DOCUMENT RESUME

ED 097 489 95 CE 002 257

AUTHOR Trice, Harrison M.; Roman, Paul M.

TITLE Evaluation of Training: Strategy, Tactics and Problems. Training Information Sources, No. 3.

INSTITUTION American Society for Training and Development, Madison, Wis.; Syracuse Univ., N.Y. ERIC Clearinghouse on Adult Education.

SPONS AGENCY Office of Education (DHEW), Washington, D.C.

PUB DATE Aug 73

NOTE 49p.

AVAILABLE FROM American Society for Training and Development (ASTD), P. O. Box 5307, Madison, Wisconsin 53705

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE

DESCRIPTORS Educational Assessment; *Educational Strategies; *Evaluation Criteria; *Evaluation Methods; Literature Reviews; Manpower Development; Organizational Development; Performance Criteria; *Personnel Evaluation; Trainers; *Training; Training Objectives

ABSTRACT

The purpose of the literature review is to outline the basic considerations involved in designing and conducting meaningful evaluation of training efforts in work organizations. First to be examined are basic questions in training evaluation strategy: the barriers to conducting evaluation of training efforts, the reasons for conducting such evaluation, where and when such evaluation should enter the training process, and who should conduct such evaluation. Following this is a consideration of six approaches whereby evaluative questions may be posed, including an outline of the assumptions of each along with their advantages and disadvantages. The tactics of evaluation of training efforts and the bases for choosing different tactical approaches are two major concerns dealt with. Concluding the document are a discussion of recent trends in training evaluation and a seven-page bibliography. (Author/MW)

ASTD

Fund for
Research in
Training and Development

TRAINING INFORMATION SOURCES

NO. 3 JUNE, 1974

EVALUATION OF TRAINING: STRATEGY, TACTICS AND PROBLEMS

COMPILED BY ERIC CLEARINGHOUSE ON ADULT EDUCATION

ERIC CLEARINGHOUSE ON ADULT EDUCATION

SPONSORED BY SYRACUSE UNIVERSITY AND THE EDUCATIONAL RESOURCES INFORMATION CENTER OF THE U.S. OFFICE OF EDUCATION

Available from the American Society for Training and Development (ASTD), P.O. Box 5307, Madison, WI 53705.

This publication was prepared pursuant to a contract with the Office of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, necessarily represent official Office of Education position of policy.

EVALUATION OF TRAINING:

STRATEGY, TACTICS AND PROBLEMS

HARRISON M. TRICE

Cornell University

and

PAUL M. ROMAN

Tulane University

CONTENTS

Page

FOREWORD

Four years ago, ASTD and the ERIC Clearinghouse on Adult Education undertook a cooperative venture to produce a series of three publications for trainers. The first two were bibliographies, *Management and Supervisory Behavior*, and *Occupational Training for Disadvantaged Adults.*

The present volume, a literature review on evaluation of training, completes this series.

Evaluation of training continues to be a big topic among trainers in business and in industry, just as it is with other educators. For this reason, we asked Professors Harrison M. Trice and Paul M. Roman to prepare a literature review on the strategy, tactics and problems in the evaluation of training.

The authors have diligently and assiduously searched the current literature in order to present an up-to-date synthesis of research findings and exemplary practices. They have approached their task with the trainer in mind; consequently this monograph is of practical use both to the new trainer as well as the one with experience.

Since the completion of this manuscript, two other publications on evaluation commissioned by the ERIC Clearinghouse on Adult Education have appeared, both of which will be of interest and usefulness to the trainer in business and industry.

One of these is *Contemporary Approaches to Program Evaluation*, by Sara M. Steele (available from Education Resources Division, Capitol Publishers, Inc., Suite C-12, 2430 Pennsylvania Avenue, N.W., Washington, D.C. 20037, 272 pages, $9.00).

The other is *Evaluating the Attainment of Objectives: Process, Problems and Prospects*, by Sara M. Steele and Robert E. Brack (available from Syracuse University Publications in Continuing Education, 224 Huntington Hall, 150 Marshall Street, Syracuse, N.Y. 13210).

The ERIC Clearinghouse on Adult Education is grateful to Professors Trice and Roman for their scholarship and efforts in preparing this monograph. Also, we are grateful to ASTD for their cooperation in this joint venture.

August 1973

Stanley M. Grabowski, Ph.D.
Director
ERIC Clearinghouse on Adult Education
Syracuse University

INTRODUCTION

Training based in work organizations attempts to *change* the attitudes or behavior of personnel in some desired direction (Roy and Bolke, 1971). Training may be designed to alter undesirable attitudes, increase specific types of knowledge, modify current behavior, or create tendencies toward certain behavior choices in the future. Training evaluation attempts in various ways to discover if desired changes occur, particularly when the trainees return to their jobs.

While training evaluation appears simple and fundamental, it offers many challenges, particularly in the work-world setting. The purpose of this monograph is to outline the basic considerations involved in designing and conducting meaningful evaluation of training efforts in work organizations. We shall first examine basic questions in training evaluation *strategy:* the barriers to conducting evaluation of training efforts, the reasons for conducting such evaluation, where and when such evaluation should enter the training process, and who should conduct such evaluation. Following is a consideration of six approaches whereby evaluative questions may be posed, including an outline of the assumptions of each along with their advantages and disadvantages. We then turn to our major concern, the *tactics* of evaluation of training efforts, and the bases for choosing different tactical approaches.

Barriers to Training Evaluation

Despite many good intentions, effective evaluation of training appears to be infrequent. A reason for this, often unknown to training specialists, is that good training evaluation may demand as much planning and resource input as the actual training effort. Another reason for giving evaluation shortshrift is the difficulty inherent in determining the desired training outcomes (Bond, 1970; Oliver, 1971). Bass and Vaughn (1966) describe this problem succinctly: "Many of the outcomes of training are difficult to anticipate or predict, difficult to measure accurately, and difficult to relate to the training objectives or broader organizational goals."

A basic resistance to the evaluation of training may lie in the threat of embarrassing results. Evaluation data which indicate a lower-than-desired degree of effectiveness may reflect badly on the trainers and planners, possibly threatening the resource allocation they receive or their overall prestige in the organization. Those responsible for the training may fear that evaluative results will not be understood by top management, and that such results will be used "against" them. A dramatic illustration of such potential embarrassment is Bremer's (1968) careful evaluation of a management development program which revealed changes in a direction opposite to that desired by top management.

The contrasts between the roles of trainers and evaluators can produce resistances and misunderstandings. Evaluators often want standardization of training procedures and assurance that their evaluative strategy can proceed without interruption. Many trainers believe that evaluators are a necessary evil, but some view them as "snoops" and, at worst, evaluation is viewed as a threat to the future of the training program, and the training which trainers feel is needed by employees. The data needed for evaluation must be collected within a practical work situation that often does not fully fit the systematic approaches of evaluators. On the other hand, trainers must recognize the basic necessities of evaluative procedures. The upshot is that training evaluators are often forced to compromise with training procedures, adjusting their strategies to work-world realities. The necessity of accommodation must be recognized in light of the dependence of training and evaluation on one another (Mesics, 1969).

Why Evaluate Training?

Although improving the effectiveness of a specific training effort is often the chief reason for evaluation, there are other reasons. Numerous signs point to the 1970's as the "decade of accountability" (Nadler, 1971:2). Work organizations which support training efforts are becoming increasingly concerned with training effectiveness. Like other corporate investments where returns are expected, management looks for returns from investments in training.

Despite recent austerity in many quarters, the past decade has seen an unparalleled growth, innovation, and diversity in training programs, strategies, and techniques. In the face of this diversity and as more is invested in training, management now asks which specific approach is likely to be most effective for their specific situation. Training specialists likewise are atune to the question of the effectiveness of different procedures. Thus the plethora of available techniques has heightened the relevance of evaluative questions.

Another reason for evaluation is its potential "side-effects" in pointing up training needs and organizational problems. Evaluation may reveal, for example, felt needs for better communication skills or other fundamentals that are not the primary goals of a particular training effort (Belasco and Trice, 1969). Problems in performance appraisal or wage and salary administration may likewise be revealed as a side-effect of a specific evaluative effort.

A further argument for systematic evaluation lies in the simple observation that evaluation is a fundamental, everyday aspect of organizational life. Involvement in any change-oriented effort leads to conclusions about its effectiveness. While such common-sense evaluation will occur "anyway," it is usually unsystematic, haphazard and impressionistic. Training specialists will be on much safer ground with systematic information, regardless of whether such data supports or refutes opinions generated via informal appraisals.

Finally, sheer survival of a training department and its programs may be a major reason for hasty, last-ditch efforts to collect evaluation data to justify training expenditures. Evaluation can become central in a fight for the existence of certain training efforts. Under less critical circumstances evaluation data have been used to improve a program's financial support. If a sound monetary return from training can be demonstrated, evaluation data may actually enhance the training department's position in the organization.

Where Should Evaluation Start?

Sound training evaluation begins with description before moving to assessment. It should start with an exhaustive, detailed description of the training "action," or the orientation, content, and context of the particular training effort. In short, it is impossible to evaluate a change effort unless that effort can first be described in its specifics. Such a description should go beyond the specific training content, including as much information as possible about the history of the training, backgrounds of the personnel performing the training, the organizational setting within which it takes place, along with as much information about the trainees as is practically possible to obtain. Training specialists often believe such descriptions are simple and readily available. Substantial evidence contradicts this. Like change agents of all types, training specialists are usually so "close" to their training "action" that they have difficulty in thoroughly describing it. Much evaluation misses the mark simply because a detailed description of the training did not precede the attempt at assessment, producing erroneous notions about objectives which then led to inappropriate criteria for gauging effectiveness. The absence of descriptions of the action can lead to assessments that may portray the training effort in an unfavorable light.

The need to get as full a description of the change intervention as possible is especially acute in efforts to evaluate Organizational Development programs. They included such a wide-range of intervention efforts that the importance of specifying precisely the nature of the interventions cannot be overemphasized.

By beginning with this kind of description, evaluators can get a picture of how much "muscle" the effort has in terms of resources: trainees' time, personnel, and equipment. When evaluative results are interpreted, such data can be invaluable in bringing expectations into line with investments. For example, supervisory training involving one hour per week for eight weeks justifies lower expectations than a program lasting five full working days. Another example: training done by the organization's selection and placement specialist who has little training experience should generate lower expectations than a program involving an outside selection and placement specialist with long training experience. Obviously, expectations are contingent upon the entire set of factors comprising the "action."

Turning to trainees, the description of the "action" should contain an analysis of the selection factors operating to bring trainees to the training. Do the training people work under the frequent handicap of "We have to take them all"? Clearly, careful selection of any type of trainees may markedly influence outcomes in a favorable direction in contrast to situations where the trainees must accept all who come, by whatever means.

A description of the "action" should also focus on goal consensus. Do trainers and line managers agree on the practical objectives of the training? Do trainees tend to accept these? How much disagreement is there among the trainers themselves? If there is sizeable disagreement in any quarter, interpretation of evaluative results take this into account.

The rapidly expanding varieties of training devices and strategies underscores the need for evaluators to know "what the action is" so that evaluation can be fitted to it. Some of these strategies vary sharply from traditional training methods, and understanding of them may require more than cursory attention. For example, attempts to improve interpersonal skills via T-group and laboratory types of training have proliferated (Argyris, 1963; Bunker, 1965), generating intense controversy (Odiorne, 1963; Coghill, 1967). Business games and simulation techniques (e.g. supervisory conflict situations) are competing for popularity with their proponents claiming to turn written cases into "living cases" (Coleman, *et al.*, 1966) by comparing trainees' actual decision-making, in a simulated work situation with an ideal model representing a performance assumed to be effective (Wasmuth, 1970; Lamp, 1970). Use of television and videotape has burgeoned (Adams, 1971; Patten, 1971) along with other more esoteric devices such as telelectures (Glueck, 1971) and candid camera shots that catch executives dealing with others under pressure in real situations (Marcus, 1971).

"Management by objectives" continues to be a popular technique (Gill and Molander, 1970) and the more traditional conference leadership pattern, including written "cases" which set the stage for role-playing, remains very much a part of the training scene. Intermingled with these have been a plethora of innovations: peer instruction by past trainees or current employees (Smith, 1971; Westley, 1971; Seingarten, 1971), "high intensity training" (Davis, 1968; Nadeau, 1969), "action-centered leadership" training (Wilie, 1971), training by trivial tasks (Oates, 1971), Group Feedback Analysis (Heller, 1970), and listening training (Ross, 1969). Evaluation of programs including innovative techniques may call for reading, consultation with others who have used such techniques, and even observation of these techniques in action in settings outside one's own organization.

When Should Evaluation Start?

Within the training setting, evaluation is more effective when it is a normal and accepted part of program planning from the beginning. Along with identification of training needs, selection of training methods, and logistical planning, evaluation should be an initial consideration. An unfortunate but common experience for an evaluator is to be approached with a request to evaluate an effort near the end of a training sequence, or after it is over. Although an evaluative attempt can be made at this point, it is sharply handicapped and usually fraught with difficulties.

Some training specialists conduct evaluation "in mid-stream," i.e. before training action is over, with the hope of correcting deficiencies during training and before opportunities are lost. Such "midstream evaluation" usually leads to altering the training "action," in one way or another, thus contaminating later evaluation. Subsequent data on training effectiveness will not answer basic evaluative questions since it will be impossible to ascertain whether the original procedures or the "midstream" innovations produced the revealed results.

Collection of evaluative data seems to typically occur immediately after training efforts conclude and before trainees fully "reenter" their regular work roles. Since reentry may either reinforce or weaken changes produced by the training, evaluation at this point can be misleading. Evaluators selecting this time should qualify their results to the effect that reentry impact is unknown. By the same token, evaluative measurement taken only after reentry does not allow distinctions between training effects and reentry effects.

Other evaluators have collected information just before reentry and soon thereafter. This obviously throws more light on effectiveness than the single measures mentioned above. At the same time, however, it introduces the problem of the

"Hawthorne effect." e.g. people alter their behavior in response to the researcher giving them "attention" via data collection instruments. The processes involved in two collections of data so close together may produce more change than the training itself. It is helpful to recall here the almost universal necessity to compromise in evaluative studies. When data are collected in this before-and-after-reentry manner, one should point out the strong possibility that research effects may be partly responsible for results.

Other evaluators focus on the extent to which changes have been sustained at some reasonable time (six to eight months) following conclusion of training. In this instance it is difficult to account for observed change in terms of the training, reentry, or organizational changes that may have occurred between the end of the training and the time of data collection. While such data may fail to answer basic evaluative questions, the revelation of sustained changes over a relatively long period of time may prove persuasive to others seeking to understand the value of the training effort. Such "good" findings are not always forthcoming, however.

It should be obvious that any "one-shot" and most "two-shot" collections of evaluative data are fraught with problems. Fleishman *et al.* (1955) offer evidence that the time of an evaluation will influence recorded results. Their measures, taken immediately after the conclusion of a training experience, indicated a positive change in the desired direction. The same measures administered at a later date indicated that much of the change had been eliminated; some supervisors had even become negative about the subject matter of training.

Warr *et al.* (1968) urge that evaluation should take place at several stages of training, not just at the end of a course. They argue for the collection of information at four points: (1) trainees' initial reaction to training; (2) immediate consequences (changes in trainees' knowledge, skills, or attitudes which can be identified immediately after completion); (3) intermediate consequences (changes in trainees' actual behavior which result from training); and (4) long term consequences (changes in the functioning of the organization due to changes in work behavior). Kirkpatrick (1956) suggests that increased knowledge can be an immediate objective, changes in job behavior an intermediate one, and changes in production, turnover, absenteeism, and morale an ultimate objective.

Thus the time at which evaluation materials are to be collected to indicate progress toward any objective is a vital concern. Another dimension of timing evaluation concerns the stage of development of the unit mounting the training. During the formative period when objectives and goals are in flux, enthusiasm will probably be high and result in greater effort. Baseline evaluation data collected at this point in time can be markedly affected by the fact that goals and objectives may sharply change by the time the training is finished and the eagerness of the trainers can likewise diminish. In such circumstances it may be appropriate to employ two types of evaluation suggested by the American Institute for Research (1970): "formative" evaluation, which serves primarily to guide improvement in a new and evolving program; and "summative" evaluation which seeks to determine whethei the goals were achieved after the program has crystalized.

Who Evaluates Training?

Some strongly believe that only outside specialists evaluate training while others, with equal emotion, contend that someone inside the organization should do it. Still others argue that the trainees (such as supervisors) should conduct the major part of an evaluation study (Scott, 1971).

Outsiders are presumably more objective; deliberate choice of an outsider is also more likely to produce someone more experienced in evaluative studies. Outsiders are not involved in internal power cliques and struggles, and usually have no vested interest in the outcome of the evaluation. But as outsiders they often are seen as aliens who have uncontrolled power to impose evaluation strategy and who do not "really understand" the goals, meaning or content of the training effort. Negative reactions may be reduced by the outsiders' "professional" image if they are academics or members of an established management consulting firm. In any event, while such specialists probably have more expertise, they often find it difficult to reduce the suspicious concerns of training practitioners.

Evaluators from inside the organization may have almost as many detractions. Training specialists are typically action-oriented and know little about evaluative strategy. They usually are deeply involved in the training and consequently tend to regard it as effective. Trainers themselves are in a difficult position if they must conduct objective studies with the "risk" of negative results (Weiss, 1970). Unless company policy strongly underscores the *training improvement function* of evaluative studies, inside evaluators may be viewed as threats and placed in an awkward and unhappy position indeed.

Some large training departments have a specialized "training evaluator" position, but this is impossible for most.. Even if an organization can afford such a position the role will still suffer from the subjectivity and potential threats created by "insider" status.

The best way out of this dilemma is to find an academic evaluator or an experienced evaluator in a management consulting firm whose mode of operation harmonizes with the evaluative ideas of training practitioners. The key elements leading to his acceptance are professionalism and objectivity. These characteristics may be maximized in the case of an academic-based evaluator since his image probably does not carry any connotation of the "right way" to conduct training. The evaluator from the management consulting firm, on the other hand, may represent a particular type of "training package" which his organization is attempting to market. Furthermore, the management consultant may be attempting to build up rapport with an organization for future contracts such that he may guide the evaluation in a way that pleases those in significant power positions. While we fully recognize the professionalism of most management consultants, the salient point is that most of them fully depend on this work for their livelihood and advancement, whereas the academic is primarily committed to and rewarded by his university affiliation.

The approach of either type of professional evaluator calls for them to help training practitioners improve those often crude and flawed evaluative methods which are already in use or being considered. Rather than immediately implementing evaluative approaches that to him seem appropriate for a training program, he starts by carefully studying the existing ideas and actions which trainers have regarding evaluation. From his body of evaluative knowledge he can then assist practitioners in refining and improving their own strategies and tactics.

HOW CAN TRAINING EVALUATION BE APPROACHED?

There are several different ways of approaching evaluation. These "approaches" set forth the various ways in which the viability of a training program can be examined; they are not specific evaluative strategies but instead are "points of entry" into basic evaluative questions. In other words, a given training effort has numerous dimensions, and evaluation designs vary in the degree to which they take these various dimensions into account. Most discussions of evaluation do not consider the assumptions underlying particular strategies, but we believe that some focus on them will show that training evaluation is somewhat more than the simple question, "Did it do any good?"

One experience indicates six approaches to evaluative questions: the goal-attainment approach, the cost-benefit approach, the ceremonial approach, the organizational support approach, the outside validation approach and the native methods approach. In most instances it is desirable to employ a combination of several of these approaches, and it is possible for an evaluation study to incorporate all six approaches in the consideration of a single program. We shall now examine the assumptions of each of these approaches and examples of their use, and then consider specific evaluative strategies and tactics appropriate for their implementation.

The Goal Attainment Approach

Within this approach it is assumed that the objectives of training are clearly defined and that both line management and staff trainees support them. It is assumed that the training effort is directed singly toward these goals and that success on the extent to which these are reached can be assessed.

There is no doubt about the goal-attainment approach being the most traditional, the most widespread, the most obvious, and seemingly the most practical way to assess training. For example, program evaluation in adult education continues to emphasize the 1952 statement that evaluation is determining the extent to which objectives have been obtained (Adult Education Association, 1952). It is safe to assume that the majority of training specialists have never thought of anything other than the goal attainment approach.

The model assumes that trainers have goals that are sufficiently clear so the evaluator knows where to focus his measurements, such as the goals of increased knowledge, specific behavior change, or increased performance effectiveness in the "back home" work situation. It is assumed that the goals are sufficiently specific that their achievement will produce measurable changes in the trainees. Obviously the degree of goal attainment cannot be determined unless some conveniently measurable criteria can be derived from the goals. This means some form of reliable and valid criteria that, like a ruler, have equal intervals of measurement to gauge the degree to which the goal has been reached. One of the valuable spin-offs of this approach is to urge training people to be clear, specific and uniform about the changes toward which they are aiming.

The goal attainment model is very prominent in training evaluation practice. Ferrari (1970) argues that: the objectives and goals of management training programs *must* be explicitly known before the programs can be evaluated. The first goals of a program for updating continuing professional training for physicians was the development of objectives, both immediate and long range (Dean, 1969). Forty-five percent of the 110 respondents in Catalanello and Kirkpatrick's (1968) survey of how companies assessed training indicated that they attempted to measure goal results, not trainee reaction, learning, or post-training behavior. Higher wages were both the goal and the yardstick for evaluating four manpower training programs and control groups in sixteen firms that hired program graduates (Greenberg, 1968).

But this deceptively simple model has disadvantages. Unless carefully planned in advance, training objectives often change as the program unfolds. If an evaluative strategy has been attached to the initial goals, it becomes irrelevant as the goals change. Even if goals are stated and are not changed during the training process, they are often fuzzy and overly idealistic, e.g. "improve human relations," "improve the quality of work life." This should not be surprising, however. Official goals are "purposely vague and general" (Perrow, 1961:855); it may be more meaningful to examine the "operative goals" that represent the give-and-take of informal influences and which come closer to the actual goals of the effort, these too can be difficult to define. Even if goals are "nailed down," they may be numerous with the consequence that the evaluator must select from among them and leave a number unassessed. Finally, there may be considerable conflict over goals among trainers and those involved in the training. This will be exascerbated if there are numerous goals, which in turn increases the probability of conflict among goals, i.e. the attainment of one goal may require the abandonment of another goal.

A classic example of such conflict is between those seeking long-term improvement in management-worker relations and those desiring short-term, rapid improvements in production. Each may argue that the achievement of his goal will automatically lead to achievement of the other's goal, e.g. happier people will be more productive vs. productive people will be happier. Such potential combat over priorities in the goal attainment approach can prevent the evaluation effort from getting "off the ground."

Evaluators using the goal attainment approach must spend an inordinate amount of time and effort ascertaining and making explicit the actual operating objectives toward which training resources are aimed. In its ideal form, however, the approach permits the identification of intermediate targets, allowing for fast feedback to effect reprogramming. A further advantage is that it is readily understood by training people who can participate intensively in its implementation. If goals can be ascertained, specified and agreed upon, the goal attainment approach may be relatively simple, indicating that in many instances it is probably less costly than other approaches, requiring fewer additional high-priced specialists or consultants.

The Cost-Benefit Approach

In this approach, which is an intense refinement of the goal-attainment approach, the evaluator tries to estimate the benefits of a training program and the direct and indirect costs of the program, which in this instance are invariably calculated in dollars and cents. The ratio of benefits to costs shows the return that the organization receives from its investment.

Historically, this approach has been related to decision making about alternative training techniques via comparison of their alternative cost-benefit ratios. Because of this highly explicit criterion, those employing the cost-benefit approach must deal intensely with the basic problem of the goal attainment approach: What are appropriate and desirable outcomes, which in this instance will be defined as "benefits"? Thus the approach calls for a very explicit consensus on the goals of the training, assuming of course that such consensus is essential if the evaluative results are to be meaningful to the various personnel who have involvements in the training. Because it is based upon quantification, the approach requires "benefits" that not only can be tabulated but can be expressed in monetary terms. Thus while increased knowledge, improved human relations skills, or positive changes in attitudes may be tabulated from questionnaire or interview data, it is difficult if not impossible to translate these values into dollars. Thus the "benefits" must invariably be defined as some form of output of production to which specific value can be assigned.

When the cost-benefit approach is employed in training evaluation, one of two basic strategies are typically followed. One formula calls for the benefit cost ratio to be calculated by subtracting current and capital costs in each year from current benefits which are "discounted," i.e. the discount rate reduces benefits by some common percentage to make up for future benefits that were foregone by current use of time and effort. Obviously, different discount rates can sharply influence cost-benefit results and the subsequent policy decisions. Thus estimated discounted benefits minus estimated costs equals the cost-benefit index.

The second method is to construct a ratio by dividing the gross benefits of the training for a typical year by the current and capital costs. Some state the formula as dividing the discounted present and future benefits by the estimated value of "the resources that would have been available for other uses" had training not taken place (Conley, 1969). It is more common, however, to divide gross benefits calculated for a typical year by current and capital costs to produc a benefit-cost ratio of gross annual benefits to total annual costs.

Once the bases for calculating benefits and costs have been established. the next step in this approach is the listing of the sources of benefits and costs, attaching specific dollar values to each. In company-based training, benefits may be estimated in terms of increased productivity, using some appropriate yardstick of improvement of performance. Efforts are then made to establish the sources of increased productivity. If training is being evaluated, it will be essential to have comparison groups of non-trainees so that the effects of training may be separated from the possible effects of other factors. We elaborate further on experimental design in a later section.

Evaluations of manpower training programs tend to use the securing of employment or increased earnings as ways to estimate benefits. In training programs. cost estimates tend to center around salaries of training personnel, time of trainees lost to the company, and costs of training materials. To these costs are often added charges for periodical contributions to a "sinking fund" adequate to discharge debts or make replacements of major items. A large array of specific benefit and cost inputs can go into any given formula (cf. Oatey, 1970).

The evaluations of government-financed manpower programs offer illustrations of the cost-benefit approach. In estimating training and costs for Navy enlisted personnel, such items as travel to school, initial uniform issue, pay and allowances, and accrued leave for students and staff were considered (Clary, 1970). A typical example of cost benefit evaluation of manpower training programs is a study of 501 West Virginia Area Redevelopment and Area Vocational Training Program Trainees and a comparison group of 453 randomly chosen nontrainees (Somers, 1964). By the summer of 1962, 60 percent of trainees were employed versus 56 percent of the dropouts and about 30 percent of the others. A year later, employment rates were up for all groups, with only slight differences between the trainees and the other groups. One of the key indexes in Somers' study was "pay-back periods," defined as time required for differential earning gains to offset total retraining costs. In another evaluation of a government-financed manpower development program, Somers (1965) emphasized the criteria of gains in worker productivity, employment, earnings, and gains to society in reduced amounts of welfare payments, and cost benefits of governmental programs. It was concluded that the economic gains derived from retraining by the trainees, and perhaps by society, seem to outweigh economic costs. The cost benefit approach suffers from the obvious deficiency that many benefits are intangible and defy reduction to a dollar and cents value. However, where benefits and costs can be legitimately put in these terms (as in measurable outputs in vestibule training), it is an impressive tool. Unfortunately, the data for many dollar estimates of benefits is most speculative. At the same time, the approach is quite attractive since it attempts to put evaluation into a frame of reference easily understood by management and, in the case of government support, it "makes sense" to fiscally-oriented legislators.

Another disadvantage is that it is a very complex process that can become so intricate that often only the end results can be comprehended. Probably the greatest risk in using this approach lies in its complexity: possible refinements can be manipulated in such a way that a specific desirable outcome can be secured, leaving the non-specialist unable to detect the defects in the complex accounting. While one need not expect dishonesty, this approach offers many opportunities. Its exclusive emphasis on the "final" calculations can raise anxieties more sharply than other approaches, giving a sense of an "all-or-nothing" outcome.

While cost-benefit analysis is attractive if results can be used to persuade managers oriented to "hard facts," its exclusive use can detract from the basic goal of training evaluation: improving training. It may encourage one-shot use of several particular training strategies, with evaluation solely oriented to the best cost-benefit ratio. But more importantly, its procedures are not oriented to the basic issue of *why* a particular training effort succeeded or failed.

Finally, it is obvious that a single cost-benefit ratio is meaningless without some base or alternative for comparison. This in turn generates the ever-present issue in evaluation of what is a "good" outcome as compared to a "moderate" or mediocre outcome. Given the complexities involved in cost-benefit determination, the possibility of illegitimate comparisons is great when the choice between different training strategies is at issue. The growth of emphasis on "data" and the widespread availability of computer analysis in work organizations augur for increased attraction to the cost-benefit approach.

The Ceremonial Approach

Another modification of the goal-attainment approach, sharply different from cost-benefit analysis, is to focus upon certain "side-effects" of the training process which may be viewed as the "ceremonial" impacts. To some extent, ceremonial effects are similar to Hawthorne effects in that they are usually unanticipated results that stem from social interactions directed at accomplishing something else. Speaking of Organizational Development programs, Margulies is more dramatic about the approach: "It seems that a powerful and neglected force in the effectiveness of organizational development has been the magical quality which is derived from the client-consultant relationship, and from the psychological belief in O.D. values and processes" (Margulies, 1972:182).

The ceremonial approach attempts to evaluate training by estimating the extent to which the training (1) legitimizes the trainees in the eyes of others, (2) provides "rites of passage" for trainees in transition from one status in the organization to another, (3) acts to maintain organizational stability following technological or cultural change, or (4) changes trainees' self-perceptions in the organization (Trice, *et al.*, 1969).

The approach assumes that many organizational members experience anxiety about upgrading and continuity of staff. It posits tha' training serves to reduce these anxieties by legitimizing those trained. It transmits cues to others in the organization that those who are trained are competent and eligible for tasks and responsibilities which they undertake after the training. It is assumed that the training serves to convince organizational members that selection and other decisions about trainees were appropriate and useful even though such selection might have been equally effective if it had simply involved random choice.

To illustrate, an organization may select a group of individuals for promotion to supervisor. Those in the group participate in a training effort before assuming their new positions. The ceremonial approach posits that the simple fact that the new supervisors have undergone training may legitimate their assumption of supervisory roles, *independent* of the degree to which they actually acquired the requisite skills for supervision or were "changed" by the training in other ways. The approach posits that training may constitute a *labeling process* in the eyes of the non-trainees whereby the trainees are believed to possess the competency required for their new tasks.

Secondly, the approach attempts to assess how effective the training may have been as a rite of passage, changing the organizational role perceptions of trainees for both the trainees themselves and for those who interact with them. Thus training may fulfill another labeling process, verifying the fact that the trainees have actually moved to a new status in the organization. There is a growing belief that the peers, supervisors, and subordinates of the trainee are legitimate sources of evaluative data (Bolor, 1970). Such a strategy makes this part of the ceremonial approach operative. The ceremonial approach focuses upon the extent to which the training removes the member in transition from interaction with those performing organizational roles he will abandon. It appraises how often and how intensively the training brings the trainee into close and frequent contact with persons occupying positions related to, and perhaps identical with, those he will enter. Finally, the approach assesses the extent to which the "changed" person is introduced into a new system of on-the-job associations, and new role expectations and behaviors, either through associations with others in the training process or in the reentry period. Thus it is possible that training may have more self-concept impact on nonparticipant than on trainees (Levinson, 1966).

The ceremonial approach may be also employed to examine training as a stabilizing force during organizational change. For instance, management training programs may clarify new role expectations for both individual role performers and other related system members, while facilitating the emergence of new informal groupings through which expectations can be fulfilled. Similarly, as a result of providing the trainee with an opportunity to share his problems with others, training may reduce feelings of isolation, frustration, and anxiety about new tasks. This evaluative approach may also focus upon the extent to which training indicates the organization cares about the problems of people, producing in the trainees a feeling of organizational inclusion and identity (Belasco and Trice, 1969).

This leads to a final possible focus of the ceremonial approach: changes in identity among the trainees themselves. Training efforts vary in their intensity, the degree to which they emphasize complex learning, and the degree to which they provide tangible signs of participation, such as graduation certificates. In any event, it is certain that the mere fact of being chosen for participation in a particular training effort can transform the trainees' image of himself. Behavioral science research clearly indicates that the degree to which such transformations are initiated and sustained are contingent upon the images of the person held by those around him.

It seems reasonable to believe that such a "payoff" might come from the formal managerial training courses recommended for Black professional talent. One study revealed significant differences between the career development of black and white salaried employees, concluding that the development of Blacks for professional employment calls for special training efforts, especially a variety of formal managerial courses (Crane, 1970). Emphasis on the ceremonial dimensions discussed above would appear to be important considerations in this type of training. Zeller reports a positive impact on labor union members of training provided to union leaders about how to develop support for anti-poverty programs and generate self-help activities. This suggests a "spin-off" onto non-trainees of the type envisioned by the ceremonial approach.

This approach highlights the fact that there may be many desirable unanticipated consequences of training that are not detected by evaluators. It could easily be that no results are achieved on the formal goals, yet the indication of ceremonial results could justify the entire effort, i.e. if persons think of themselves as supervisors and are defined by others as such, they may act effectively as supervisors. In short, "ceremonial payoffs" should always be looked for regardless of what other approaches might be used.

One of the approach's disadvantages is that training specialists tend to view its explicit formulation as reducing their efforts to insignificant ritual. Americans like to think of themselves as pragmatic, passing over "ceremony" for concrete results. Thus trainers are apt to overlook the possibilities in the model. Much more difficult, however, is operationalizing the approach. How can anxieties about the quality of staff be gauged and how can reductions in it be attributed to training? To our knowledge such measures have never been attempts. Even more elusive are methods for demonstration that training acts to stabilize organizational life during change. On the other hand, collecting information from non-trainees regarding their impressions of trainees or demonstrating increased feelings of organizational commitment among the trainees as a result of training are not too difficult from a measurement point of view. Similarly, measuring shifts in role identity, such as from operative to foreman, has not been easy to carry out, but, again, is by no means impossible.

The Organizational Support Approach

Numerous dissatisfactions that have accumulated regarding the goal attainment approach (as well as the cost-benefit approach which is a specific type of goal attainment model) have recently led to the advocacy of the organizational support approach, sometimes referred to as the "systems model."

Weiss and Rein (1970) believe that the goal attainment model leads to attempts to use experiments that practically always fail. The approach they propose conceives of all parts within a collective effort as being reciprocal or interdependent upon one another. These systems can be either "closed," i.e. self-contained units that can be evaluated by looking at the system's internal integration of parts and resources; or "open," i.e., the activities are interdependent with forces outside the system's boundaries (Thompson, 1967). The "open" system concept is probably more appropriate in considering training activities in work organizations, implying that any organized unit's effectiveness should be seen as a reflection of how well it is integrated into other systems that penetrate it at many points, i.e. sources of new knowledge, new training techniques, etc. Systems of either the open or closed type can be assessed in terms of: "a pattern of interrelationships among the elements of the system which would make it most effective in the service of a given goal . . ." (Etzioni, 1964).

In training evaluation, the organizational support approach involves assessing the extent to which a training effort develops within a work organization such that it gains the support of other system components and prevents the overall organizational system from returning to its former "untrained" condition. Obviously, efforts to develop these support systems for training and for the maintenance of training results call for the use of the training department's resources for purposes other than the specific achievement of training goals. Herein lies this approach's chief difference with the goal attainment approach. Evaluation of the training efforts consists of how thoroughly training specialists develop and sustain a support system for their efforts within the organization.

The most thorough example of the use of this approach is reported by Nadler (1971). He sets forth five elements of a support system for training, which focus on improving the effectiveness of the employee in his present positic·, not on the broader notion of human resources development. These five elements are: (1) organizational involvement, (2) pretraining preparation, (3) training activity and training period, (4) job linkage, and (5) followup. Unfortunately Nadler was unable to point to a training program in which all the elements were present, but he does cite many individual actions that exemplify the various parts.

Examples of organizational involvement are: development of specific training policy at policy-making levels in the organization which are then put into concrete action. "When training the hard-core, the Boeing Company made a specific provision for allowing a 7½ percent production differential upon completion of the training program. This action put the company's production schedule and financial resources behind the particular training program so that it would be obvious that there was company involvement" (Nadler, 1971:3). He also points to "training to support training," i.e. other related groups must be trained if they are to understand trainees and the changes that training renders in their behavior.

Such "secondary training" is especially appropriate for supervision wh· will work with the products of training programs for the hard-core unemployed (Niederfrank, 1970). Organizational involvement and support are clearly present when future supervisors of such trainees come to understand (through training experiences designed especially for them) the fatalism, orientation to the present, and extreme concreteness that often characterizes those from deprived backgrounds. The usefulness of secondary training is not limited to this type of program, but is applicable to any training where it is anticipated that the success of trainees is contingent upon other organizational members' understanding of and accommodation to their post-training behavior, styles and attitudes. Furthermore, training-to-support-training need not be limited to those who will be the immediate supervisors of the trainees, but can be extended to other levels of line personnel as well as staff people.

In an even broader context, use of organization resources to generate community support and direct involvement of influential business executives in a Work Experience Program illustrates the organizational involvement dimension (Levinson, 1966).

The pre-training preparation dimension of organizational support attempts to prepare "the various individuals and groups who are most directly related to the training activity" (Nadler, 1971:4), including the trainees, supervisors and even peers. In terms of the system concept, this initiates the innovation of the training outcomes into the existing system, and prevents the necessity of the trainees' "going it alone" when they return to the job after training. Nadler described how the Ford Motor Company exposed supervisors of trainees to a shortened version of the training program given to the trainees in order to improve "cultural climate" for training. Short "run-throughs" for trainees, exposure of them to videotapes of actual sessions, and ready availability of training personnel near the work location of trainees for discussion and questions would be other preparation techniques. Also, specific planning for the trainees' replacement during his training period without producing insecurity makes for more preparation. Similarly, the selection of trainees is also a part of the pre-training preparation. The degree of investment in such preparation is of course contingent on the type, level, and anticipated impact of the training effort.

In the organizational support approach, selection involves more than individual testing. The whole question of who is given training and how they are selected for this experience may be pivotal in determining subsequent support for training outcomes. Selection of trainees in the organizational support context "includes the involvement of a variety of individuals in selecting the proper trainee for the appropriate training opportunity" (Nadler, 1971:4). While outsider resources for selection such as the assessment center program (McConnell and Parker, 1972) may be attractive, the assumptions of this approach make it especially important to consider how those who will work with the trainee when he "gets back home" to his job participate in his selection.

Continuity for the trainee while he is in training and when he re-enters his work situation also illustrates the organizational support approach. Relatively long-term absence from the job for training purposes may undermine training results if the work situation changes during the trainees' absence or if the work group system is altered by its adjustment to the absence or to the trainees' replacements. Nadler (1971) cites "spaced learning" as a device to keep the trainee partially in the system even though he is in training. The New York State Department of Labor conducted a program where the trainees returned to jobs and homes regularly during the three phases of their training, enabling them to devise and plan their own "back home" programs without losing contact with either the training or their own unique job situation.

This tactic relates to the job linkage dimension of organizational support. Much training cannot use spaced learning and "reentry" difficulties become the overriding question. Has the support system for the training been strong enough to provide the trainee with opportunities to actually use his changed attitudes and behaviors on the job (assuming the training produced such results)? Has the training produced expectations that probably cannot be met (Quinn, *et al.*, 1970)? This becomes the crucial and disturbing question. The extensive study by Fleishman *et al.* (1955) showed that although training resulted in immediate changes in the self-perceptions of the trainees, this change was overwhelmed by the leadership style of the trainee's supervisor as soon as he reentered his job context Nadler suggests providing trainees with checklists for use back on the job and the inclusion in the training of specific sessions for role-playing situations which illustrate the incompatibilities that may arise back on the job. These devices do not, however, substitute for organizational involvement and pre-training preparation as a way to reduce reentry problems.

Training followup may act to reinforce the changes hopefully produced by training. It strengthens the trainee in retaining the results of training even though his work milieu is somewhat incompatible. According to Nadler (1971), Boeing Aircraft trainers ask the trainee to select and write down a specific behavior change that he expects to carry out when he returns to the job. The trainee makes a carbon copy which he puts in a sealed envelope. Several months later the training staff mails the statement to him as a reinforcement device.

More specifically related to the organizational support approach are reinforcement sessions for supervisors, subordinates and peers regarding the post-training behavioral changes that have taken place and these persons' evaluation of how the changes fit into the improvement of the organization's efforts. The stage might be set for such reinforcement sessions for trainees' "significant others" by a pretesting period before training in which a "dry run" or "walk through" of the entire supportive system, including a short mock-up of the training itself, would take place. Eastman Kodak (1971) does some of this type of analyzing, developing, and pretesting training programs through its Marketing Education Center. All in all, one of the most obvious indexes into training effectiveness, using the organizational support approach, is to measure the extent to which trainers have prepared the "back home" work situations for training changes and the extent to which "significant others" actually behave compatibly with the changes rendered by the training.

The organizational support approach in training evaluation has as its basic strategy the determination of the existence of supports without which training will in all likelihood be ineffective. It is a dynamic approach, catching those organizational factors that are the natural everyday processes surrounding training and job performance. It is especially appropriate for the initial stages of a training program when goals are apt to shift and objectives remain fuzzy. From this standpoint it can be the approach used first, followed by the addition of a goal-attainment approach as the program matures.

The organizational support approach calls for a perspective that is radically different from traditional approaches to evaluation. It draws attention away from specific techniques in training to the extent that it focuses on the fact that long-term training successes are contingent upon the systems in which such training is undertaken and sustained. The approach can be valuably applied to an understanding of an organization's entire training component, aside from its use to assess specific program successes or failures. In this broader context, the approach can draw attention to situations, for example, where the support for training efforts is almost totally based on a training director's charismatic and persuasive personality, pointing up the fact that the program could collapse if he vacated this leadership position. On the other hand, a systems analysis may reveal unknown and underutilized sources of organizational support. The basic value of the approach is its broad focus on "the big picture."

Yet the approach's disadvantages are many. It has never been fully operational and its use will likely reveal many unforseen problems. It places unusual demands on evaluators by requiring they learn a great deal about the organization within which the training takes place. Furthermore, objective criteria for measuring the strength of the social support are impressionistic rather than quantitative in most instances, i.e. how does one measure the "amount" of pretraining efforts and their results in preparing trainees and their "significant others" for the training to come? There is a large array of research techniques that can be used to explore the usefulness of the approach.

The systems model, however, appears to call for complex techniques and this might be substantially more costly than other approaches. The ambiguities and difficulties in collecting "hard data" within this approach may create resistances among both trainers and evaluators who are committed to the idea that quantification is the solely acceptable product of evaluative studies. Especially difficult is the notion of trying to decide how close a given program comes to an "optimum" allocation of resources (Etzioni, 1960) for training among goal and non-goal functions.

Despite these drawbacks, the emergence of Organizational Development programs and their current popularity may well force a wider use of the systems model. By emphasizing such techniques as systemwide multitechnology, teambuilding and long-range "organizational training laboratories" (Friedlander, 1967), O.D. programs become less and less subject to evaluative, experimental designs such as those encouraged by Dunnette and Campbell (1968). More appropriate to O.D. are those non-experimental techniques such as interview data that will pick up changes otherwise missed by experiments (Argyris, 1968a, 1968b).

The Outside Expert Approach

This strategy needs little elaboration. Its use assumes that specialists can be found who can carefully review the training content and "action," its milieu, and the training personnel, and then render a judgment about effectiveness. Obviously such "clinical" rankings cannot specify whether techniques, personnel, timing, etc. secured desirable or undesirable outcomes. The approach, however, does not assume this. It is oriented to bringing expert scrutiny into the program from outside; hopefully the expert will offer detached suggestions and questions, as well as provocative comparisons.

In the approach it is assumed that such outsiders will have systematic guidelines for judgment that have been developed from successful and objectively evaluated programs. In sum, the approach calls for a knowledgeable and experienced training specialist to review a program, judge it in the light of his experience, offer guidelines for "shaping it up" from a perspective unavailable to organizational "insiders." Lester (1971) discusses the use of a check list in the selection of training materials, an item that an outside specialist could use in appraising a training program. With an eye to the outside evaluator strategy, Tracey (1968) devised a comprehensive manual of guiding principles and elements of evaluation that provides means for identifying strengths and weaknesses, judging trainers and rating practical exercises.

An example of the use of this approach involved an independent jury of adult educators who evaluated the teaching plans of a random sample of other adult educators in a Cooperative Extension Service (Evans, 1970). Another example is Schmidt (1970) who describes how United

Airlines submitted its training committee findings concerning training needs, procedures, and strategies to a panel of outside training specialists.

It is difficult to gauge the effectiveness of this approach simply because of the variance in the evaluative skills of "outside experts." The outside specialists' assessments remain essentially judgmental, but they have the virtue of experienced input that broadens perspective and firms up objectives and goals.

The Native Approach

In contrast to other approaches which impose assumptions and techniques on training programs, the native approach sees the task of evaluation as learning as much as it can about the indigenous, subjective, evaluative processes at work among training people. The object of the native approach is to supplement these rather than replace them with "imposed" models. It is assumed that training personnel "keep score" of their activities in some fashion; this "native" pattern of evaluating can be improved by diplomatic inputs from the evaluator. Thus it is assumed that existing methods are already adopted and accepted; building upon these methods may sidestep the difficult problems of "accepting" evaluation within the training context and encourage healthy attitudes of participation in decision-making on the part of training personnel. In a real sense, this approach to evaluation is based on certain *training* assumptions: the adoption of new ideas may proceed more rapidly if they are based on existing attitudes and behaviors.

For example, a training specialist recently told us, "This program has to be good; we maintained our 'student body'; they kept coming back." He kept score by counting how many continued to return to the sessions. In this setting the stage is set for a sample of all the trainees to be interviewed, from which a questionnaire can emerge. Such an instrument could then seek out the differences between those who continued to come to the training and those who did not, tapping directly into the training specialist's "native" technique of evaluation.

A larger example of an opportunity for the native approach comes from a study started by 50 state directors of education for gathering information from graduates of public post-secondary vocational and technical programs. These officials named over 168 local administrators who used systematic followups of such students at the local level (Goff, 1968). Thus this method revealed existing elements of a substantial outcome study.

The native approach calls for the evaluator to find compatible ways to introduce sampling, comparison groups, reliable and valid instruments, and sound statistical devices to improve existing evaluative strategies. Even when these strategies are primitive, the approach calls for building on the crude techniques and assumptions that are evident.

Such an approach can sharply reduce the resistance to *use* of evaluative results found among trainers, as well as members of support systems. It builds on their own method and approach, sharply increasing the likelihood of using the results to alter current programs. Bolar (1970) puts the matter succinctly: "It is necessary that the methods employed to measure the effectiveness of training be meaningful and acceptable, not only to researchers and trainers, but to the wider group in the organization concerned with any organizational activity."

At the same time, the approach has the problem of accentuating biases in favor of one's own work. Moreover, even if the training staff is amenable to the improvement of its methods, the approach calls for an unusual evaluator, well-equipped and versed in many evaluative techniques that he can call upon as he encounters a wide variety of native schemes. The approach also assumes a substantial degree of consensus among training personnel about how to keep their own score, i.e. the evaluator can only cope with a limited number of native methods. Hopefully, however, native schemes can be improved to more efficiently answer the question of training effectiveness.

Integration of Approaches

For the purposes of understanding, we have separated the descriptions of the different approaches to evaluation of training. In reality, however, the evaluator does not choose a single one of these methods, but designs his effort to include the *combination* of these approaches most appropriate to the demands and limits of the training situation.

Nearly all evaluations involve the goal-attainment approach, either implicitly or explicitly. Depending on the goals, the more specific approach of cost-benefit may or may not be desirable and/or feasible. While the ceremonial approach is a means of dealing with the question of goal attainment, it can be easily combined with more formal goal attainment approaches. Its primary value is in drawing attention to "Hawthorn effects" and to the functions of training in status and role passages.

The organizational support approach constitutes a broad evaluative strategy which can be combined with any of the goal attainment approaches; it may have the valuable spin-off of clarifying goals, their origins and their supporters. As may be evident, we see a great deal of value in the organizational support approach in terms of understanding systems and sub-systems within organizations, as well as their viability and integration. It seems inconceivable that any attempt to assess organization support for a program or for a specific training effort could be a "failure"; it would doubtless yield valuable information about the organization regardless of the training outcome. Most importantly, the organizational support model focuses attention on *why* training efforts succeed or fail, whereas "straight" goal-attainment or cost-benefit designs may tend to be overly concerned with specifying outcomes than with the sources of outcomes.

The outside expert approach likewise can be combined with any version of a goal attainment or organizational support approach, either as an additional criterion of outcome, as a means for providing a broad evaluative perspective, or as a pre-evaluation input to establish program goals. An intriguing combination is the assessment of the impact of the outside expert via the ceremonial approach. The outside expert approach may lend itself to the development of the systems perspective required by the organizational support approach, particularly since an outsider may be in a good position to identify internal subsystems and the degree to which they are mutually supportive.

In employing the native approach, the evaluator will most likely identify some existing version of the goal attainment approach. Here the task is to systematize and improve on the goals designated in the approach, and perhaps combine the approach with a version of the organizational support model. Our experiences indicate that the systems concepts contained in this approach are easily communicated to training personnel; they quickly follow the logic of the approach and are eager to "test it out" once they have been exposed to it. Finally, the use of the native approach implies some sort of outside evaluator, making the strategy readily adaptable to at least partial application of the outside expert approach, assuming the evaluator has the adequate experience and appropriate orientation to make such judgments.

FUNDAMENTAL TECHNIQUES IN IMPLEMENTING TRAINING EVALUATION

There are several behavioral science techniques which are fundamental to conducting training evaluation. The use of them varies depending on the approach or combination of approaches that the evaluator employs. These techniques include record-keeping, observation, experimental design, interviewing, questionnaires, and statistical analysis.

Record Keeping

Although at times tedious, and often viewed of as mundane, good records about training efforts and techniques used are invaluable. Unfortunately, good records of the training "action" are rare. As a first step, descriptions of the *process* whereby a training program was mounted and sustained within a work organization should be kept as thorough as possible. The evaluator should develop systematic files on: (1) a natural history of how a training program started, its original objectives, sources of initiation and support, and patterns of participation in the planning phase; (2) data on the trainees themselves, how they initiated involvement or were recruited into the training program, their age, sex, tenure, company division, and occupational background; (3) the trainees' attendance pattern and their reactions to various aspects of the training action; (4) how much time was involved in planning and executing the training effort, the monies and other resources invested, the company personnel who participated in various phases of the effort, the training strategies used, the duration of each strategy, and the organizational conditions under which the training occurred; (5) the impressionistic assessments of those conducting the training, including their feelings about its main defects and strengths, how they would change it if it were repeated, their own frustrations and satisfactions sensed in conducting the effort.

Accuracy, specificity, and thoroughness should characterize record keeping. Forms for specific types of records should be prepared, and it is desirable to have them completed by several different persons in order to check on their reliability and precision (Belasco and Trice, 1969). Standardization of forms should be minimized so that records may be adapted to the differences that inevitably mark one training effort apart from another. Records must be up-to-date and relevant to a specific training program.

A simple but effective device for maintaining a thorough natural history of a training effort is a "work diary" kept by the training personnel who participate in the formulation and launching of a specific program. Patterned after a typical diary, the writer daily records the highlights of his experiences. The more pivotal an individual is to a particular training effort, the more important his diary is for the evaluation. Like a personal diary, the entries are frequently terse, impressionistic, and often personal. Consequently, they are often retained in the sole possession of the writer until edited to be shared with others. Diary writers should be encouraged to spend twenty to thirty minutes per work day recording what they believe to be the main activities of the training effort at that point, their own reactions to these activities, and their perceptions of the reactions of others.

Observation

Direct observation is a very important means of operationalizing the evaluative approaches we have outlined. The outside evaluator is probably the best equipped to observe since he has had minimal previous experience with the particular training situation and the personalities of those involved. The outside evaluator should carefully note how a training technique is carried forward, how a trainer conducts his efforts and how the trainees respond to the training setting and contact. He listens to verbal exchanges and watches the physical action. While this is usually done in conjunction with testing (Weingarten, 1971), observation is nonetheless a major tool of evaluation. All too often evaluators regard data collection as the sole basis for evaluation. While other measures may provide external or independent checks of the observation records, systematic, detailed notes can pick up a great deal missed by testing instruments.

Attempting to be as detached as possible, the observing evaluator should be marginal to both trainees and trainers so he can objectively observe, even participate, in their "back region," hopefully tapping the various sentiments and behaviors which group members usually keep to themselves (Berreman, 1962). He watches and listens fo. signs of use or rejection of training ideas, and becomes a "participant observer" as much as possible without developing commitments or losing his objectivity. Such an observer should never attempt to disguise who he is or what he is doing: "Be who you are" is a simple rule. The observer should avoid cross-examination type questions such as "who," "when," or "why." He should "hang around" and learn the answers to his questions without creating the resistance often produced by direct questions. He should never directly reveal the specific comments and behavior of particular trainees to anyone else, including the trainers. He can paraphrase comments and analyze trends, but he should make every effort to respect the privileged nature of his observations. In sum, observation requires a delicate balance of "insider" and "outsider" status, collecting as much unbiased information as possible without betraying his ethics. The effectiveness of observation can be greatly enhanced by having multiple observers who compare notes after completing their observations: in this way they may independently check the validity of each other's preliminary conclusions. Dralle (1969) describes how two observers recorded verbal behavior during training laboratory sessions: he indicates that they had few problems agreeing on the direction and content of communications, but generally disagreed on the affect or tone of these exchanges.

Accuracy is a basic problem in observation. By careful recording of his observations in systematic field notes, the observer develops a "field diary." By recording in diary style as many observations of trainee behavior as possible, the observer has details to be reviewed in broader perspective later. The timing of recording such notes is important, and oftentimes it is necessary for the observer to briefly absent himself in order to record his notes or attempt to recall as many points as possible for recording at an appropriate break in the training. While such recording should not be secretive, the observer who writes down a great deal in front of the observed may create a "Hawthorne effect" on the trainees. The act of recording field notes is important in that it helps the observer to concentrate on his detached, objective role. Hopefully he can, without deception, become a part of the routine of the training so that he can get a feeling of "what it is like" to be in the course. Observations of critical incidents that occur during the training period may give him additional materials. For example, in one training program a particularly sharp exchange between a chief shop steward and a general foreman during a panel discussion on grievance procedures set the stage for a subsequent intense and revealing discussion about the value of training.

Observation can be combined with research interviews and questionnaires to provide multiple perspectives on training results which may be compared with each other. Dada (1970) reports programs in an Adult Education Center which were evaluated in this way; Tolela (1968) evaluated T-group training by combining observations of group interactions with solution analysis and questionnaires. Becker and Geer (1963) studying the training of physicians, discovered that participant observation enabled them to "check description against fact" following research interviews with medical students.

Basic to many evaluative studies is experimental design. This constitutes a relatively rigorous attempt to ascertain the specific sources of training outcomes.

The Experiment

Rather than depending solely on the trainees' or trainers' subjective reactions, experimental designs attempt to be more objective via "strategic comparisons." For example, "before-after" is a commonplace design calling for comparing the behavior or attitudes of trainees on a criterion of change *before* training with their performance on the same yardstick at some point *after* the training experience, i.e. at the end of training, at the point of reentry, or at some point following. An experimental design, however, becomes more powerful if trainees can be compared before-and-after with another group (often called "control group") which did not receive the training. Such a design tries to ascertain if revealed changes can be attributed to the training. Further refinements of experimental designs can grapple with the important but often neglected question of whether the revealed changes were effected by the training or are the results of the evaluative study itself. The impact of evaluation, or of any kind of research, upon the subjects being studied is a very significant issue to be kept in mind in training evaluation.

Different types of experimental designs include:

1. The single group before-after comparison, discussed above. If "after" measures on the criteria of success show desirable changes, success is usually attributed to the training. Such results can conceivably stem from research effects, but this design cannot distinguish between the effects of training, the effects of evaluation, or the effects of other changes elsewhere in the environment which may have impacted upon the trainees during the training period (Borus and Buntz, 1972).
2. The two-group before-after comparison, with one group being administered the criterion yardsticks before and after the training period, but receiving no training. If the before-after changes in the training group are greater than before-after changes in the comparison group, the training is assumed to have produced these desirable results.
3. The two-group after-only-comparison design compares the training group with a group which has not received the training on the same criterion yardstick *after* training is completed. If the training group shows more desirable changes on the yardstick the training is regarded as successful. This design is especially dependent on the assurance of similar "starting points."
4. Finally, a "four-way" design comes closest to detecting the effects of the training vs. the effects of the evaluation study. Two more comparison groups are included with the two before-after-comparison groups. One of these additional groups receives the training and the "after" measures only. The other receives *only* the "after" measures. Comparisons of the four groups on the criterion yardstick allows for specification of the effects of simply filling out the criteria questionnaires (or being interviewed). If those receiving the training rank higher on the criterion yardsticks than those who simply were administered the yardsticks, then this difference can be specified as the result of the training.

The designs indicated in (2), (3) and (4) require legitimate comparisons. This means that the characteristics of the groups which do and do not receive the training must be as similar as possible. Thus, if the group that is administered training has some unique characteristic (e.g. higher rank, previous training experience, a stated desire to participate in the training, etc.) which is *not shared* with the non-training groups against which comparisons are made, such comparisons are illegitimate because the groups had different "starting points" when the training commenced.

Two ways of achieving legitimate comparisons among groups in experimental designs are (1) matching or (2) random placing of trainees in comparison and training groups. For example, in evaluating a management development program, Valiquet (1968) used a comparison group of non-trainees each of whom were matched with a specific trainee on such features as sex, age, occupational level, etc. Randomization requires sampling from the same population to secure both the training and the compaison group (or groups); it can be achieved by alphabetizing trainees, starting at a random point, and placing names in trainee and comparison groups in alternating order.

Each of these experimental designs suffers from specific flaws. The before-after design, lacking any control groups, cannot come to grips with the basic evaluative question regarding the training. Its results are contaminated by "research effects"; the experience of being tested about what the training aims at producing often brings about as much change as the training itself. If trainees are tested on the same criteria both before and after training, the differences revealed may well be due to the sensitivity and "test-wiseness" elicited by the "before" testing (Trice and Belasco, 1968). The simple before-after design without comparison groups is widely used, however, and in at least one instance its use has been shown to overstate effectiveness (Borus and Buntz, 1972:235).

A major drawback of the after-only design is that it requires very large numbers. 250 to 300. This is necessary to assure the randomization of all possibly relevant factors within the comparison and trainee groups. This design does. however. have the distinct advantage of avoiding the research effect of most other designs and the complexity of the four-way design.

The four-way, before-after-three-comparison group design is quite complicated. It calls for the development of not just one comparison group, but *three*. with all the sampling problems involved in such an effort (Trice and Belasco, 1969). Not only must the evaluator sustain access to the four sampled groups on a strict basis, but he must also effectively explain to the comparison groups *why* they are experiencing what they are, e.g. no training. For example, one comparison group will receive no "before" measures, no training, and only "after" measures. One must be a master salesman to convince those having these experiences that they have meaning. In most instances, this design is simply impractical for training evaluation.

The biggest potential defect in any of the designs employing comparison groups is illegitimate comparison between trainees and non-trainees. Although random placing of trainees and others in one or the other group is relatively straightforward and an adequate way to get acceptable contrasts, this is usually difficult. Work schedules force some persons randomly placed in the comparison group to be put in the trainee group instead. Individuals in comparison groups can make intense demands to be placed in training. Frequently the pressure to "take them all" makes randomization nearly impossible. Even if it does stand initially, a management decision may withdraw some of those in one or the other two groups. Dropouts from comparison groups can alter their representiveness.

Matching may yield meager comparisons. such as the comparison group in Trice's (1959) study of conference leadership training. Although it is not immediately obvious, matching persons on several characteristics requires a large pool of potential candidates, even if one is using only three or four characteristics in the match. As the number of matched characteristics desired increases, the necessary size of the pool in which to locate matched persons for comparison also increases. Furthermore. matching assumes one can decide on the pivotal characteristics that might affect training response before the training begins. Overall, randomization is a more desirable and clearly a more simple procedure, but it presumes a relatively large number of candidates to be placed in both the training and comparison groups. A suggested compromise means to achieve matching is to ask trainees after they arrive at a training site to pick two peers back on the job who would be asked to rate the trainee after he returned in contrast with another trainee who the peer knew had not been in the training (Friedlander, 1967; Miles, 1960). Such peer judgments can, however, generate other problems.

Methods of Data Collection

The most widespread method for collecting data in training evaluation studies is to query trainees in some fashion, the subjective method. In contrast to the objectivity of observations or specific measures of work performance, queries directed at the trainee must be filtered through his own systems of perception, motivation and affect. We can (1) try to determine the trainee's favorable or unfavorable reaction to the training; or (2) attempt to measure if he has changed his knowledge, his attitudes, or his behavior as an outcome of his training experience. Strictly speaking, (1) is not evaluation since it does not focus on some aspect of change, but it is helpful for future planning to know what trainees liked and disliked, what suggestions they have for altering the training format, and what scheduling and arrangement problems they encountered in attending the training. Even though intense observation and performance measures are attractive, the questionnaire, or the research interview often turn out to be the most feasible.

The research interview is not a "non-directive" experience (Whyte, 1969), but is a structured, talking relationship characterized by degrees of directiveness. The interviewer decides in advance how much structure he wishes to use in guiding the verbal interchange. Whyte (1960) catalogs these degrees of interviewer directiveness as follows:

1. **Minimal**: "Uh-huh," a nod of the head, or "That's interesting." Such responses simply encourage the informant to continue and do not exert any overt in-

fluence on the direction of the conversation.

2. **Reflection**: Let us say the informant concludes his statement with these words: "So I didn't feel too good about the job." The interviewer then says: "You didn't feel too good about the job?" repeating the last phrase or sentence with a rising inflection. This adds a bit more direction than response 1, since it implies that the informant should continue discussing the thought that has just been reinforced.
3. **Probes**: The probe may be directed at the last remark by the informant, an idea preceding the last remark by the informant but still within the scope of a single informant statement, or on an idea expressed by informant or interviewer in an earlier part of the interview (that is, not in the block of talking that immediately preceded the interviewer's probe). Each of these probes represents the introduction of increasing structure into the interview.
4. **Introduction of a new topic**: Here the interviewer raises a question on a topic that has not been referred to before. The more such new, specific topics the interviewer introduces, the more structured the interview becomes.

The interviewer selects one of these strategies and then prepares some guideline questions. Even (1) calls for questions to direct the subject toward a topic ("What was the effect of the training on your schedule?"). More directiveness calls for more guideline questions as well as probing following a response, or an *interrupting* probe. Here the "how," "what," "why," and "where" kinds of questions apply. The interviewer wants to keep the interview "on the track," sometimes by summarizing what was just said, but adding a specific question ("You say the training interfered with a tight schedule. How, specifically, did it interfere?"). Belasco and Trice (1969) have set forth other features and refinements.

It may be possible in an evaluative study to conduct a series of relatively unstructured research interviews which yield enough information to construct specific questions which can be tabulated in contrast to the narratives developed in open research interviews. Such "closed questions" are those which call for either yes or no answers ("Did you have to choose between going to the retirement party and the training session?"), have a limited number of alternative responses ("Which of the three conference leaders did you like the most?"), or ask for specific facts ("On what days was the training suspended?"). Even though directiveness, in some degree, is a basic feature of the research interview, experience suggests its degree of directiveness should be fitted to the overall evaluative strategy. If it is the only method that is to be used in the study, then more control is appropriate. Use with observation and questionnaires suggests less directiveness. One may desire to use research interviews with the training staff and other pivotal people involved in mounting the training effort, and employ relatively structured interviews with the trainees.

In any event, relatively unstructured research interviews are essential for the construction of more structured and "close-ended" instruments. The most frequently used instruction in evaluation studies is the questionnaires which may or may not include specific scales. Even through questionnaires are widely used, they are also much misused. Evaluators and training specialists are especially prone to dash off a questionnaire or a scale in the quiet of their offices largely unaware of the often tedious process involved in producing a good instrument to get at subjective feelings and meaningfully relevant responses while such "armchair" instruments may have considerable "intuitive validity" to their creators (and oftentimes prove to be good instruments), a series of steps are essential if an instrument is to meet minimal scientific criteria.

Questionnaire construction should begin with observation or research interviewing. While most questionnaires are made up of "closed" questions, there are numerous "open-ended" types. Thus, "Please tell me in your own words how you feel about the supervisory training course you have just finished?" is a subjective open question. Put in closed torm this question might be, "Would you say the supervisory training course you have just completed was: (a) a good one, (b) a poor one, (c) one that didn't matter to you?" Although the open type does not confine the respondent as does the closed type, the open type is less often used since responses are varied and require a substantial time to classify and code for tabulation. Furthermore, respondents are often unable to exoress themselves in their own words in writing. Even if they can, they frequently are unwilling to take the time to do it well. For these reasons questionnaire construction is typically of the closed type despite the richei subjective items to be found in open questions. For these reasons it is essential that questionnaire development be linked with observation and research interviewing. The results of these methods provide the varied raw materials from which items can be made.

When questionnaires are based on research interviews conducted in the same respondent population, it will provide respondents with items that more closely fit their unique feelings and experiences than would items based on intuition or the work of researchers in other populations.

After the questionnaire maker extracts questions and appropriate response categories from observation and research interviewing, he frames a "dry run" instrument. This he tries out on a small sample of trainees. asking them what words are confusing, what difficulties they experience in responding, and how they would reword items or responses. He should make a deliberate effort to find out if the questions and response items provided force the pre-test respondents into reactions that distort or minimize how they truly feel. Such pretesting usually results in numerous changes in the questionnaire, followed by a second trial run, often with the same pretest subjects or with employees who have had similar training but not in the specific program being evaluated.

The careful questionnaire maker also wants to know the reliability of his instrument, i.e. if it produces the same answer when used over and over with the same individual. If it is not reliable, its instability may give the illusion of change when groups are compared, or can even produce the impression of change when no change has actually occurred. Closed questions provide for a "test-retest" reliability check. Once reliability has been established the question of the validity can be raised, i.e. does the questionnaire reflect the true feelings of the respondent as indicated in the pretest? Often closed types produce a "response set." That is, respondents may fall into an automatic response pattern without thinking especially if he tends to be a "yea-sayer" or "nay-sayer." Thus lower validity. Questionnaire makers look to how much observation and research interviewing has preceded the closed instrument, how many "open" questions were used in pretests. In sum, how valid the form is rests on the extent it reflects the *range* of feelings and experiences being tapped.

Scales attempt to set up degrees of reaction or experience. Scales are comprised of a cluster of questions which tap dimensions of the same general concept, i.e. degree of liking for supervisory roles. Questions which strongly correlate with one another are often placed together to form scales, and specific techniques such as Guttman scaling require a series of statistical manipulations on pretest data to derive the ordering of questions which comprise a scale. Once a scale is constructed in this relatively complex manner, one can assume equal intervals in the attribute being measured, much like a ruler. Most evaluators are limited to constructing scales on the basis of correlations, which is short of the ideal and which clearly prevents the assumption that persons who, for example, score answer 4 items positively in a given scale possess "4 times as much" of a certain attitude or attribute than a person who answers only one item positively.

Scales usually call on the respondent to select a degree of agreement or disagreement with an item. Belasco and Trice (1969) illustrate the simple Likert type scale with the following items:

1. I feel I shared my anxieties about the job of supervisor with the others present so that I feel better about my job:

Strongly agree Undecided Strongly disagree
1 2 3 4 5

2. During the training sessions I made, or renewed, a friendship with another supervisor with whom I later discussed supervisory problems:

Strongly agree Undecided Strongly disagree
1 2 3 4 5

In contrast to most structured questions, scales of this type do not force respondents to take a position even though they may not have one. These features suggest a modest degree of freedom to express a range of feelings.

Statistics

In practically any evaluation effort, statistics sooner or later are very helpful, if not essential. There is no way to avoid, at the very least, measures of central tendency such as the mean, the median, and the mode. Complex evaluation designs, such as experiments, cannot be completed with technical statistical tools.

For example, an evaluator may want to group the responses of trainees on a questionnaire according to certain kinds of reactions, such as the extent to which they agree or disagree with certain statements. He must classify the range of responses into categories, and then use these categories to calculate averages for trainees grouped according to some characteristic which may be predictive of the response in which he is interested. Ideally, such groupings are mutually exclusive with no item falling in between the groups, although missing information or "don't know" responses can make a series of calculations more difficult.

Such grouped information sets the stage for graphic analysis and illustration: bar graphs, "pie" charts, and line graphs, as well as presentation of percentage tables. Distributions revealed by grouped data can be portrayed by charts of frequency distributions, such as the number of discussion participations over ten minutes by tra ees according to their status level in their organization. Such a chart might reveal that fewer lower status trainees contribute to discussion in the presence of higher status personnel, suggesting that training be confined to peer levels.

Averages show central trends. The arithmetic mean (total of scores divided by number of scores) tells the point in a distribution around which the most instances cluster. The standard deviation indicates the amount of spread around this point, indicating that scores are relatively similar or different from each other. Simple means reveal trends in responses to questionnaires, background information, recorded observations, interview results, and other data generated in training evaluation. Also useful are the median (the point in a frequency distribution that falls in the middle) and the mode (the score that occurs most often in an array).

Once means of groups have been developed, the evaluator can calculate differences between means to ascertain their "statistical significances", i.e. an estimate of the probability that the differences simply occurred by change. This statistic enables him to compare, for example, trainees with non-trainees on criteria of desired change. A similar measure is a "significance of difference between proportions" statistic (Yoder, 1956). There is a wide variety of statistical devices available for comparing groups for evaluative research purposes. The widely used Chi-square statistic often enables an evaluator to decide if there are significant differences between relatively comparable groups who differ on such experiences as training vs. non-training or different types of training experiences. Knowledge of the various types of correlation, such as the pitfalls and proper uses of Pearson's "r" and the ranking method can be used to great advantage in evaluation studies. For an excellent set of practical examples of the use of statistics in training evaluation see Reeves and Jensen (1972).

As mentioned, an evaluator should have a basic understanding of sampling so he can ascertain the limits to which his findings can be generalized. A sample is an attempt to *represent* a *population*; the evaluator must be sensitive to how this population is arbitrarily defined. If sampling is ignored or only considered "after the fact," it is likely that the evaluator will not know for what kinds of employees training is effective.

The evaluator or trainer should not be overwhelmed by the range of different data collection and analysis techniques that can be employed in training evaluation. None of them constitutes an absolute ideal and each is flawed in its own way in terms of the completeness of data collection. While typical limitations usually lead to the use of questionnaires on trainee populations because of the cost and time involved in interviews and observation, there is every reason to argue that a combination of these methods at different points in the evaluation will produce a more comprehensive evaluative picture than restriction to a single method. Likewise, the use of consultation to develop more sophisticated evaluative designs and analyses (without substantially increasing the costs of the effort) may lead to evaluative outcomes which not only strengthen the specific training activity but which also enhance the image of the training components in the eyes of organizational decision-makers.

The Criterion Problem

We have frequently referred to "desirable changes" as a hopeful outcome of training. A fundamental problem in the methodology of evaluation is the yardstick for gauging desirable changes. Training can conceivably produce changes in trainees' knowledge, attitudes, behavior, and job performance. Criteria to judge a training program could come from any or all of these areas. Clearly some of these are more cogent than others in a given situation. Criteria indexes for gauging changes in job *results* (Roy and Bolke, 1971) are usually more meaningful in the work organization than those measuring changes in knowledge or attitude. The comparative value of one program when contrasted to another is more relevant than measuring the amount of effort invested in a single program, i.e. "trying does not equal success."

Regardless of the evaluative model used, a success criterion must be applied. Criteria used in evaluation which employs the organizational support model are less precise and less quantifiable than in the goal-attainment model. They can, however, be quite comprehensive (Farmer, 1970).

By contrast, the goal-attainment and cost-benefit approaches call for pinpointed criteria. Such precise criteria can of course be limited; as Williams (1969) cautions, "only specific aims are measurable and open to evaluation but management finds it easy to jump from specific to more general aims." But beyond doubt, any experimental design calls for criteria that are measurable and exact. One of the valuable "spin-offs" of evaluators' exploring experimental approaches is forcing training officers to come to grips with the precise purposes of the training. One of the simplest examples is the development of training time measures for retraining programs.

Good criteria in training evaluation have: (1) range, (2) quantification, (3) reliability, (4) validity, (5) relevancy, and (6) independence. Range means that the yardstick used should yield different scores from individuals in a trainee group, ranging up and down the index. A criterion is quantifiable if its range breaks into equal intervals, much like a ruler: this allows for arithmetic manipulation. Criteria are reliable if, within certain confidence limits they yield approximately the same results if re-administered to the same group, i.e. measures of intelligence are reliable if they produce the same score when re-administered to the same individual. Criteria are valid if they actually measure what they are designed to measure, i.e. the evaluator's measure of job performance should correspond to the performance standards typically applied in the work organization. Criteria are relevant if they stem directly from training objectives or program planning, are related to the organization's goals. Criteria are independent if they are free of possible changes occurring external to the training program.

Validity is the single most important characteristic of a criterion. It is, however, the most difficult to achieve. Subjective methods often come to rely on "face validity," i.e. do the changes detected *seem*, on the "face" of the situation, to be logically related to the training?

For example, if a training program contains some very unique materials and information which are quite unlikely to be encountered elsewhere by trainees, these set the stage for the evaluators' reliance upon face validity. If subsequent research interviewing shows that trainees have adopted and used these unique features, a basis is set for arguing that it (training) has been effective, using a criterion that is "obviously" valid (Bell, 1968). Bell and Honour (1969) suggest, however, that when trainees are asked to recall how their situation after training has changed from what it was before, they tend to overemphasize desirable changes.

Reliability, which is a necessary forerunner of validity, is more easily established (Belasco and Trice, 1969) by simply applying the criterion to a group of potential trainees (usually before training) and then repeating it within a reasonable lapse of time (again, before the training).

The validity problem in criterion development must be dealt with in nearly every training evaluation but must be resolved in accord with the particular circumstances. A major aspect of validity is the problem of inference. How well do criteria which measure changes in knowledge or attitudes actually predict what will occur on the job? While an obviously simple resolution to this problem is to measure job performance directly, this is oftentimes impossible. Trainers and evaluators may not have access to performance situations because such access is disruptive to the work flow. More important is that job performance cannot always be measured, particularly when dealing with whitecollar or semi-professional work, when dealing with highly specialized technical work, or when dealing with "intangibles" such as dimensions of supervision. While performance measures are the ideal, they can usually be approximated at best. The evaluator should endeavor to cope with this basic validity problem in terms of minimizing the "distance" between the actual and ideal criteria. One example of such reduction is through the use of *projected* job behaviors as a criterion, i.e. gauging the trainees' reactions to hypothetical situations which represent both the real world of their jobs as well as the training goals. Care must be taken so that such criteria do not have obviously "correct" answers.

Follow-up Measurements

While many evaluative activities occur before and during the training activity, the major "payoff" lies in post-training measurement. Two major problems are involved in considering follow-up: timing and legitimate data comparisons.

While evaluation activities are often summarized under the question, "Did it do any good?", the accompanying query usually is, "For how long?" Timing follow-up can be a cause of major consternation. The longer effectiveness can be shown to be sustained, the more likely the training will be strongly supported. On the other hand, evidence of the "wash-out" of results after a period of time may undermine and discount earlier evidence of effectiveness.

For a variety of reasons, including research and statistical effects, most measurements of effectiveness tend to drop with increasing time-distance from training. Measuring such patterns of decline, which is infrequently done, can provide valuable information on the appropriate timing of refresher courses or other efforts to reinforce training experiences. If anything, training evaluation is usually conducted too soon after the completion of the training, often due to trainers' anxieties about "looking good." Immediate follow-up may tap "halo effects" of the training experience. Furthermore, immediate follow-up does not provide the trainees adequate opportunity to encounter the situations and demands where they might "try out" their newly acquired knowledge or skill. Thus, the timing of follow-up should be realistically gauged to the training goals, and above all, should be a matter for careful consideration. Longer-term follow-ups may be considered in the light of training resources invested in a particular effort, as well as in light of the resources available for refresher activities should data indicate such a need.

Evaluators typically use the goal-attainment approach in measuring post-training criteria. For example, vocational program evaluators frequently use length of employment and amount of earnings as criteria. These follow-up data are of three types: (1) those with no "before-training" indexes; (2) those with "before-training" information using the same measures as the follow-up; and (3) those with both "before" data and data collected from a comparison group, using the same measurements.

Obviously, evaluators using follow-up with no baseline criteria information for comparison have no basis at all for conclusions about the training's effectiveness. Those with baseline "before" data are in a better evaluative position; most of the published reports employ this strategy (Solomon, 1969; Goff, 1968). Without comparison groups, however, such evaluation has no way to assess the ever-changing, complex factors that may alter the trainees' behavior in the *real* world. A simple example is the effect of labor market conditions on Manpower Development and Training Act trainees' job opportunities. If the job market is poor, this will be reflected in criterion focused on post-training employment. These forces are obviously beyond the control of the training program, but with measures of a comparison group's employments experiences, conclusions about the training's effectiveness are more soundly based. It is important to note that follow-ups using comparison groups, whether secured by matching or sampling, face the problems of legitimate comparisons previously described (Greenberg, 1968; Smith and Honour, 1969). For example, in studying the effectiveness of vocational training, school dropouts or academic course students are scarcely acceptable comparisons from which to reach conclusions about training effectiveness (Somers, 1971).

Follow-ups face some other common problems. If mail questionnaires are used, the response rate may be low, calling for efforts to stimulate non-respondents. Even with efforts to increase response, respondents who return the questionnaire may not be representative of the trainee population. While samples of trainees produce the same problem, sampling makes such follow-ups more manageable. All follow-ups have the "detective problem," i.e. how to efficiently locate trainees after the training for the administration of measurements (Belasco and Trice, 1969). Many follow-ups have floundered because evaluators did not carefully plan and test out a way for locating former trainees.

Follow-ups may be plagued by ever-present research effects, although these may be difficult to detect. By interviewing or observing trainees, by sending them scales, or with any other research method, the evaluator unwittingly may affect criterion measurement. Probably more frequent is the inflation of measured effectiveness due to typical attempts by trainees to give socially desirable responses. On the other hand, repeated contacts by evaluators may increase negative reactions to the training. Research effects are increased when two or three contacts, involving several different instruments, take place during the follow-up. It is nearly impossible to adjust for research effects in data analysis.

Fitting Method and Approach

At this point it may be valuable to return to two basic sets of categories as a means of summarizing our outline of methods and techniques in training evaluation. These techniques comprise two types: subjective and experimental. How do these techniques fit the six approaches to evaluation which we previously delineated?

The goal-attainment approach can employ either experimental or subjective methods. Experiments are ideal since the approach assumes training goals to be firmly established. Research interviews and questionnaires can be aimed at getting trainees to describe the extent to which they have changed in the direction of the fixed goals of the training. Where goals are considered largely in terms of measuring effort, records can be appropriately used. Generally speaking, the goal attainment approach can be implemented through using all of the techniques described.

The cost-benefit approach is an "open and shut" experimental situation with heavy emphasis on statistical indexes and comparisons. Since it has been primarily used with manpower training programs, the approach has not perfected the use of experimental designs. Control groups are, however, being introduced and the problems of simple follow-ups when attached to the model have been clearly discussed (Somers, 1971). Statistics play a central role in the approach since it is based largely on dollar estimates, indexes, and ratios.

Both subjective and experimental techniques fit the ceremonial approach. The focus of this approach is how the trainees see themselves, and on the reduction of their anxieties about new statuses. As in the goal attainment approach, changes in these feelings and sentiments can be measured to some degree via questionnaires, preceded by research interviews to establish parameters. Subjective methods rather than indices of job performance permit more flexibility to register subtle changes in self-concept and anxiety.

Records are of central importance for the organizational support approach. This approach calls for a high degree of direct contact with the training program for a sizeable period of time. Evaluators think in terms of how resources are distributed, how interacting units perceive the training program, and how the program fits trainees' needs. Experiments are largely inappropriate for this approach while records, observation, and subjective methods fit nicely. Whyte (1971) believes that unless substantial participative observation is introduced into evaluation efforts, the process and dynamics of a program's efforts cannot be accurately assessed. Research interviews complement observations and records, and are particularly relevant to get at the amount of consensus on operative goals, i.e. those on which day-to-day decisions about training resources are made.

Beyond doubt, the outside validation approach depends almost entirely on observation, records, and subjective techniques, especially the research interview. In many instances the outsider may have a structured observation check list and guide against which he judges various dimensions of a training effort. Thus the Los Angeles City Schools (1971) have a handbook for evaluating instruction. Good records often afford the outside specialist the only opportunity he has for a time perspective on the training effort. They are, so to speak, a platform from which he can more accurately observe and talk with trainers and trainees.

Much the same can be said for the native methods approach. Since it focuses on trainers' own ways to "keep score" about the success of their efforts, the evaluator must get as accurate an account of these native evaluative devices as possible. Questionnaires would probably miss these native techniques. Structured instruments could not accurately disclose what kinds of formal evaluation devices would, in the trainer's eyes, fit with what he already does. Research interviewing alone would be hampered because many times the native evaluative devices are partly unconscious. Thus records and skilled observation are of primary importance in this approach.

Toward Improving Subjective Methods

Despite the elegance and power of experiments, they are often impossible to use. Repeatedly, training directors and specialists have reluctantly abandoned the experiment or have adapted it for more practical use. This leads to the consideration of how subjective methods which admittedly are weaker, can be strengthened. Illustrative is Reeves and Jensen's (1972) suggestion that *participants'* evaluation of adult education programs could be an effective tool for refining future programs.

Using research interviews, observation and questionnaires in conjunction with one another is a major strategy for strengthening these methods. "Far from being competitive, surveys and interviewing and observation are actually *complementary* methods: the strengths of one compensate for the weaknesses of the other" (Whyte, 1969). This principle first calls for action-oriented trainers to *take the time* necessary to use several different subjective methods. An example lies in the preparation of questionnaires. As described earlier, these instruments emerge too often from trainers sitting down in their offices and devising items which they, not trainees, believe are relevant. To make matters worse, these questionnaire-makers often limit the respondent's range for expressing their reactions to the training by giving them forced-choice "closed" responses, thus running the risk of missing trainees' true sentiments or behaviors. Observation and the research interview provides means for generating the "raw materials" out of which to fashion a questionnaire. More important, research interviewing may provide new materials that the evaluator does not think of, improving the questionnaire. In an effort to evaluate organizational training, Friedlander (1967) took detailed notes during interviews and recorded verbatim comments made by group members which were rephrased into items for a questionnaire.

Statistical findings from questionnaires must be interpreted, and research interviews at this point in the process may provide a means for more creative and comprehensive analyses (Belasco and Trice, 1969). Careful observation can likewise expand the trends detected in research interviewing. Observation provides a tangible reality dimension for the results of research interviewing. Observational data may provide even better "raw materials" for questionnaires than research interviews alone.

Subjective methods can be improved by sampling. An adequate and straightforward sampling approach begins with a complete alphabeticized list of trainees to be evaluated. By simply counting off along the natural ordering of the alphabetically arranged list and taking every nth name after starting at a random point in the list, a "systematic sample" is produced. One study indicated this procedure produced the same results as a stratified random sample in using weighted application blanks for hiring purposes (Trice and Penfield, 1961). If a systematic sample is taken from a *clearly defined population* and subjective methods used on all persons in that sample, the evaluator knows far more than if he relies on intuitive assumptions that the trainees on whom he has measurements constitute a representative sample. Such sampling, if it is sustained, reduces the dollar costs of evaluation efforts appreciably since no more than 30 or 40 percent of the trainees, at the most, must be contacted for data collection.

Another improvement of subjective methods is their complementary but separate use on the *same* evaluative question, a strategy called in technical terms "multiple triangulation." For example, an evaluator develops a questionnaire asking a sample of supervisory trainees if and how they adopted specific techniques of grievance handling. Completely independent of this data collection he uses research interviews to explore the same basic points on a smaller (but different) sample of the same population of trainees. Whether or not the conclusions of these two approaches coincide, subjective methods have been improved by using one to check on the other and by unavoidably generating a greater breadth of information.

Would these two methods produce discrepant results, a third method, observation, could be employed to resolve the differences. One large-scale example of this combination of methods is the Ritzer and Trice study of the professionalism of personnel managers (1969).

A powerful argument against rigorous experiment designs and for the integration of various subjective approaches to evaluation comes from Argyris (1968a, 1968b). Taking sharp exception to the experimental design *only* position of Dunnette and Campbell (1968) regarding laboratory training, Argyris argues for the use of various subjective methods and their combination as a more viable, less offensive way to conduct training evaluation. He is especially insistent that "control" groups are highly unrealistic, producing harmful unintended consequences and failing to perform their alleged experimental function.

An additional way to improve subjective techniques would be a concerted effort to focus them on *behavior* change rather than, for example, on knowledge or attitudinal change. Attitude change may have practically no relationship to behavior change; consequently it is important to concentrate as much as possible on actual, desirable, on-the-job behaviors that result from training. A pertinent example of this emphasis is Kirkpatrick's (1969) interview study aimed at measuring such behavior changes. Interviewing both participants and their immediate supervisors two to three months following a training institute on "Developing Supervisory Skills," he concentrated on how specific job behaviors differed *after* the training in contrast to before the institute. These job behaviors were exhaustive, including interview data on changes that ranged from order-giving to employee turnover. Numerous desirable changes were found: for example, "probably the most significant behavior change appears to have occurred in the participant's success in satisfying complaints before they become formal grievances" (Kirkpatrick, 1969:35). The results indicated that the training was, beyond doubt, resulting in favorable changes in on-the-job behavior, although the trainees tended to indicate more positive changes than did their supervisors.

Turning to other aspects of questionnaire improvement, validity is clearly the most important aspect. Validity will be enhanced by pretesting questionnaires in order to get a fix on their validity. Pretesting calls for the training evaluator to resist the temptation to respond to pressures to do a "quickie" questionnaire which he "feels" is a valid criterion.

First the questionnaire developed from observation and research interviewing is submitted to a small but representative group of trainees, asking them to locate confusing wording, and directions, and to suggest how the instrument can be improved.

Next, the pretest includes efforts to validate the questionnaire against outside criteria. This calls for responses to be checked by some independent source of the same information to see if there is congruence. Thus it a questionnaire indicates supervisors have reduced the "halo" effect in performance appraisals, i.e. typically rating high-status subordinates high and low ones low, a pretest check of actual behavior can determine if this took place. Another example is a comparison of reported adoptions of managerial techniques such as clearer budgetary planning gained from "management by objectives" training checked against records in accounting offices. A third example: respondent-managers may report they participate differently (such as interview, study records, and seek job analysis information from the personnel office) in the hiring process following participation in "management games." Such reported behaviors could be checked, at least in part, against personnel department records.

The "behavioral" theme in validation could be carried a step further to the use of direct observation of the trainees' post-training behavior as an evaluative criterion. In communication training, trainees can be rated by observers on the accuracy with which they communicate complex bits of information to a third party. Similarly, those who have experienced supervisory training could be asked to role play a typical problem situation with subordinates while observers assess the extent to which specific supervisory concepts were actually used.

There are several other kinds of validity in addition to external validation. "Face validity" has already been described. "Predictive validity" refers to the extent a questionnaire distinguishes between trainees who will change in some desirable direction compared to those who will show little change. Thus a training evaluation questionnaire's predictive validity would consist of correlating its behavioral change scores with actual behavioral changes, as in the example discussed above. When the questionnaire comes to consistently identify responders and nonresponders to the training, it has predictive validity. When large investments are to be made in a training effort, pretesting during a pilot period may establish this type of validity.

An imaginative approach to the problem of "construct validity" of a questionnaire has been reported by Reeves and Jensen (1972). Focusing upon participant evaluation of management training programs at the University of Wisconsin, they assumed that there should be (1) comparable subjective evaluations of identical programs by separate groups of participants with similar training needs, (2) participants' subjective evaluations would be consistent over time; and (3) trainers' evaluations of their own programs would be compatible with those of the participants. The results showed considerable consistency across these three categories. Thus construct validity (Sellitz, *et al.*, 1959) raises the question of whether questionnaire results relate to other relevant factors in a fashion one would reasonably expect.

Questionnaires that attempt to measure behavior are more easily validated than questionnaires dealing with beliefs, attitudes, and perceptions which are less amenable to validity checks. As Whyte (1963:10) points out, such data "remain within the subjective world of informants and do not allow us to break out and connect the subjective with the objective." Questionnaires can only tell us feelings the respondents think that they have. If such data are collected, for example, in an evaluation of human relations training for supervisors, their validity might be checked against such "hard" post-training data as work-group with figures such as absenteeism, turnover, and productivity.

Despite the numerous efforts that can be made to establish and improve the validity of evaluative criteria, a major (and usually unavoidable) source of validity problems is the respondent group. Evaluators might as well accept the fact that respondents commit a sizeable amount of both intentional and unintentional errors. For example, Bell and Buchanan (1966) discovered that 30 percent of respondents in a general population gave inaccurate replies to a question about voting. Cannell and Fowler found that 10 percent of their respondents had inaccurately reported whether they had or not had surgery. Apparently such errors are an attempt to make reported behavior compatible with personal and group norms (Clark and Tifft, 1966). Consequently trainees may be motivated to exaggerate reports of desirable behaviors resulting from the training because these are perceived to be consistent with expectations of trainers and of the organization. In sum, we must expect some invalidity through both error and lying, but should try to learn how much of it there is.

Reliability, which is a necessary first step toward criterion validity, is all too often forgotten. It can be estimated in the pretest by administering the questionnaire a second time to the same sample of trainees some three or four weeks later. Various forms of correlations between the items administered at the two points in time can then be used to determine if respondents are consistently responding to the items. Correlations of .85 or above suggest reliability although even this much disagreement can damage interpretations of questionnaire results. Sampling fluctuations may contaminate correlations to an unknown degree (Lord, 1970), sometimes artificially inflating reliability. One's evaluative information will, however, be more justifiable if simple reliability indexes are computed during pretest.

A scale is a gauge for arranging responses in order to assign numerical distinctions of degree to them. Scales usually comprise the "meat" of criterion questionnaires. The best most evaluators can expect is the use of crude ordinal scales such as the previously described Likert-type (Likert, 1932). These scales permit ranking of responses or descriptions of behavior in some clearly discernable order that may, or may not, be assigned numerical values. An example of ordinal scale in organizational life is foreman, supervisor, director, manager, vice president, and executive vice president. In terms of a questionnaire, items can be devised so that trainees' responses could be arranged on an ordinal scale that reflects degrees of behavioral change they attribute to the training.

Although ordinal scales are an improvement over a simple list of different responses (sometimes called "nominal scales"), they do not permit a standard unit of measurement because there is not a standard *interval* between points on the scale such as would come from information on income level. Thus, while interval scales are difficult to develop, some "naturals" may be present in questionnaire material. Age is the classic example. In supervisory training, the actual number of subordinates, salaries and salary changes, or the total dollar budgets are examples of interval items. The evaluator should be alert to these true intervals that may permit a classification of subjective responses into a more precise description. The evaluator should not fear a sizeable number of categories in the belief that simplicity and 2 x 2 statistical tables are the only basis for analysis (Blalock, 1964).

It is often helpful during a pretest to set up "dummy" tables for how the results will be analyzed, using pretest results as a "dry run." The evaluator can prepare in advance how he will use the results of the actual administration of the questionnaire and drop items which do not appear to be useful. Furthermore, these mock-ups raise questions regarding how additional scales may be incorporated into the questionnaire, or how those already there might be improved.

A practical issue involved in many follow-up studies using questionnaires is how response rates can be increased. Champion and Sear (1969) suggest use of personalized; hand-stamped letters; they found special delivery helped materially; and that "longer" questionnaires do not seem to stymie returns. In terms of different training groups, these researchers found greater appeal to lower-class respondents if the covering letter emphasized the benefits the respondent would receive by responding while a similar, but muted, point was included in those follow-ups sent to upper-middle and upper-class respondents; a strong appeal to help the evaluative effort achieve its goals appeared to increase response in these groups. In sum, follow-up appeal to low status respondents was based on egotism, while appeal to high status persons was based on altruism.

Until now we have assumed that the fixed-choice questionnaire is more likely to be the best, with items and response choices developed through the use of research interviewing. Numerous reasons justify the closed strategy: it insures that answers relate to the specific evaluative focus; it forces the respondent to make a judgment about the training experience; it avoids the necessity of having responses on an "open" instrument interpreted, classified and coded; it facilitates responses from those who are not articulate; and it reduces differences between reticent and verbose respondents. But there may be situations where evaluators should consider an open type questionnaire. If research interviewing reveals that respondents lack sufficient information, are uncertain about feelings and attitudes, or are widely variable in their reactions to a training experience, an "open" approach may be indicated. This may provide depth information on the *process* through which to arrive at particular subjective feelings about the training experience. Such a questionnaire could be made up of questions like "Describe the differing opinions you have had about this training course"; "As the training progressed tell us about your emotional feelings"; "What aspects of the training experience were particularly meaningful to you?" "How had you defined the situation in your own mind before the training?" If training issues are complex, the open type may be particularly appropriate with evaluation emerging from a detailed content analysis to form judgmental ideas about effectiveness.

There are several ways in which research interviews might be improved in order to supply the data for the fixed-choice questionnaire. Evaluators can try to keep the interview constant from respondent to respondent by intensive training designed to standardize interviewers. Role playing between interviewers and video-taping serves to familiarize interviewers and provides for reduction of variations in questioning strategy. It is helpful to have only *one or two* interviewers do all the interviewing. Where a large number of interviews are needed, Friedman (1967) suggests a different tactic: if there are a number of potential interviewers, they can be representatively sampled for interview participation in order to randomize the interviewer's influence on respondents' replies. Regardless of the number of interviews, it is good policy to have enough interviewers to complete the work in the shortest time possible.

In terms of the conduct of the interview, Whyte suggests that indirect types of questions may help when people do not respond to a direct one. Thus, "How do you feel about A?" can be altered to "How do you feel about A, compared to B,C,D, etc.?" (Whyte, 1959). Furthermore, he found that the .se of cards with questions printed on them drew the "attention of the respondent away from the interpersonal situation with the interviewer and pointed it toward his experiences and sentiments" (Whyte, 1959:23). Becker (1954) suggests it may be helpful for the interviewer to respond to abstract statements with mild skepticism and ask for specific examples. This forces respondents to come forth with more personal, subjective feelings in place of vague abstractions. He also reports "playing dumb," thereby eliciting more detail than if he immediately accepted statements as if he fully understood them.

Another suggestion is the "tandem interview." Kincaid and Bright (1957:309) report that two interviewers "can more effectively explore an unchartered field than a single interviewer" as well as produce "gains in validity and reliability." The interviewers perform as colleagues, but develop a pattern for questioning such that "recording is efficiently done without struggling to keep pace with the respondent, thus freeing the interviews of breaks in continuity." This further suggests the possibility of a tandem research interview on small *groups* of trainees. Chandler (1956:27) reports close correspondence between materials collected in group and individual interviews, but indicated that "the group opinions that could be distilled were not a completely accurate reflection of the private feelings," with the conclusion that interviewing small groups could be "a valuable supplement to individual interviewing."

Recording the interview is an ever-present problem. While the goal is full recording, this is obviously impossible without a tape recorder. While some report they have been .ole to introduce tape recorders without threat, the majority opinion seems to be that they stymie responses – especially among those tho are not articulate. Without a recorder the ideal is to memorize responses and record them later. Accurate memorization may, however, be a difficult skill to acquire, and dependence upon memory is inevitably risky. The development of a rapid note-taking ability, even shorthand can surmount this problem. Such recording can be made either during or immediately after the interview. Rapid writing during the interchange may help to demonstrate confidentiality if the interviewee can readily see that the notes are largely unintelligible. But such note-taking can disrupt eye contact, questioning patterns, the respondent's perception that the interviewer is listening. One solution is "taking notes with a partial shorthand during the interview in order to get on paper a summary of main points and key verbatim sentences . . . These notes become the basis for a longer written account to be made as soon as possible after the interview" (Belasco and Trice, 1969:30). At this point a tape recorder can also be used so that important points are captured while they are still fresh.

Unfortunately evaluators usually have less opportunity to use criteria based on direct observation than to use other subjective devices. The most direct improvement of observation is the use of two observers if at all possible. Observation suffers primarily from inaccuracies and from the conscious and unconscious bias of the observer. If two observers of different backgrounds carefully observe the same behavior, they can compare and check their recorded materials. Beforehand they can develop recording methods that differentiate actual events from the interpretation of events. Observers usually assume they will remember more than they do; thus a training effort should be directed toward getting observers to separate actual observations from interpretation in their recordings. Furthermore, materials can be coded relative to the degree of accuracy of recorded observation: exact recall, reasonable recall, or only an approximation of what actually occurred.

IMPROVING ADAPTATIONS OF THE EXPERIMENTAL METHOD

After-only Strategies

One of the major problems with the experimental method lies in the fact that typically it calls for "before" measures of some kind. These measures tend to have a very potent effect of their own, namely the "research effect" described earlier. Administration of these measures can often produce as much change as the training itself (Trice and Belasco, 1969). The experimenter himself, as well as the instruments he may use in an experiment, may generate self-fulfilling prophecies that produce changes in trainees which may naively be attributed to the training (Rosenthal, 1966). For example, Anderson and Anderson (1971) found that behavioral change was accomplished by the use of questionnaires that called for participants to rate management and each other.

The impact of this flaw can be reduced if a design can be constructed that avoids collecting "before" information from trainees. While "before training" data are necessary to establish a starting point from which to compute change, direct collection (and its contaminating effects) may be avoided if enough training candidates are available. Two steps can be taken: first, the use of a large number of potential trainees and, second, use of numerous strata for sampling them so that representiveness is assured. With a large sample (200 or more) and *random* division is made of the sample into two groups, neither of which gets "before" measures. One group receives training while the other does not and is used as a comparison group. The probability of drawing comparable groups increases greatly if the sample is stratified on a maximal number of those variables which are believed to lead to differential attitudes, knowledges, skill, and behaviors. In addition to demographic variables, stratification can hopefully include personality traits. If these steps are taken, the training and comparison groups should have nearly identical "starting points," eliminating the necessity of "before" measurement. If this strategy is followed as an alternative "before" measurement, a simple two-group comparison contrasting the "after" scores of trainees with the "after" scores of the non-trainees will constitute a legitimate measure of change. Incidentally, these non-trainees can receive the training after the evaluation study so that both evaluative and training goals are met. It bears repeating that this adaptation of classic experimental design relies heavily on *randomized placement* of the population into a trainee and a comparison group.

Legitimately Compare Several Types of Training

Rather than attempting to generate a control group which does not receive the training, it is often more practical and sometimes more revealing to compare *types* of training, assuming such comparisons can be done legitimately. Ideally this can be accomplished by randomizing trainees into one of the several types of training, which should be as different in content and strategy as feasible, but whose objectives are as similar as possible. Randomization, as described previously, will assure common starting points, using common "after" criteria for all of the training. In situations where fewer trainees are available, matching of trainees in the several different types can also be attempted to assure common starting points on the evaluative criteria. Such matching can be done via "twins" in which a trainee from each type is matched with one from the other. Unfortunately matching requires large numbers of trainees before individuals can be found that match each other as closely as possible.

Current literature about training evaluation abounds in examples of this approach, suggesting it may be more feasible than the traditional experimental design. Often, however, these studies do not appear to meet minimal requirements for legitimate comparison since neither randomization nor matching assured common starting points. A "univex net" which transmits audio and visual signal via telephone lines from one campus classroom to another was compared with auto-tutorial carrel units which were set up for independent study (Everly, 1970). In a study which compared programmed instruction with older, more traditional methods for retail staff, it was concluded that programmed training seems to have its greatest application in fields where the subject matter is clearcut and where trainees are required to learn in a routine way (Pickett, 1970).

Another study contrasted computer-assisted instruction about basic electronics with typical classroom lecture and demonstrations, with achievement and time scores as common criteria, and the computer assisted group showing greater improvement on the post-test measures (Ford and Dewey, 1970). The question of T-group training effectiveness has been approached in this fashion. Heck (1968) compared the effectiveness of T-groups and the more traditional program of the Human Development Institute in changing interpersonal perception styles and communication skills. These skills improved in both types of training but neither altered interpersonal patterns. A comparison of a lecture-discussion approach to interpersonal relations in organizations with T-group training produced roughly the same change in beliefs about interpersonal behavior, but the T-group experience produced greater changes in the trainees' perceptions about themselves (Bolmen, 1971). Arnoff and Litwin (1971) gave executives a program designed to strengthen their need for achievement and matched them with executives chosen to attend the corporation's executive development course. Motivation training produced significantly better results in both job level and salary achievements.

Although these examples deal with comparison of only two techniques of training, an experimental design, using a Latin Square notation, could compare three or four techniques. Thus one group participates in business games, another receives lectures, while a third engages in role-playing. These techniques might be directed toward a series of specific objectives such as improved communications, employee-oriented supervisory styles, and knowledge of union-management relations. Each group would participate in three training experiences, with one of the three training techniques used for each of the three objectives. Scores of the nine resultant exposures could then be compared on a common criterion.

Comparing Randomized with Non-Random Groups

Although the classic experimental design calls for random placement of trainees into trainee and control groups, this is rarely accomplished. Practically speaking, groups that have not been selected on a random basis are often the only ones available for comparison even though they theoretically introduce bias and make comparisons less legitimate. Borus and Buntz (1972), writing primarily about evaluating manpower training programs, implicitly raise the question: Are non-random control groups as contaminating as believed? There seems to be little actual empirical evidence for such contamination. "To determine the relative merits of the various types of control groups, an empirical investigation should be conducted in which data from each type of control group (random and non-random) and their labor market experiences compared" (Borus and Buntz, 1971:328). Training managers wishing to improve training evaluation procedures and at the same time create a potentially large saving of resources, could make a notable contribution by mounting such a study. They might discover we can reasonably operate with non-random groups with less damage than previously believed. Perhaps a weighting factor formula might be devised whereby results could be corrected by the degree of contamination typically present in non-random control groups.

Predicting Effective Training Results

Although predictions of desirable training results without comparison groups do not answer the basic question whether changes were due to the training, they do use criteria against which predictor items can be assessed; thus, such efforts are at least marginally experimental. For example, turnover and absenteeism, both during training and after job placement, are criteria against which predictor items of training effectiveness can be judged. Subsequently, other training approaches can be devised and evaluated for those who do not respond. Numerous studies of the hard-core unemployed have focused on predictor items that could serve to identify those hard-core persons who respond to traditional training, setting the stage for systematic study of those who do not. As might be expected the younger, unmarried hard-core trainee who is relatively free of family responsibilities is more likely to drop out of training and show job turnover (Quinn, *et al.*, 1970; Gurin, 1968; Hodgson and Brenner. 1968; Rosen, 1969). These, and the numerous other predictive variables that have been isolated in this type of training might be effectively used to develop different types of training as well as more effective selection for training (Purcell and Cavanagh, 1969; Shlensky, 1970; Greenberg, 1968; Allerhand, *et al.*, 1970; Teahan, 1969).

Use of Subjective Methods

Finally the principle of complementary research techniques should be explored. Experimental results can be broadened and better understood by research interviews, questionnaires and observation. Whyte (1969:47) puts it succinctly: "I have been arguing that before-after measures of the effects of a given governmental program are not good enough. We need to know what went on within the program that may be presumed to account for the differences . . . who is to provide such data? Someone who is out in the field observing what is going on, perhaps even a participant observer."

RECENT TRENDS IN TRAINING EVALUATION

Traditional concerns for evaluating craft and technical training remain prominent among training specialists, despite trends toward a wider variety of training methods and new evaluation approaches. For example, Cook (1971) discusses the advantages of full-time training courses for building craft apprentices. He assesses the cost of effectiveness of aptitude tests in such training and presents a comparison of costs and results between day release and full time training. A summary of 22 research reports assessing the technical proficiency of U.S. Navy aeronautical support personnel describes a "matrix method" for evaluating training (Siegel, 1967). Kayloe (1971) uses this method to evaluate how technical training prepares a person to perform defined tasks after graduation from a training program, with estimates of "suitability for the job" as the basis for training evaluation. Tasks were sorted into a matrix in which the two divisions represented technician proficiency and task importance. Task frequency entries produced a Training Index, an Overtraining Index, and an Undertraining Index. Comparison of a traditional and a sharp revision of a radio operator training course provides another example of continued concern for evaluation of technical training (Goffard, 1970). Using the same method, Greenberg (1970) compared graduates of on-the-job training technical skills programs conducted by business firms themselves with Neighborhood Training Centers trainees. Over the past decade, hundreds of training programs for the hard-core unemployed have emerged. Many of these efforts have reported their evaluations which frequently turn out to be a predictive strategy in which the characteristics of successful trainees are described. A review of some of these results provides a flavor of this literature.

Quinn *et al.* (1970) used an experimental design and contrasted employees with pre-job training that was company-oriented (rather than skill training) with those not trained but hired directly. Although trained individuals were more likely to value work and show positive attitudes toward time schedules, their job skills were not any greater. This study also produced a caution against optimistic hopes for training programs for the disadvantaged. Trained individuals were found to want more autonomy than they probably will get and they viewed supervision more unfavorably than did those who did not receive the training; in short, the training tended to lead to greater expectations than job circumstances provided.

Rosen (1969) studied training for hard-core unemployed that focused on changing the trainee's attitudes about himself and his relationships with others in contrast to training that focused on the company and job adjustment, using matched groups. Those in the company-oriented sessions showed less turnover than did those in the attitude-change sessions. But when this study went a step further and compared turnover with normal hires, it was revealed that company orientation training was not necessarily more effective than no training at all.

Much of this program evaluation is at a rudimentary level, but nonetheless shows a clear trend toward accepting the need for accountability. The employment history and present job status of educable mentally retarded students who terminated their schooling in New York City from 1960-1963 was studied by Tobias (1960). Despite free resources for training and placement available through the Division of Vocational Rehabilitation, less than 40 percent of the interviewed sample and less than 20 percent of the total population were known to the D.V.R.

What appears to be basically an outside specialist approach characterizes numerous evaluations. Rowan and Northrup (1972) examined the impact of adult basic education programs on the upward mobility in the paper industry of disadvantaged workers (especially Black) initially hired for unskilled jobs. Observation and testing showed that few completed the courses and those who did showed little improvement in arithmetic and reading. Programmed teaching methods were shown to be largely unsuccessful. An analysis of six manpower development and training programs in five cities directed toward unskilled slum residents recommends specific techniques for trainee recruitment and selection, staffing, and job placement (Nellum, 1969). Elliott (1972) urges adult educators to evaluate their own evaluation efforts, providing a sequence of features of a program for an evaluator to assess in the content analysis of a training effort for disadvantaged families.

Simple follow-ups without control or comparison groups is often found in attempts to assess the value of training among the disadvantaged. Frequently these are cost-benefit oriented (Conley, 1969; Wood and Campbell, 1970), but they may tend to overstate the effectiveness of the training because even without the programs disadvantaged groups may have had greater-than-average improvements in their work situations during this time period. Similarly, the follow-ups of a Manpower Training and Development Act program (Prescott and associates, 1971), of Training Incentive Payment Programs (Institute of Public Administration, 1971), and of Job Opportunities in the Business Sector (Greenleigh, 1970) suggest the beginnings of evaluative efforts even though they suffer from the flaws of most follow-ups.

An evaluative trend can be discerned here, namely that assessment of M.D.T.A. programs has taken a cost-benefit form, usually without control or comparison groups. A South Carolina study (University of South Carolina, 1968) projected estimated lifetime benefits of these programs against estimated monetary training costs, concluding that benefits greatly exceeded costs. Mangum (1967) makes estimates of overall costs of both quantifiable and non-quantifiable accomplishments of M.D.T.A. programs, compares these with the overall benefit contributions, and concludes that the programs should be expanded. Young (1970) focuses only on the cost dimension of M.D.T.A. programs, hoping to provide training decision makers with a thorough range of training costs involved. His list of cost dimensions provides evaluators with an exhaustive list of both direct and indirect dollar costs against which to compare estimated benefits. A U.S. Department of Labor report (1967) concludes that M.D.T.A. training programs result in a general upward shift in overall hourly earnings despite variations in different industries. Sewell (1971) criticizes many such training efforts for selective admission, thereby "stacking the cards" in their favor. When, however, he used a cost-benefit analysis on a manpower training program that did not select on the basis of aptitude or intelligence, he found a higher benefit cost ratio for on-the-job training than for institutional training.

Efforts to evaluate the retraining of the unemployed is another trend. Somers' 1967 study is illustrative. He describes a 1962 interview survey of employees in West Virginia who had hired trainees under the Area Redevelopment Act, a 1964 nationwide questionnaire survey of members of the American Society of Training Directors, and a questionnaire mailed to 1,000 employers in Wisconsin about the apprenticeship pattern of on-the-job training. These represent an awareness of the need for evaluation but must be viewed as embryonic at best. Borus (1966) computed cost-benefit ratios (for individual trainees, for the government, and for the economy) for trainees and three non-random control groups. Hardin and Borus (1969) used cost-benefit ratios in a predictive pattern. Another evaluative effort (Somers, 1968) used three non-random control groups (rejected applicants, trainees who dropped out before completing the course, and unemployed workers who did not apply) with the evaluative criterion of success in placing workers in useful employment. Solie (1968) used practically the same kinds of control groups, concluding from the comparisons that retraining programs do improve job prospects for the unemployed, but these benefits probably vary directly with changes in the general level of employment.

Although much less sophisticated, some studies of women in the labor market contain evaluative themes. Orth (1971) carefully examines the extent to which women have penetrated managerial ranks and concludes that even though the long term outlook for the next decade is for a shortage of male managers, male attitudes toward women at the professional and managerial level continue to block change. This suggests the absence of a support system for training of female managers, i.e. they simply will not be accepted regardless of their training. This occurs despite recent facts about labor costs (absenteeism, turnover, tenure, and mobility) of the female manager which show no cost differential between men and women in terms of their contribution to the work effort (Wells, 1969).

There has been some attention to a potential training audience among senior citizens. Although not formal evaluation in the strict sense, a Chicago study established that there is, indeed, an elderly audience that is interested in quality continuing education (Sanfield, 1971). Supporting this finding is a Massachusetts study showing that it is feasible, in a case-study orientation, to redirect and reactivate older workers toward employment (John F. Kennedy Center, 1969). These studies, while only forerunners of actual evaluation, suggest that training for women and older employees is both desirable and practical.

As occupational obsolescence becomes more and more a fact of work life, continuing professional education has emerged as a way to deal with it. Robertson and Dohner (1970) have described the need for lifelong learning for physicians and offer their opinions about the effectiveness of recent continuing education programs to combat obsolescence. One study attempted to devise evaluation criteria for continuing training by observing physicians' practice and then suggesting educational programs to meet the needs revealed, although actual evaluation did not take place (Meyer, 1970). Concern reached the point where an entire medical conference devoted itself to the possibility of evaluating continuing medical education (Rising, 1970), surveying methods and approaches and emphasizing the need for more systematic selection of strategies. Continuing training of state and federal judges, and its evaluation by outside experts, is another example of the trend. In a general vein, adult educators dealt with the criterion problem at a recent conference dealing with gauging the adaptability of a profession (Nattress, 1969).

The Threat of Evaluation to Trainers

In conclusion, we return to our earlier discussion of resistances to evaluations. Even though most evaluators are aware of the threat evaluation poses to training people, they probably are not aware of some of the reasons for the concerns they encounter. Much of the trainer's resistance comes from the explicit intent of evaluation to find out if training is effective. Evaluation is usually an effort to "find out what's wrong with us." Anxieties mount when trainers discover that evaluation often produces negative results that mav put them in a bad light, and quite naturally these feelings bring out defensive measures to protect against unfavorable findings.

Furthermore, evaluative studies aim at generating information that will probably be used to change the training program. Since the changes that may come from evaluation are largely unknown, the typical anxieties associated with change are aggravated. One careful study of professional employees and supervisors in two large hospitals concluded that "only a minority expressed an attitude favorable to taking the risks which are believed to exist when information with evaluative implications is widely disseminated" (Eaton, 1962:421).

Training people often feel that a more balanced picture could be developed if they could participate in the evaluation, even design some of the methods used and define the outcomes to be measured. Trainers justifiably tend to believe that everyday experience and practical judgments are far more realistic than results generated from "scientific" methods. They feel evaluation will find little that they did not know before.

Trainers are more apt than evaluators to be concerned with immediate, specific use of knowledge while evaluators think more in long range, problem-solving terms. "Action" people in general and trainers in particular, are prone to commit themselves to evaluation strategies without realizing their full importance. When they do discover the evaluative study's meaning and implications, they may sharply reverse their acceptance.

As a result training staff may devise ways to counteract these risks and threats. Negative results can be interpreted as evidence that the program just doesn't have enough resources; or that changes occurred that were not measured. Other rationalizations are plentiful and may be more than excuses: (1) training effects are long range and cannot be gauged immediately; (2) measuring instruments are too crude and cannot pick up important, but subtle, effects; (3) withholding training for experimental purposes is unfair to those so used. These sentiments sometimes even reach the point of putting direct pressure on evaluators as to how the study should turn out.

Much can be done to overcome these natural resistances if evaluation is viewed as a joint activity for the benefit of the training function. The evaluator must be accepted as more than a voyeur. To a large extent, mutual understanding of the concepts, approaches and methods of training evaluation will reduce the social distance between trainers and evaluators.

BIBLIOGRAPHY

Adams, Paul (1971) – Evaluating Non-Commercial Television: A Study. Austin: University of Texas, Center for Communication Research. Available from Eric Document Reproduction Service, 107 Roney Lane, Syracuse, New York.

Adult Education Association of the U.S.A. (1952) – Program Evaluation in Adult Education, Committee on Evaluation. Chicago. 38 pages.

American Institute for Research (1970) – Evaluative Research: Strategies and Methods. Pittsburgh, Pennsylvania. 160 pages.

Anderson, Stephen and Nancy E. Anderson (1971) – Human Relations Training for Women 25 (No. 3, August):24-27.

Argyris, Chris (1964) – T-groups for organizational effectiveness. Harvard Business Review (March):60-74.

Argyris, Chris (1968a) – Some unintended consequences of rigorous research. Psychological Bulletin 70 (No. 3):113-122.

Argyris, Chris (1968b) – Issues in evaluating laboratory education. Industrial Relations 9 (No. 3):96-103.

Aronoff, Joel (1971) – Achievement motivation training in executive advancement. Journal of Applied Behavioral Science 7 (No. 2, March-April):215-233.

Auerbach Corporation (1971) – The W.I.N. System: Analysis: Final Report and W.I.N. Model. Philadelphia, Auerbach Corporation, April 30. 135 pages.

Bass, Bernard M. and E.M. Vaughn (1967) – The anarchist movement and the T-group: Some possible lessons for organizational development. Journal of Applied Behavioral Science, April:212-213.

Becker, Howard S. (1954) – Field methods and techniques: A note on interviewing tactics. Human Organization (Winter):31-33.

Becker, Howard S. and Blanche Geer (1962) – Participant observation in interviewing: A comparison. Human Organization 16:28-32.

Belasco, James and Harrison M. Trice (1969) – The Assessment of Change in Training and Therapy. New York: McGraw-Hill Book Company.

Belasco, James and Harrison M. Trice (1969) – Unanticipated returns of training. Training and Development Journal 23 (No. 7, July):12-17.

Bell, Charles and William Buchanan (1966) – Reliable and unreliable respondents: party registration and prestige pressures. Western Political Quarterly 29:37-43.

Bell, Gordon (1968) – The adoption of business practices by participants in the small business management training program. M.S. thesis, British Columbia University, Vancouver. 128 pages.

Berreman, Gerald D. (1962) – Behind Many Masks. Ithaca, New York: Society for Applied Anthropology. Monograph No. 4.

Bjorkquint, D.C. (1970) – Technical education for the underemployed and the unemployed. Vocational Guidance Quarterly 18 (No. 4, June):264-272.

Blalock, Hubert M. (1964) – Casual Inferences in Non-Experimental Research. Chapel Hill: University of North Carolina Press.

Bolar, Malathi (1970) – Evaluating management development programs in industry. Training and Development Journal 24 (No. 3, March):24-40.

Bolman, Lee (1970) – Laboratory versus lecture in training executives. Journal on Applied Behavioral Science 6 (No. 3):323-335.

Bond, Nicholas (1970) – Measurement of Training Outcomes. University of Southern California at Los Angeles, Department of Psychology. Available from the National Technical Information Service, Operations Division, Springfield, Virginia.

Borus, Michael and Charles G. Buntz (1972) – Problems and issues in the evaluation of manpower programs. Industrial and Labor Relations Review 25 (No. 2, January):234-245.

Bremer, John (1968) – Management development: Multiple measurement of its effect when used to increase the impact of a long term motivational program. D.B.A. Thesis, University of Michigan, Ann Arbor.

Brophy, John W. (1971) – Television video-tape recorder – New tool for training in business and industry. Personnel Journal 50 (No. 9, September):716-719.

Bunker, Douglas (1965) – Individual applications of laboratory training. Journal of Applied Behavioral Science (April):131-148.

Clary, James N. (1970) – Training time and costs for navy ratings and NECS. Naval Personnel Research and Development Lab., Washington, D.C.

Campbell, Donald and D.W. Fiske (1959) – Convergent and discriminate validation by the multitrait-multimethod matrix. Psychological Bulletin 56:81-105.

Canealy, John (1968) – Management development training: Multiple measurement of its effects when used to increase the impact of a long term motivational program. D.D.A. thesis, Washington University, Seattle, Washington. 249 pages.

Catalanello, Ralph and Donald L. Kirkpatrick (1968) – Evaluating training programs: The state of the art. Training and Development Journal 22 (No. 5, May):2-9.

Champion, Dean and Alan Sear (1969) – Questionnaire response rate: A methodological analysis. Social Forces 43 (Spring):335-339.

Chandler, Margaret (1956) – An evaluation of the group interview. Human Organization (Spring):26-29.

Clark, John and Larry Tifft (1966) – Polygraph and interview validation of self-reported deviant behavior. American Sociological Review 31:516-5231.

Clary, James (1970) – Training Time and Costs for Navy Ratings. Naval Personnel Research and Development Laboratory. Available from National Technical Information Service, Operations Division, Springfield, Virginia. (July)

Coghill, Mary Ann (1967) – Sensitivity Training: A Review of the Controversy. Key Issue Series, No. 1, School of Industrial and Labor Relations, Cornell University, Ithaca, New York. (December) 26 pages.

Coleman, James, Sarane Boocoock and E.O. Schild (eds.) (1966) – Simulation games and learning behavior. The American Behavioral Scientist (October):1-32 and (November):1-35.

Conley, Ronald (1969) – A benefit-cost analysis of the vocational rehabilitation program. Journal of Human Resources 4 (No. 2, Spring):226-252.

Connell, Charles and Floyd Fowler (1963) – Comparison of a self-enumeration procedure and a personal interview: A validity study. Public Opinion Quarterly 27:250-264.

Cook, Alan (1971) – Share success – Full-time training for building crafts. Industrial Training International 6 (No. 6, June):164-167.

Cordery, Michael (1971) – Evaluation of external supervisory courses. Industrial Training International 6 (No. 7, July):204-205.

Cottis, J. (1971) – Training for the youth employment service. Careers Quarterly 23 (Spring):187-194.

Couch, Peter, and George B. Strother (1971) – A critical incident evaluation of supervisory training. Training and Development Journal 25 (No. 9, September):6-12.

Crane, Donald (1970) – Qualifying the Negro for professional employment. Ph.D. thesis, University of Georgia, School of Business Administration, Athens, Georgia.

Dada, Paul (1970) – Evaluation of Courses and Programs Offered Under the Auspices of Wayne State University and the University of Michigan at the University Center for Adult Education, Detroit, Michigan. University of Michigan, Department of Community and Adult Education, Ann Arbor, Michigan. 16 pages.

Davis, Earle E. (1968) – A Study of Low Wage Workers and Their Response to High Intensity Training. New York: Skill Advancement, Inc. (August) Available from Eric Document Reproduction Service, 107 Roney Lane, Syracuse, New York.

Dean, Gary and others (1969) – Regional Medical Programs: Guidelines for Evaluation. University of Southern California at Los Angeles, School of Medicine. Available from Eric Document Reproduction Service, 107 Roney Lane, Syracuse, New York. 24 pages.

Dralle, Penelope (1969) – The measurement of change during a laboratory training experience: Person perception and verbal behavior patterns. Ph.D. thesis, Louisiana State University, Baton Rouge, Louisiana. 166 pages.

Dunnette, M.D. and J.P. Campbell (1968) – Laboratory education: Impact on people and Organizations. Industrial Relations.

Eaton, Joseph (1962) – Symbolic and substantive evaluative research. Administrative Science Quarterly 6 (March):421-442.

Eastman Kodak Company (1971) – A system to create training systems. Training in Business and Industry 8 (No. 9, September):40-46.

Elliott, Elizabeth (1972) – A model for evaluating educational programs aimed at disadvantaged families. Paper presented at Adult Education Research Conference, Chicago, Illinois, April.

Etzioni, Amitai (1964) – Modern Organizations. Englewood Cliffs, New Jersey: Prentice Hall, Inc., pages 16-19.

Evans, Virginia (1970) – An Analysis of the Teaching Plan of the Adult Educator and Its Relationship to Teaching Effectiveness. Columbus, Ohio: Ohio State University, Cooperative Extension Service. (Abstract of a Dissertation.)

Everly, Jack (1970) – Instructional systems for extramural courses. Paper presented at the Adult Education Research Conference, Minneapolis, Minnesota, February 27-28.

Farmer, James A., et al. (1970) – Western Region AMIDS Evaluation: A Description of Evaluative Research Design and Methodology. Los Angeles: University of California, Division of Vocational Education. 19 pages.

Ferrari, F. (1970) – The open problem of management evaluation training. Management International Review 10 (No. 4-5):39-44.

Fleishman, Edwin, Edwin Harris, and Harold Burt (1955) – Leadership in Supervision in Industry. Columbus, Ohio: Ohio State University, Bureau of Research. Pages 29-54.

Florida State University (1972) – Self Study Report of the Department of Adult Education. Tallahassee: Florida State University, Department of Adult Education. (February, 44 pages.) Available from Eric Document Reproduction Service, 107 Roney Lane, Syracuse, New York.

Ford, John and Dewey Slough (1970) – Development and evaluation of computer assisted instruction for navy electronics training. Clearinghouse for Federal Scientific Information, Springfield, Virginia. 38 pages.

Friedlander, Frank (1967) – The impact of organizational training laboratories upon the effectiveness and interaction of ongoing work groups. Personnel Psychology 20:289-307.

Friedman, Neil (1967) – The Social Nature of Psychological Research. New York: Basic Books

Gill, J. and C.S. Molander (1970) – Beyond management by objectives. Personnel Management 2 (No. 8, August):18-20.

Goff, Maurice (1968) – Survey of present methods of follow-up of public post-secondary school graduates in cooperative and preparatory vocational programs and development of a follow-up. Ed.D. thesis, University of Wyoming, Laramie, Wyoming. 221 pages.

Goffard, S.J. and associates (1970) – Development and Evaluation of an Improved Radio Operators Course. Springfield, Virginia: National Technical Information Service, Operations Division (June).

Goodman, Paul S. (1969) – Hiring, training and retraining the hard core. Industrial Relations 9:54-66.

Goueck, William (1971) – Management training using telelectures. Training and Development Journal 25 (No. 11, November):12-16.

Greenberg, David H. (1970) – Employing the training program enrollee: An Analysis of employer personnel records. Rand Corporation, Santa Monica, California.

Greenberg, D.H. (1968) – Employers and manpower training programs: Data collection and analysis. Office of Economic Opportunity Memorandum RM-R740-OEO Santa Monica, California: The Rand Corporation (October).

Gurin, G. (1968) – Inner-city youth in a job training project. Institute for Social Research, University of Michigan (December).

Hardin, Einar and Michael Borus (1969) – Economic benefits and costs of retraining courses in Michigan. East Lansing: Michigan State University.

Heck, Edward (1968) – A study concerning the differential effectiveness of two approaches to human relations training in facilitating change in interpersonal communication skills and style of interpersonal perceptions. Ph.D. thesis, Syracuse University, Syracuse, New York. 180 pages.

Heller, F.A. (1970) – Group feedback analysis applied to training and learning situations. Journal of Management Studies 7 (No. 3, October):335-346.

Hinrichs, J.R. (1970) – Implementation of manpower training: The private firm experience. IBM Corporation, White Plains, New York.

Jaffee, Cabot, Stephen Cohen and Robert Cherry (1972) – Supervisory selection programs for disadvantaged or minority employees. Training and Development Journal 26 (No. 1, January):22-27.

John F. Kennedy Family Service Center (1969) – The Aging Worker. John F. Kennedy Family Service Center, Boston, Mass.

Kayloe, Alvin (1971) – A method for evaluating the effectiveness of technical training. Training and Development Journal 25 (No. 6, June):24-30.

Kincaid, Harry and Margaret Bright (1957) – The tandem interview: A trial of the two-interviewer team. Public Opinion Quarterly XXI (No. 2, Summer):304-312.

Kirkpatrick, Donald L. (1966) - How to start an objective evaluation of your training program. American Society of Training Directors 10 (No. 3, May-June):53-57.

Kirkpatrick, Donald (1969) – Evaluating a training program for supervisors and foremen. Personnel Administrator, Sept.-Oct.:29-38.

Lamp, B. and S. Hargreaves (1970) – The Esso students' business game. Technical Journal 8 (No. 5, June)4-16. New series.

Lawrie, L.W. and Clayton W. Boringer (1971) – Training needs, assessment, and training program evaluation. Training and Development Journal 25 (No. 11, November):6-10.

Lester, Richard (1971) – Criteria for evaluating training materials. Training and Development Journal 25 (No. 8, August):12-15.

Levinson, Perry (1966) – Evaluation of social welfare programs: Two research models. Welfare in Review 4 (No. 10):5-12.

Likert, Rensis (1932) – Technique for the measurement of attitudes. Archives of Psychology No. 140.

Little, J. Kenneth and Whinfield, Richard (1970) – Followup of 1965 Graduates of Wisconsin Schools of Vocational, Technical and Adult Education. Madison, Wisconsin: University of Wisconsin, Center for Studies in Vocational and Technical Education. Available from Eric Document Reproduction Service, 107 Roney Lane, Syracuse, New York.

Lord, Frederic (1970) – Problems arising from the unreliability of the measuring instrument. Pp. 79-93 in Research Strategies for Evaluating Training, Philip H. DuBois and G. Douglas Mayo (eds.) Chicago: Rand McNally and Company.

Los Angeles City Schools – Handbook for Evaluating Instruction. Available from Eric Document Reproduction Service, 107 Roney Lane, Syracuse, New York.

Management Technology Incorporated (1967) – Abstract of a Conceptual Model of Adult Basic Education Evaluation Systems. (January) Available from the Division of Adult Education Programs, Adult Education Branch. U.S. Office of Education, Washington, D.C.

Mangum, Garth (1967) – Contributions and Costs of Manpower Development and Training. Policy Papers in Human Resources and Industrial Relations, No. 5. Ann Arbor: University of Michigan, Institute of Labor and Industrial Relations. 95 pages.

Marcus, Alan (1971) – Cope training: What it does for managers. Continuing Education 4 (No. 3, July):47-49.

Marguilies, Newton (1972) · The myth and magic in organization development. Proceeding of the 31st Annual Meeting of the Academy of Management, 177-182.

McCarthy, Philip J. (1956) – Sampling: Elementary Principles, Ithaca, New York: New York State Schcol of Industrial and Labor Relations, Bulletin No. 15. 33 pages.

McCarthy, Philip J. (1957) – Introduction to Statistical Reasoning. New York: McGraw-Hill Book Company.

McConnell, John and Treadway Parker (1972) – An assessment centered program for multi-organizational use. Training and Development Journal 26 No. 3, March):6-14.

McCord, Bird (1971) – Identifying and developing women for managerial positions. Training and Development Journal 25 (No. 11, November):2-6.

Mesics, Emil A. (1969) – Education and Training for Effective Manpower Development. Bibliography Series No. 9, School of Industrial and Labor Relations, Cornell University.

Meyer, Thomas (1970) – A feasibility study in determining individual practice profiles of physicians as a basis for continuing education of these physicians utilizing a postgraduate perceptor technique. Final report. Wisconsin University, Madison.

Miles, N.B. (1960) – Human relations training: Prophecies and outcomes. Journal of Counseling Psychology V:l:301-306.

Mollenkopf, William (1960) – Some results of three basic skills training programs in an industrial setting. Journal of Applied Psychology 53 (No. 5):343-348.

Morgan, R.G.T. (1971) – Air Transport and Travel Industry Training Board, Staines, England. (March) Available from Air Transport and Travel Industry Training Board, Staines House, 158-162 High Street, Staines, Middlesex, England, 245 pages.

Nadeau, Richard (1960) – The Worker Three Months After High Intensity Training. (December) Springfield, Virginia: National Technical Information Service.

Nadler, Leonard (1971) – Support systems for training. Training and Development Journal 25 (No. 10, October):2-7.

Nattress, Leroy W. (ed.) (1970) – Continuing Education for the Professions. Proceedings of the Sections on Continuing Education for the Professions at the Galaxy Conference on Adult Education, Washington, D.C., December 8-10.

Nellum, A.L. and associates (1969) – Manpower and rebuilding. A study of six manpower development and training programs operating with rehabilitation and construction of housing. A.L. Nellum and Associates, Washington, D.C. 230 pages.

Niederfrank, E.J. (1970) – Working With the Disadvantaged, Washington, D.C.: U.S. Department of Agriculture, Federal Extension Service. Available from Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 11 pages.

Oates, David (1971) – Training by trivial tasks. International Management (April): 14-17.

Oatey, M. (1970) – The economics of training with respect to the firm. British Journal of Industrial Relations 8 (No. 1, March):1-21.

Odiorne, George (1963) – The trouble with sensitivity training. Training Directors Journal (October):9-21.

Oliver, Stanley (1971) – Dissecting the supervisor. Industrial Society (March-April):22-25.

Orth, Charles E. and Frederic Jacobs (1971) – Women in management: Pattern for change. Harvard Business Review 49 (No. 4, July-August):139-147.

Patten, George (1971) – The role of TV in training. Industrial Training International 6 (No. 6, June):180-182.

Perella, Vera (1968) – Women in the labor force. Monthly Labor Review (February – Special Labor Force Report No. 93):1-12. Available from U.S. Government Printing Office, Washington, D.C.

Perrow, Charles (1961) – The analysis of goals in complex organizations. Sociological Review 26:854-866.

Pickett, J.B. (1970) - Programmed instruction for the training of sales staff. Personnel Practice Bulletin 26 (No. 4, December): 238-245.

Prescott, Edward, and others (1971) Training and employability: The effects of MDTA on AFDC recipients. Welfare in Review, Jan.-Feb.: 1-6.

Quinn, R., B. Fine, and T. Levitin (1970) – Turnover in Training: A Social-Psychological Study of Disadvantaged Workers. Survey Research Center, University of Michigan (September).

Reeves, Elton and J. Michael Jensen (1972) – Effectiveness of Program evaluation. Training and Development Journal 26 (No. 1, January): 36-41.

Rising, Jesse (ed.) (1970) – Proceedings of the Conference on Evaluation in Continuing Medical Education, August 25-26, 1970, Kansas University, Kansas City Medical Center, Kansas City. Available from Eric Document Reproduction Service, 107 Roney Lane, Syracuse, New York. 102 pages.

Ritzer, George and Harrison M. Trice (1969) – The Personnel Manager: An Occupation in Conflict. Ithaca, New York: New York State School of Industrial and Labor Relations.

Roomkin, Myron (1970) – An Evaluation of Adult Basic Education Under the Manpower Development and Training Act in Milwaukee, Wisconsin. Madison: University of Wisconsin, Industrial Relations Research Institute. Available from National Technical Information Service, Springfield, Virginia.

Robertson, William O. and Charles W. Dohner (1970) – Study of continuing medical education for the purpose of establishing a demonstration center for continuing education in the Pacific Northwest. Final Report. Washington University, School of Medicine, Seattle.

Rosen, Howard (1969) – A group orientation approach for facilitating the work adjustment of the hard core unemployed. Final Report, U.S. Department of Labor.

Rosen, Ned (1970) Open systems theory in an organizational subsystem: A field experiment. Organizational Behavior and Human Perfotmance 5 (No. 3, May): 245-265.

Rosenthal, Robert (1966) – Experimental Effects in Behavioral Research. New York: Appleton-Century-Crofts.

Ross, T.C. and H.A. Shoemaker (1969) – An evaluation of listening training against job relevant criteria. National Society for Programmed Instruction 8 (No. 4, April): 14-18.

Rowan, Richard L. and Herbert R. Northrup (1972) – Educating the employed disadvantaged for upgrading. A report on remedial education programs in the paper industry. Wharton School of Finance and Commerce, Pennsylvania University, Philadelphia.

Roy, S.K. and A.M. Bolke (1971) – Evaluation of a supervisory training program. Training and Development Journal 25 (No. 12, December): 35-39.

Sanfield, Ronald (1971) – Senior Studies. Chicago Department of Human Resources.

Schmidt, Warren (1970) – How to evaluate a company's training efforts. California Management Review 12 (No. 3, Spring): 49-56.

Scott, Ted (1971) – Supervisors' evaluation of staff development activities. Training and Development Journal 25 (No. 9, September): 12-15.

Sellitz, Claire, Marie Jahoda, Morton Deutsch, and Stuart W. Cook (1959) – Research Methods in Social Relations. Revised edition. New York: Holt, Rinehart, and Winston.

Sewell, D.O. (1971) – Training the Poor. A Benefit-Cost Analysis of Manpower Programs in the U.S. Anti-Poverty Program. Industrial Relations Center, Queens University, Kingston, Ontario. 153 pages.

Siegel, Arthur (1967) – Post Training Performance Criterion Development and Application. Wayne, Pennsylvania: Applied Psychological Services. 15 pages.

Smith, Alvin (1971) – How the Toronto Dominion's Banklab makes the training of administration officers faster and more effective. Canadian Training Methods 4 (No. 2, May-June): 10-11.

Smith, P.B. and T.S. Honour (1969) - The impact of phase I managerial grid training. Journal of Management Studies 6 (No. 3, October):318-330.

Solie, Richard J. (1968) – Employment effects of retraining the unemployed. Industrial and Labor Relations Review 21 (No. 2, January):210-225.

Solomon, Herman (1969) – After Training: A Followup Report on MDTA-Course Graduates. New York State Department of Labor, Albany, New York, Division of Employment. Available from Eric Document Reproduction Service, 107 Roney Lane, Syracuse, New York. 21 pages.

Somers, Gerald (1964) – A Benefit Cost Analysis of Manpower Retraining. Proceedings of the 17th Annual Meeting, Industrial Relations Research Association, Chicago, December 28-29, pp. 172-185.

Somers, Gerald (1965) – Retraining, an evaluation of gains and costs. Chapter 9 in Arthur Ross (ed.), Employment Policy and the Labor Market. Berkeley: University of California Press.

Somers, Gerald (1971) – The Effectiveness of Vocational and Technical Programs: A National Followup Survey. Madison, Wisconsin: University of Wisconsin, Center for Studies in Vocational and Technical Education. Final Report. 264 pages.

Somers, Gerald G. (ed.) (1968) – Retraining the unemployed. Ford Foundation, New York.

Somers, Gerald G. (1967) – Our experience with retraining and location. Chapter 8, Toward A Manpower Policy, edited by Robert A. Gordon. New York: John Wiley.

Suchman, Edward A. (1967) – Evaluative Research. New York: Russell Sage Foundation. 178 pages.

Thompson, James D. (1967) – Organizations in Action. New York: McGraw-Hill Book Company.

Tobias, Jack and associates (1969) – A survey of the employment status of mentally retarded adults in New York City. Association for the Help of Retarded Children, New York.

Tolela, Michele (1968) – Effects of T-group training in cognitive learning on small group effectiveness. Ph.D. thesis, Denver University, Denver, Colorado. 144 pages.

Tracey, William R. (1968) – Evaluating Training and Development Systems. New York: American Management Association. 304 pages.

Trice, Harrison M. (1959) – A methodology for evaluating conference leadership training. ILR Research V (No. 2 and 3, Fall):2-6.

Trice, Harrison M. and James Belasco (1968) – Supervisory training about alcoholics and other problem employees: A controlled evaluation. Quarterly Journal of Studies on Alcohol 29:282-298.

Trice, Harrison M., James Belasco, and Joseph Alutto (1969) – Role of ceremonials in organizational behavior. Industrial and Labor Relations Review 23 (No. 1, October).

Trice, Harrison M. and Robert V. Penfield (1961) – Use of Application blank data in a study of job quitting. ILR Research VII (No. 2, Summer):9-14.

University of South Carolina (1968) – A Benefit-Cost Analysis of the South Carolina M.D.T.A. Program. Columbia: University of South Carolina, Bureau of Business and Economic Research.

U.S. Department of Labor (1967) – Earnings Mobility of M.D.T.A. Trainees. Washington, D.C.: U.S. Department of Labor, Manpower Administration. (April)

Valiquet, Michael (1968) – Individual change in a management development program. Journal of Applied Behavioral Science 4 (No. 3):313-325.

Warr, Peter, et al. (1968) – Evaluating management training. Association of Teachers of Management Bulletin 8 (No. 2, July):1-13.

Warriner, Charles (1958) – The nature and functions of official morality. American Journal of Sociology 64:165-168.

Wasmuth, William (1970) – Workshop: A dynamic simulated training program. Rehabilitation Record (July-August):12-16.

Weingarten, Kenneth and others (1971) – The APSPRAT Instructional Model. Washington, D.C.: U.S. Army, Office of the Chief of Research and Development (May).

Weiss, Carol (1970) – Politicalization of evaluative research. Journal of Social Issues 26:57-68.

Weiss, Robert and Martin Rein (1970) - The evaluation of broad-aim programs: Experimental design, its difficulties and an alternative. Administrative Science Quarterly 15:97-109.

Wells, Jean A. (1969) – Facts About Women's Absenteeism in Labor Turnover. (August) Available from U.S. Government Printing Office, Washington, D.C. 13 pages.