or four units. One group of program participants would be presented with one sequence of these units (e.g. Unit l, Unit 3, Unit 2), a second group of participants being presented with a second sequence (e.g. Unit 3, Unit 2, Unit 1) and so forth. Ideally, each possible sequence of units should be offered. Outcomes are assessed at the end of each unit.

An entrepreneur 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 a new set of training modules or a new public relations strategy. It would yield information not only about the overall success of the new set of modules or strategy but also suggest which sequence of modules or press releases 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.

Program Diagnosis

Many times, we find that program assessments are 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, that is if we have not fully appreciated the complex dynamics operating within and upon this program unit, 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 so rapidly that programs that were once successful may no longer be so.

Michael Quinn Patton (1990) is among the most influential evaluators in his emphasis on the pragmatic value inherent in a diagnostic focus. Coining the phrase "utilization-focused evaluation," Patton (1990, p. 105) 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 assessment that is appreciative in nature? First, this type of evaluation necessarily requires qualitative analysis. (Patton, 1990, Chapter 2 and 3) Whereas evaluation that focuses on outcomes or that is deficit-oriented usually requires some form of quantifiable measurement, diagnostic evaluation 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 human resource development 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. The program must be examined in its 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 will 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.

Appreciative focusing also takes place in a progressive manner during the information collection phase of the diagnostic evaluation process. As the developer of a diagnostically oriented procedure called "illuminative evaluation," Malcolm Parlett describes appreciative focusing as a three-stage information collection process. (Parlett and Deardon, 1977; see also description and critical analysis offerd by Worthen et all, 1977, pp. 158-159; and Scriven, 1991, p. 190).

During the first stage (Parlett and Dearden, 1977, p. 17):

> . . . 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 (Parlett and Dearden, 1977, p. 17):

> 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, a whole variety of tools and concepts must be considered. In attempting to better understand the workings of a specific program or culture, the evaluator, like the anthropologist, uses a variety of data collection methods, ranging from participant-observation and interviews to questionnaires and activity logs. A compendium of qualitative methodologies and disciplinary frameworks is offered by Michael Quinn Patton (1990), Chapter 3) 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.

———————

References

Astin, Alexander (1990) "Educational Assessment and Educational Equity," *American Journal of Education*, 1990, vol. 98, pp. 458-478.

Bergquist, William (1995) *Quality Through Access, Access with Quality*. San Francisco: Jossey-Bass, 1995

Bergquist, William (2003) *Creating the Appreciative Organization.* Sacramento, CA: Pacific Soundings Press, 2003.

Campbell, Donald and Julian C. Stanley (1966) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1966.

Isaac, Stephen. *Handbook in Research and Evaluation*. San Diego, CA: Edits, 1979

Kelly, Edward F. and others (1977) *A Story of Essence: The Portrayal of a Science Curriculum*. Syracuse, New York: Center for Instructional Development, Syracuse University, 1977

Ost, David and others (1977) *Interactive Evaluation of CAUSE, LOCI and ISEP.* A report prepared for the National Science Foundation, Directorate of Science Education, Office of Program Integration, December 28.

Parlett, Malcolm and Garry Dearden (1977) *Introduction to Illuminative Evaluation*. Washington D.C.: Council of Independent Colleges.

Patton, Michael Quinn (1990) *Qualitative Evaluation and Research Methods*. Second Edition. Sage Publications: Newbury Park, CA.

Scriven, Michael (1991). *Evaluation Thesaurus*. Fourth Edition. Newbury Park, CA: Sage Publications, 1991

Worthen, Blaine, James Sanders and Jody Fitzpatrick (1997) *Program Evaluation: Alternative Approaches and Practical Guidelines*. Second Edition. New York: Addison Wesley Longman, 1997