

A CASE STUDY OF TRAINING EVALUATION

by

Terrence J. South
B.A., Hope College, 1976

Submitted to the Department of Speech
and Drama and the Faculty of the
Graduate School of the University
of Kansas in partial fulfillment
of the requirements for the degree
of Master of Arts.

Redacted Signature

Professor in Charge

Redacted Signature

Redacted Signature

Committee Members

Redacted Signature

For the Department

ABSTRACT

This study examines in depth the area of corporate training evaluation, and evaluates one training evaluation program that is actually in use in a large airline corporation's Reservations Department.

"Corporate or industrial training" is viewed as any formal training sponsored by an organization with an explicit goal of improvement in job performance or productivity. "Evaluation" is any systematic activity designed to identify goals of training and measure progress made toward those goals.

Comments about and a description of the nature of training and evaluation in organizations are provided, and the existing literature on training evaluation is reviewed. Based on this review, a model for evaluation is developed. The model consists of four sequential components--rationale, formulation, implementation, and use--and five descriptive characteristics--planned, unified, flexible, appropriate, and complete--that should be included in a corporate training evaluation program.

The model serves as a framework with which to evaluate the "real" training evaluation program under consideration. A detailed description of the organization, the training system, the program to evaluate training, and its results are provided. Using this description, the airline corporation's Reservations Training evaluation program is critiqued.

The Reservations evaluation program was found to be, on balance, a worthwhile evaluation effort, under the circumstances. It provided a sound base for an evaluation program which could be refined in the future, and it provided significant organizational benefits outside the specific area of training evaluation. The major shortcomings of the program were that its focus was not sufficiently limited, and that its experimental design was ineffective, with respect to training evaluation. The program performs well against the five descriptive characteristics of a good training evaluation program.

This study concluded that the development of a training evaluation model is possible and that it is a useful tool in improving training evaluation in a corporate setting. It was also pointed out that there is a distinct need for further theory, research, and experimentation in the area of training evaluation.

In addition, the conclusion was drawn that informal or "non-official" factors play an important role in official organizational activities, but that they are difficult to discover and document.

ACKNOWLEDGEMENTS

My thanks go to Professor Cal Downs, who was a great help in this project from beginning to end, and who has taken a much appreciated involvement in my professional growth.

Also appreciated are the efforts of Professor Bill Conboy and Professor Kim Giffin, who agreed to serve on the advisory committee for this study.

I am grateful to Don Adams, Paula Smith, Al Doeve, Mark Wolf, and Mike Kelly for making this study possible.

Finally, thanks to my wife Ginger who helped in a number of ways.

T.J.S.

CONTENTS

| | |
|--|-----|
| ACKNOWLEDGEMENTS. | ii |
| Chapter | |
| I. Introduction. | .1 |
| Purpose and Scope | 2 |
| Statement of the Problem | 4 |
| Definition of Terms | 4 |
| II. Background and Perspective. | .7 |
| Emergence of Evaluation Research | 9 |
| Nature of Training | 11 |
| Nature of Evaluation | 13 |
| Rationale for Not Evaluating Training | 16 |
| Rationale for Evaluating Training | 19 |
| III. Review of the Literature. | 26 |
| Issues in Training Evaluation: Formulation | 27 |
| Issues in Training Evaluation: Implementation | 42 |
| Issues in Training Evaluation: Use of Data | 55 |
| Conclusion | |
| IV. Model for Training Evaluation | 62 |
| Components | 62 |
| Key Characteristics | 66 |
| V. Methodology for Evaluating the TNA Program. | 68 |
| Organizational Background | 68 |
| The Training Needs Analysis (TNA) Program | 76 |
| Data Collection and Analysis | 80 |
| Summary | 86 |
| VI. Results of TNA. | 89 |
| Item Analysis | 89 |
| Comparison to MAP | 93 |
| Analysis of Variance Between Locations | 96 |
| VII. Evaluation of TNA | 99 |
| Rationale | 100 |
| Formulation | 102 |
| Implementation | 105 |
| Use of Data | 108 |
| Evaluation Against Key Characteristics | 111 |

| | |
|--|------|
| VIII. Discussion and Recommendations | .114 |
| Conclusions | 114 |
| Recommendations for Future Research | 120 |
| A Final Note | 124 |
| BIBLIOGRAPHY. | .126 |

LIST OF TABLES

Table

| | |
|---|-----|
| I. Results of Item Analysis. | .92 |
| II. Comparison of Results to MAP Scores | .95 |
| III. Results of Analysis of Variance | .97 |

CHAPTER ONE

INTRODUCTION

Human resource development is important because the expertise and abilities of individual organization members have such a great effect on the total effectiveness of the organization. Employee training is a significant technique for developing human resources, and if the amount of energy spent in training is any indication of its success, training works.

Evaluation is also integral to effective organizational operation. Evaluation, appraisal, or assessment is done in practically every phase of organizational activity to ensure that people, resources, and programs are making a contribution to the goals of the enterprise. Some form of performance feedback is essential to organizational effectiveness. But training and human resource development, unfortunately, are areas that present evasive targets for evaluation. As a result, recent research indicates that the identification of explicit training goals and the measurement of progress made toward those goals is done far too infrequently.¹ In spite of this, the effective evaluation of training not only can be done but can also improve training programs, save money, increase productivity, and direct resources to where they are most needed.

Many corporations in America have a great deal of faith in training as a technique to improve performance. How well-deserved

that faith is, moreover, is clearly an issue that sound training evaluation can address. Training evaluation--assessing past training, and identifying directions for future training--is part of the theory of effective training. The following study examines the options for developing and implementing training evaluation in a corporate environment.

PURPOSE AND SCOPE

The purpose of this study is twofold. First, a model of training evaluation, based on theoretical and experimental research from the literature, is developed. Second, an actual industry training evaluation program is presented and evaluated against the criteria developed in the model. Both the method and the results of one administration of a training evaluation program at a major airline are presented in depth. The actual data representing the results of that evaluation program will serve as an example of both the utility and the shortcomings of the particular evaluation method.

The unit that was studied was the Sales and Services Training division of the airline. This division is responsible for training virtually all airport support personnel from baggage handlers to airport managers, to reservations and ticketing agents. Within that unit, the training for reservations agents was the specific activity under scrutiny. A program to evaluate reservations training had been recently developed, and that program was analyzed.

The study attempted to integrate theory and practice. While the shortcomings of actual implementation vis-a-vis idealized theory were discovered, the constraints and exigencies of a profit-making

enterprise were seen to affect the practicality of certain theoretical approaches. This study examined both perspectives.

The scope is defined in part by the author's perspective. The first element of this perspective is that only training programs in an industrial setting, and with a tangible task-related goal, were considered. This is defined not as much by the surroundings or the institution in which the training occurs, as by the goals and purposes of the training. The training with which this study was concerned was that which had as its goal improved performance in a job.

Second, the entire training function is viewed as an integrated whole. Evaluation, actual training, needs assessment, and program design, all have to be considered as interdependent. Therefore, training evaluation, for example, cannot be examined apart from the other elements of the training function. Similarly, the training function must ultimately be considered within the entire firm or institution's system.

The limitations of this study also fall into two areas. First, the design necessitates that only one case of training evaluation is examined in-depth. Consequently, the results can be generalized only to similar programs occurring in industry. In spite of this, the model presented will not be limited solely to the type of training evaluation at this corporation; it will be a broad theoretical model.

Second, the scope of this study was not limited to the actual outcome of this particular training evaluation. How effective or ineffective the training was, was important only insofar as it was relevant to the evaluation and needs assessment functions. In other words, this study was not an evaluation of training; it was an evaluation of an evaluation of training. The evaluation methods in the program

which was examined were of far more interest than their training outcomes.

STATEMENT OF THE PROBLEM

A succinct statement of the research questions under consideration in this study is as follows:

- A.) What are the criteria and components of an effective training evaluation program for training in an industrial setting?
- B.) To what extent does the Sales and Services "Training Needs Analysis" (TNA) meet these criteria and involve these components?

DEFINITION OF TERMS

The assumptions involved in the definition of key terms in the research questions follow.

1.) "Training evaluation program" is any systematic activity designed to both identify goals of training and to assess the progress made toward those goals through training.

2.) "Effective" indicates the degree to which the program is meeting three goals:

--supplying reliable information as to whether training has brought about desired changes

--identifying in what specific areas those changes have been brought about

--providing information on how training could be changed to make it more effective.

3.) "Criteria" are key characteristics or necessary qualitative features.

4.) "Components" are actual elements, phases, or structures of a

training evaluation program.

5.) "Training in an industrial setting" is any formal training sponsored by an organization with an explicit goal of improvement in job performance or profitability.

6.) "Training Needs Analysis" (TNA) is a series of tests used by this corporation's Sales and Services Training Division to assess an employee's level of knowledge in various areas of Sales and Services functions. There is a different set of tests for reservations, airport services, and maintenance employees; in this study only the reservations TNA tests were considered.

NOTES: CHAPTER ONE

¹Michael A. Raphael and Edwin E. Wagner, "Training Surveys Surveyed," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 295.

CHAPTER TWO

BACKGROUND AND PERSPECTIVE

The amount of training in corporate America is growing. In a Department of Labor survey conducted in 1962, 70% of firms with 500-999 employees, 85% of firms with 1000-2499 employees, and 96% of firms with 5,000 or more employees, had training programs.¹ Seven years later, in 1969, in a Bureau of National Affairs survey of over 200 member industrial representatives, it was reported that 75% of the companies had some sort of formal training program.² Almost every large institution, public or private, conducts some training for their employees, and many do it on a large scale.

Furthermore, training has become big business. One author quoted a price of a one-week management training session for twelve persons, at \$3500, in 1967 dollars.³ Anyone even tangentially involved with training in industry will testify that that figure, even translated into 1980's inflated dollars, is conservative. When one considers also the on-the-job work time sacrificed during training, it is clear that corporations in America are willing to invest a great deal of time, money, and resources in training.

Is this investment justified? Those of us in the field of organizational communication think so. Richard Farace, for example,

suggests communications training programs as one of the principal job responsibilities of the corporate communications manager.⁴ The entire human relations school in business views various types of interpersonal training as necessary to "humanize" the business environment, increase morale and ultimately improve productivity. Certainly, effective training in technical job requirements will also reap benefits in job performance, especially in highly skilled or knowledge-intensive jobs. Other benefits that can accrue through training are improved productivity, customer service, sales effectiveness, and morale.

The key to all the benefits cited above, however, is not training, but effective training. If changes in information, skills, or attitudes are not produced, any amount of training may be accomplishing very little. And in spite of the large amounts of money being poured into industrial training, most companies are quite lax in assessing the effectiveness of their training. Training evaluations, "are, like the weather, things that everyone talks about but few do anything about."⁵ A study by Blumenfeld and Crane (1973) investigated the perceptions of training effectiveness and the evidence used in evaluating training, and reported that "there was very infrequent report of quality evidence of effectiveness, and the relationship between the perceptions and evidence was essentially zero."⁶

However, the benefits of a good training evaluation program are clear: three benefits, in particular, are extremely important. First, training evaluation can provide information on whether the training is doing what it is supposed to be doing, i.e., if it is effective. Second, an evaluation program can ascertain in what specific areas time, energy, and other resources are being wasted on an effort

that has been proven to be ineffective. In an era of threatening recession, this is a valuable and timely concern, both to business and education. Third, and perhaps most important, certain evaluation programs can suggest ways to remedy whatever problems exist. Of the three, this step is probably the most valuable, and correspondingly the most difficult.

The likelihood of achieving any of these three benefits of training evaluation, however, is dependent on the quality of needs assessment, and goal identification that accompany the training and the training evaluation. In the model that is presented in Chapter Four, this needs assessment may precede, coincide with, or follow the actual evaluation. But it is, and should be, a conceptual part of the evaluation process. Further, it will be shown that the three benefits of training evaluation cited above will be constrained by the corresponding element of training needs assessment because training evaluation can only be as good as the needs assessment that goes along with it. In other words, training evaluation and needs assessment are interactive and cannot be usefully separated, from the perspective of the entire training function.

EMERGENCE OF EVALUATION RESEARCH

In the early 1960's, as the management development push in training was gaining momentum, another field of study was emerging: program evaluation. The massive influx of federal money into human services programs began to prompt a need to determine how effective those programs actually were. Later, in the 1970's as federal resources

became more scarce relative to the competition among programs, and the American tax-paying public began to wonder if their tax dollars were well spent, the concept of "accountability" gained credence. Consequently, "the requirement for evaluation research was a political response to the perceived demand for increased governmental accountability,"⁷ and this response became a fixture in social research. Currently, the concept of evaluation research has become, and promises to increase as, an integral part of programs involving federal spending on human services.

Apparently the influence of the new field of evaluation research has filtered to the private sector, at least in the area of training evaluation. As employee training is one of the most intangible and "human-oriented" expenditures that companies make, it parallels closely the type of evaluation concept developed for government human services programs. The idea of evaluating industrial training really began to get off the ground in 1959-60 when a series of articles by Donald Kirkpatrick were published in the Journal of the American Society for Training and Development.⁸ These articles contained a theoretical rationale, albeit a simple one, to use in approaching industrial training evaluation. Since then the interest and activity in training evaluation has grown to the extent that in 1975, the ASTD considered it important enough to publish a collection of articles on the subject, as a book. An examination of the last 5 issues of the monthly Training and Development Journal before May 1980 reveals that each issue contains at least one article on training evaluation.

Training directors obviously consider evaluation of their programs a legitimate and necessary field of study and discussion. However, in

spite of the increased attention and talk about training evaluation, there are a number of issues around which the gap between discussion and practice is enormous. Not the least of these issues is the suggestion that evaluation is not really practiced, but just talked about; "the practice of evaluation has not kept pace with the prescription."⁹ This question, along with the various arguments for and against training evaluation, are explored in the sections following.

NATURE OF TRAINING

Most companies engage in a wide range of employee education and development activities ranging from tuition assistance plans to sponsorship of public interest speakers. However, the present study is interested in those training activities which prepare or assist organizational members or employees in performing a task and/or contributing to the organizational goals. Hinrichs defines training as "a systematic intentional process of altering behavior of organizational members in a direction which contributes to organizational effectiveness."¹⁰ The key concepts in this definition are "systematic," "intentional," and in a certain "direction."

Campbell, et al., indicate four features of training and development: 1) a learning experience, 2) planned by the organization, 3) after the individual has joined the organization, and 4) intended to further the organization's goals.¹¹ Hamblin defines training as "any activity which deliberately attempts to improve a person's skill in a job (as opposed to education which is mainly concerned with personal development as opposed to direct job-relevance)."¹² Hamblin, like many writers, sees

a subtle difference between "training" and "development," but as he considers development simply a training for a future job, he includes development under the broad concept of training stated above.

There are several key features to the definition of training used. The most important is that training can be related to some relatively tangible goal or objective of an organization. In this concept training may or may not job-specific, as long as its contribution to the organization can be identified. Consequently, like Hamblin, development is included in the definition of training used in this study. Second, the focus is on formal, intended, systematic training; while the results may not be intended or systematic, the training is. Finally, the organization involved may be public or private, profit or non-profit. Because the ultimate goals of a private, for-profit organization are usually more clear, it will appear that those organizations are the focus of this study. However, public and/or non-profit organizational training is included in this study, as long as it has as its objective a relatively tangible contribution to the organization's goal.

The training that is discussed in the literature would appear to have a significant role to play in corporate and organizational America. It is the key to developing human resources, which some theorists identify as the most valuable asset of a firm,¹³ even though it does not appear on the balance sheet. Geary Rummler calls it "probably the single most critical function in the effective use of human resources to meet today's pressing manpower needs."¹⁴ John Hinrichs specifies training as one of the key elements in the productive overall organizational system, and identifies the crucial role it plays.¹⁵ Lynton and Pareek see training on a global scale as "a means to reduce obsolescence among

people and organizations in the face of relentless technological innovation."¹⁶ And twenty years ago industrial training was described as a potent weapon in the then-raging Cold War: "Training all personnel to peak efficiency is. . . a key factor in the struggle between East and West."¹⁷

However, in spite of these almost prosaic pronouncements, there also appears in the same literature a curious contradiction regarding training in practice. Rummler describes the training function in today's organizations as enjoying little support, being poorly budgeted, and having little respect from the rest of management.¹⁸ Hinrichs, in a survey of personnel training, calls it "disorganized," rife with fads, and having little or no concern with theory; in short his picture was "not a particularly bright assessment."¹⁹ While these problems may be overstated, they do point out the need for training activities to provide evidence of their positive impact. Undoubtedly, training accomplishes significant outcomes; there is simply not enough emphasis on documenting those outcomes.

NATURE OF EVALUATION

The literature indicates that the popularity of training evaluation has greatly increased in the past twenty years, but the evidence is that much of this theory has not found its way into practice. Perhaps because of the "philanthropic" attitude toward training traditionally found in industry, there has not been a great practical impetus to conduct evaluation, assessment, or impact studies of training. In addition, the effects of training are typically difficult to measure. Consequently, two conditions have arisen: training is probably less effective than

it can be, and it is not taken seriously because its worth has not been proven.

The British Industrial Training Act of 1964 establishes two basic objectives of evaluation: validation and evaluation.²⁰ The first, validation, involves the extent to which training has accomplished what it intended to do. Internal validation attempts to determine whether trainees have learned the skills, attitudes or knowledge that the training was designed to teach. External validation attempts to determine whether what was learned in the training program has resulted in the desired behavior change on the job. It involves not only the success of the training program in itself, but also the transferability of what was learned to the job situation.

While external validation asks a broader question than internal validation, the concept of "evaluation" involves an even broader question than both. Evaluation attempts to justify the goals, processes, and outcomes of a training program in the broadest sense. Is this training relevant to the company or industry involved? Can its cost be justified relative to its benefits? Are the objectives of the training worthwhile? Could these objectives be accomplished in some other, more efficient way? In other words, evaluation involves "the assessment of the total value of a training system, training course or program in social as well as financial terms."²¹

It would appear that the British have refined the concept of evaluation considerably more than Americans, at least in the realm of theory. No doubt this is due to the fact that England has, in effect, nationalized industrial training. All firms in any industry are assessed an equal training levy, and then are subsidized for training costs, to

the extent that they can show a benefit from training. Thus, there exists a motivation to be able to measure training costs accurately, and to assess clearly the benefits (effectiveness) of training. In addition, the overall worthwhileness of the training's objectives are taken into account in the consideration of the subsidy. While the relative merits of the British system as public policy will not be considered here, it is clear that it has at least resulted in a closer and sharper focus of the role of evaluation in training.

While training evaluation can have several different aspects, it can also be done for several different reasons. The first and most obvious is simply to assess the impact of training against its stated goals. Evaluation done for this reason does not involve itself with why or how training succeeded or failed; it simply wants to know to what extent did training do its job. This, in the broadest sense, is evaluation of training. Its primary motive is justification.

A slightly more refined motive for training evaluation seeks to identify in what particular areas training was effective or ineffective. While the previous question queried the impact of a training program as a whole, this question inquires as to the effectiveness of various components within the training program. These components may be subject areas, different skills, various learning experiences, identifiable attitudes, or different instructors or methods. The reason for this type of evaluation is to gather more detailed information on where training reached its goal.

Finally, a reason for evaluation may be to answer the question: how can training be improved? This approach is qualitatively different from the first two and is what Randall calls the "refinement approach."

In this approach, "the assumption that training does make a contribution is never in question; instead, the only issue is how that contribution can be increased to the fullest degree."²² This type of evaluation seeks to gather data which will suggest tangible and specific ways in which training can more effectively meet its objectives.

Each of these three reasons for initiating training evaluation obviously asks a different question, and each will obtain different types of answers. The rationale for engaging in training evaluation involves each of these reasons to some extent, and the benefits of evaluation have implications for each of these types of evaluation. A discussion of the rationale and benefits of evaluating training, and of not evaluating training, follows.

RATIONALE FOR NOT EVALUATING TRAINING

There are a number of reasons why organizations do not institute training evaluations. The rationale for not evaluating training, like the rationale for evaluating training, depends on the nature and objectives of the training, and who is initiating and using the evaluation. Some of the common rationales for not evaluating training are outlined below.

1. Management is already convinced of training's effectiveness.

This is probably the single most entrenched belief in industrial training. Mindak and Anderson, in a 1971 article entitled "Can We Quantify an Act of Faith?," quote Dun's Review as reporting that "management training remains the American corporation's most complete act of faith."²³ They argue that the main reason evaluation is not done is because management

does not want to waste resources on something it has convinced itself is good. While I have labeled this belief a rationale, Bunker and Cohen call it a rationalization which is based on the fallacious view that "since training is developed by company experts in response to specific company needs no evaluation is needed."²⁴ Whether this belief is well- or ill-founded, it is clearly widespread.

2. Training evaluation is impossible to do correctly or well.

This is the point of view that Randall ascribes to what he calls "the negativists," and in practice this difficulty prevents many training practitioners from implementing evaluations. The rationale is that there are too many possible problems, inconsistencies, inconveniences, and unexplainables to make even attempting a training evaluation a waste of time. This view argues that scientific, rigorous evaluations have rarely any practical, bottom-line application, while evaluations connected with results and profitability suffer from too many discrepancies and intervening variables to make them valid. It should be pointed out that there are many types of training and some are easier than others to evaluate; this distinction is often ignored.

Randall cites Kenneth Andrews writing in the Harvard Business Review, who summarizes this view of management training: "To prove the quantitative contribution of formal education may remain impossible and finally be considered irrelevant."²⁵

3. Training evaluation may fail to indicate effectiveness.

This is the converse of the rationale for performing evaluations to obtain data which justifies the existence of the training program. Just as an honest and valid evaluation could indicate that a training

program is effective, the same evaluation could indicate that the program is ineffective. This information, at a time when budgets are being scrutinized, could be potentially disastrous for a training department. From an objective standpoint, it is clear that evidence of ineffectiveness is just as valuable, if not more so, than evidence of effectiveness. This information could be used to modify, adjust, or eliminate certain parts of a training program, not simply pronounce a verdict of "failure" on the training program. However, it is not always seen this way by training departments, especially when the evaluation is requested by top management. In this case, a negative evaluation can mean dire consequences, and training departments would prefer not to risk an evaluation, especially when "the training director's very occupational existence is threatened and depends greatly on the positive outcome of an evaluation of his current programs."²⁶

4. Training evaluation is too expensive. This line of reasoning is usually coupled with a conviction that training evaluation is unnecessary, impossible, or undesirable. The view is that the primary activity--training--should be receiving the resources, and that any other activity--evaluation--is diverting resources away from the most important function. Reinforcing this belief are the statements in evaluation articles which state that evaluation must be experimental, have pre- and post-measures, use control groups, and have objective criteria--all apparently expensive.

5. Effectiveness of training is irrelevant to the goal of the program. While few, if any, training departments would say they are uninterested in the effectiveness of their training, this may be the

case to some degree, in some training organizations. In these instances, "irrelevant" is defined as meaning "not central to the main goals of the training program." Although the irrelevance of training effectiveness would seem to be unlikely in profit-making organizations, it would certainly help to explain the lack of enthusiasm for evaluation that exists in many organizations.

If certain organizations are not interested in how effective their training is, why bother spending the money on training? In training as in practically any activity, there are overt stated goals, and there are underlying, unstated goals. Garbutt suggests a number of reasons corporations engage in training, two of which are "sense of social responsibility" and "spending for prestige purposes."²⁷ Belasco and Trice cite the "ceremonial" aspects of training which establish training as a necessary and desirable ritual which must take place in the organization.²⁸ All of these factors offer good reasons to have training in an organization, but do not indicate that evaluation of effectiveness is necessary, at least for the stated goals of training.

RATIONALE FOR EVALUATING TRAINING

There are at least seven good reasons for an organization to evaluate its training. All of these are not relevant to every organization but each is relevant to a large segment of organizations. Although their relevance depends on who asked for evaluation, who will use it and what the objective of evaluation is, each of the following seven reasons illustrates a reasonable rationale for training evaluation.

1. Evaluation results can justify the existence of a training department. If training did not accomplish anything, training departments would not exist. But perhaps this contention is an assumption for which there should be a greater body of support. On both the abstract and pragmatic levels, training evaluation can serve to connect more clearly the purposes of industrial training with their effects.

2. Evaluation can justify, protect, and increase training budgets. This reason takes the concept of reason #1 and operationalizes it. Resource allocations, and hence budget-makers, dictate to a great extent the power, scope, effectiveness, and freedom of a unit's activity. Since training makes one of the less tangible and less direct contributions to profit (as compared to production or sales, for example) it is more vulnerable to cutbacks of resources. However, some reliable and reasonable evidence to indicate a positive impact of training on organizational performance is a strong argument for a safe, or even augmented budget. If there are "training organizations that disappear in strenuous economic times because they are unable to show any apparent value to the organization, when asked,"²⁹ it follows that the training departments that have positive evaluation data are less likely to disappear or shrink.

3. Training programs can be improved through evaluation. To evaluate a training program for "diagnostic" or "refinement" purposes is among the strongest reasons for evaluation for a training department. Evaluations can identify weaknesses of training, suggest areas and ways in which the training could be made more effective, and indicate particular areas in which training is unnecessary. With refinement evaluation,

"the results of the evaluation are analyzed and used to redesign certain parts of the program that are apparently falling short of their mark."³⁰ In this way, evaluation becomes an integral part of the training activity and makes a key contribution to increasing the effectiveness of the training department. It is perhaps the clearest rationale for evaluation of training.

4. Evaluation is used to assess future needs in training. Implied in an assessment of the impact of training are consequences for future training. If a training program is found to have less than the desired impact, there still exists a need to have that training gap filled. This is particularly important in industry where the paramount importance is having capable, well-trained workers, rather than excellent training techniques. Along with everything else, training evaluation can give a gauge of what needs still exist which have not been met.

5. Evaluation can lend credibility to the training department and its programs. A training function which makes an honest effort to evaluate its performance will probably be seen as more credible by its customers and users, than one that does not. The evaluating of training programs indicates some degree of accountability and the existence of an established goal; this, in turn would indicate a demonstrated concern for effectiveness.

In addition, many of the evaluation efforts that are actually used, ask the trainees to evaluate the training and provide suggestions for improvement. This request for feedback ("message to the source conveying knowledge of the effectiveness of a previous communication") would indicate that the trainers are open to suggestions: "attention to

feedback implies an orientation to the receiver, a concern with whether he or she is 'getting the message.'"³¹ It would appear that this type of evaluation would increase the believability of a training program.

6. Training evaluation can help to determine whether a need is directly a training need. Although there are many organizational needs and management problems, not all are needs or problems which can be resolved through training. Gilbert, among others, offers an argument and a technique for separating training needs from management, motivation, or job design needs.³²

All needs or problems in an organization can conceivably be called training needs in that they all reflect actions made by someone, which can be altered through training. But many of these decisions, especially on the higher levels, are less directly available to training solutions. Examples are corporate policy which causes low morale, the economic or governmental environment, and ineffective hierarchical structures.

Although most of the discussion on this subject is related to training needs assessment, it is also a problem that training evaluation can help to address. The evidence that training evaluation can be designed to offer, can answer the question, "Is the problem amenable to a training solution?"³³

7. Evaluation programs can produce basic behavioral science research data. While the bulk of training evaluation done in industry is to determine training's contribution to overall organizational goals, sometimes evaluation may also collect data for less specific purposes. The corporation may assume the role of a basic researcher, and simply collect data on any of the behavioral changes induced by training. These

changes may involve behavior which has no apparent connection with explicit or implicit organizational goals. But like much of the other basic research done by U.S. corporations, the knowledge discovered may ultimately prove applicable to the furtherance of the organization's goals.

NOTES: CHAPTER TWO

¹Michael A. Raphael and Edwin E. Wagner, "Training Surveys Surveyed," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 295.

²Raphael and Wagner, p. 295.

³O.C. Lott, "Evaluating to Reduce Training Costs," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 245.

⁴Richard V. Farace, Peter R. Monge, and Hamish M. Russell, Communicating and Organizing (Reading, Massachusetts: Addison-Wesley, 1977), p. 257.

⁵J. W. Lawrie and Clayton W. Boringer, "Training Needs Assessment and Training Program Evaluation," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 291.

⁶Warren S. Blumenfeld and Donald P. Crane, "Opinions of Training Effectiveness: How Good?," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 305.

⁷Michael Q. Patton, Utilization-Focused Evaluation (Beverly Hills: Sage Publications, 1978), p. 14-15.

⁸Donald L. Kirkpatrick, ed., Evaluating Training Programs (Madison, Wisconsin: American Society for Training and Development, 1975).

⁹Kerry A. Bunker and Stephen L. Cohen, "Evaluating Organizational Training Efforts: Is Ignorance Really Bliss?," Training and Development Journal, 32, No. 8 (1978), p. 4.

¹⁰John R. Hinrichs, "Personnel Training," Handbook of Industrial Psychology, ed. Marvin D. Dunnette (Chicago: Rand McNally, 1976), p. 832.

¹¹John P. Campbell, Marvin D. Dunnette, Edward E. Lawler III, and Karl E. Weick, Jr., Managerial Behavior, Performance and Effectiveness (New York: McGraw-Hill, 1970), p. 233.

¹²A.C. Hamblin, Evaluation and Control of Training (London: McGraw-Hill, 1974), p.6.

¹³Malcolm M. Warren, Training for Results (Reading, Massachusetts: Addison-Wesley, 1969), p. xiii.

¹⁴Geary Rummler, in Warren, p. xiii.

¹⁵Hinrichs, p. 840.

¹⁶Rolf P. Lynton and Udai Pareek, Training for Development (Homewood, Illinois: Richard D. Irwin, 1967), p. vii.

- ¹⁷Frank A. DePhillips, William M. Berliner, and James J. Cribbin, Management of Training Programs (Homewood, Illinois: Richard D. Irwin, 1960), p. 4.
- ¹⁸Warren, p. xiv.
- ¹⁹Hinrichs, p. 830.
- ²⁰Douglas Garbutt, Training Costs (London: Gee and Co., 1969), p. 15.
- ²¹Hamblin, p. 8.
- ²²Lyman K. Randall, "Evaluation: A Training Dilemma," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 187.
- ²³William A. Mindak and Robert E. Anderson, "Can We Quantify an Act of Faith?," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 87.
- ²⁴Bunker and Cohen, p. 4.
- ²⁵Kenneth Andrews, in Randall, p. 187.
- ²⁶Randall, p. 187.
- ²⁷Garbutt, p. 15.
- ²⁸James A. Belasco and Harrison M. Trice, "Unanticipated Returns of Training," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 68.
- ²⁹Karen S. Brethower and Geary A. Rummier, "Evaluating Training," Training and Development Journal, 33, No. 5 (1979), p. 15.
- ³⁰Randall, p. 188.
- ³¹Everett M. Rogers and Rekha Agarwala-Rogers, Communication in Organizations (New York: Free Press, 1976), p. 14.
- ³²Thomas F. Gilbert, "Praxeonomy: A Systematic Approach to Identifying Training Needs," Management of Personnel Quarterly, 6, No. 3 (1967).
- ³³Bunker and Cohen, p. 5.

CHAPTER THREE

REVIEW OF THE LITERATURE

There are some general comments about the training evaluation literature with which it would be appropriate to orient this discussion. The first is that training evaluation seems to be a very sensitive subject among industrial trainers. No doubt, this is because it is done far too infrequently, and is usually done less than perfectly. Compared to many of the demanding prescriptions for "correct" evaluations, it appears that very few in practice measure up. This large gap between theory and practice apparently is the prime reason training evaluation is such a sensitive subject.

There seem to be two schools of thought in training evaluation which ignore each other. One is the scientific, rigorous, experimental school. This group sees training evaluation as a carefully controlled, detailed, almost academic endeavor. They seem to be unaware of cost constraints, organizational politics, and the environment in which industrial training occurs. The other group is the practical, "quick and dirty" school. Evaluation to them should be quick, fairly inexpensive, somewhat subjective, and not without a good deal of intuition. They are

very aware of bottom-lines, time and money, top management, and organizational exigencies. However, they are not too concerned with control groups, rigor, intervening variables, and experimental studies.

The comments above appear throughout the literature on training evaluation and represent some of the unresolved problems which exist. The literature also includes many studies which report the use of any number of techniques of evaluation, most peculiar to each situation, along with the results of the particular evaluation. In general, these results and techniques are not reported in this study. The broader, more important questions seem to be those concerned with the concept and application of training evaluation across many situations.

These questions or issues seem to group together according to the logical progression that one would follow if he or she were called upon to perform a training evaluation: formulation, implementation, and use. This is the pattern used in grouping and discussing the various issues that impinge on the effective evaluation of training.

ISSUES IN TRAINING EVALUATION: FORMULATION

The issues involved in the formulation of training evaluation strategies are the most crucial in the entire evaluation process. These issues involve the goals, philosophy, rationale, direction, and priorities of a particular training evaluation--in short, the focus of the evaluation. If this stage is not carefully considered, much of what follows in the design and implementation is potentially useless and valuable resources are wasted. Careful study at this stage of evaluation can uncover future problems, direct energy to the right areas, and even identify whether an evaluation is feasible or desirable.

Goals of Training Evaluation

Alden indicates that the reason for failure of most evaluation efforts lies in their lack of focus.¹ Evaluators measure many different things, without having decided what is important and what will be used as input for managerial decisions. Consequently, much of the data gathered answers the wrong questions, or answers no questions at all. Before an evaluation is conceived, much less implemented, a clear decision should have been made as to what the purpose of the training evaluation is, and to what use it will be put.

One area of input into this decision reflects the different rationales involved in training evaluation described on pages Whatever reason is behind a particular evaluation should be clearly specified as part of the formulation process, as this will have implications for the type of evaluation that is carried out. For example, an evaluation to justify resource expenditures will employ different criterion measures than a "refinement" evaluation. A justification evaluation will have to identify very specifically the desired change, and will have to supply data that indicate that that specific change did or did not occur. In this type of evaluation, a fairly rigorous design is necessary and data on the fine points of evaluation is not necessary; the entire program is being looked at as a whole, rather than being dissected.

On the other hand, in a refinement evaluation, it is necessary to gather more microscopic data. Data for individual questions, subject areas, learning experiences, etc., should be made available for analyses, along with how different groups of trainees scored. In addition, trainee suggestions or reactions are often helpful to supplement the

other information; this information can be combined to diagnose problems in training.

Finally, if the main expected outcome of evaluation is to enhance credibility for training, this should be clear to whomever is formulating the evaluation. The main focus of this type of evaluation would be a highly visible, acceptable, and credible evaluation; as strange as it sounds, the results themselves would not be as important as the communication of those results. One way to establish credibility for evaluation is to make it clear that evaluation results are being used in decision-making. A training program with public relations, or image problems would perhaps have this type of evaluation as its goal, and should be particularly sensitive to the appearance and use of its evaluation. Many of the industrial training programs which use "smile sheets" or participant feedback as their technique clearly want it to help establish credibility for their training.

As can be seen, different goals of evaluation have different implications for the strategy and focus of evaluation. However, in practice, often goals are not stated explicitly and evaluation ends up trying to do several things at one time. Chabotar identifies this as one crucial element of evaluation: deciding on the questions evaluation is supposed to answer.² Limiting the evaluation to one main question with one or two secondary questions will serve to make the evaluation more effective in meeting those stated goals.

Also with implications on the goals of training evaluation are the recipients of the evaluation data. Thompson identifies three main audiences of evaluation and says that these audiences to a large degree provide answers as to the goals of evaluation.³ The three main audiences

that use training evaluations are the trainees, the organizational decision-makers, and the trainers. If the evaluation is for the trainees' benefit, the main goal is to show whether the training made a difference both in the desired change of training, and in the performance on the job. The trainees want to know if they got their money's worth. The evaluation should be geared to making clear the evidence of benefits (or lack of same) that have accrued to the trainee through training.

If the organizational decision-makers are the audience of the evaluation, there are different goals. Like the trainees, they are interested in whether the training program made a difference. But they are also interested in what effects these changes had on overall organizational objectives. If human relations training, for example, made supervisors more satisfied and gave them better relations with their employees, did it actually improve or decrease the unit's productivity? In addition, the organizational decision-makers are interested in the effects of training relative to the costs. While the training may have been shown to be very effective, it may also have been very expensive. Finally, this audience is interested in the generalizability of the evaluation results to other company training programs. Clearly, this audience demands much broader and more extensive evaluation, with wider implications than does the trainee audience.

The audience of trainers themselves have different goals for evaluation. First, they need to know how they are doing. This type of feedback orients them, helps morale, and gives information as to which training programs are working best. Second, a goal of evaluation for trainers is assessment of current training needs and improvement of training programs. Knowing whether training was successful or not is

not enough; trainers need to know how their programs can be improved or refined, and need to know the current state of training needs among the organization's employees. Third, a goal of evaluation is to provide information on whether a training program can be expanded or developed within different subsystems in and out of the organization.

The audiences, or users, of training evaluation must be clearly identified in the process of establishing goals for evaluation. It has strong implications in the choice of evaluation method, desired information, and use of the evaluation data.

Finally, in the consideration of the goals of the training evaluation one must consider the nature of the training being evaluated. Although it is generally assumed that industrial training is undertaken for the purposes of improving employees' job performance and ultimately the organization's performance, there may be other reasons for training as well. It has already been pointed out that such intangible goals as social responsibility, prestige, and organizational ritual may be part of the motivation underlying organizational training. While these goals would indicate that organizational objectives are as important as in the standard goals of training, they have different implications for evaluation. If a main goal of spending on training programs is actually for prestige or image purposes, then an evaluation on performance criteria would be extraneous and a waste of money. If an organization were conducting training out of a sense of social responsibility, evaluation would perhaps be inappropriate. Or the goal of evaluation, on the other hand, might be to measure the adjustment, happiness, family life, or job mobility of the trained employee--far different from the goals of evaluation conducted against performance criteria.

Obviously the percentage of training programs conducted for the tertiary reasons stated above is quite small, and the percentage that would admit those reasons as stated goals is even smaller. However, it is a factor that does exist in industrial training, and should be included in an overall consideration of training evaluation.

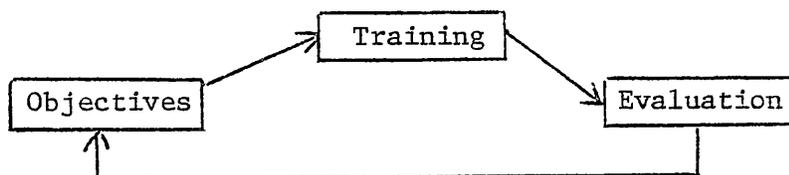
Needs Assessment/Evaluation Interface

Before any training program is developed, some form of needs assessment has been performed. As with evaluation, this needs assessment may be informal or formal, implicit or explicit, good or bad, but someone has assessed the organization and decided that a training need exists. The evaluation of the subsequent training and the original needs assessment, in theory and in practice, depend to a large extent on one another. The connection between needs assessment and evaluation occurs in at least two ways.

The first is that objectives implied in the needs assessment become the criteria used in the evaluation. A needs assessment will survey a current situation (using any of a variety of methods) and establish that certain training needs exist; if the assessment is a good one, these needs will be specific, tangible, measurable, and changeable. The measures, or criteria, that are used in the initial assessment then should be used in the subsequent evaluation, as the training should be evaluated against what it was supposed to do--not against a concept that is independent of needs assessment. Lawrie and Boringer recognize this interaction and suggest that the criteria used in assessment be used in evaluation: the assessment "develops within its own process the necessary criteria for its training evaluation."⁴

However, this process means that the criteria used in needs assessment are particularly important. If training is based on faulty or irrelevant criteria, not only will the training be misdirected, but so will the subsequent evaluation. Training cannot be evaluated on goals for which it was not designed. Unfortunately, "this first step (pre-specification of a goal or goals) is too often given too little attention, thereby imposing a ceiling on the quality of what evaluation research follows."⁵

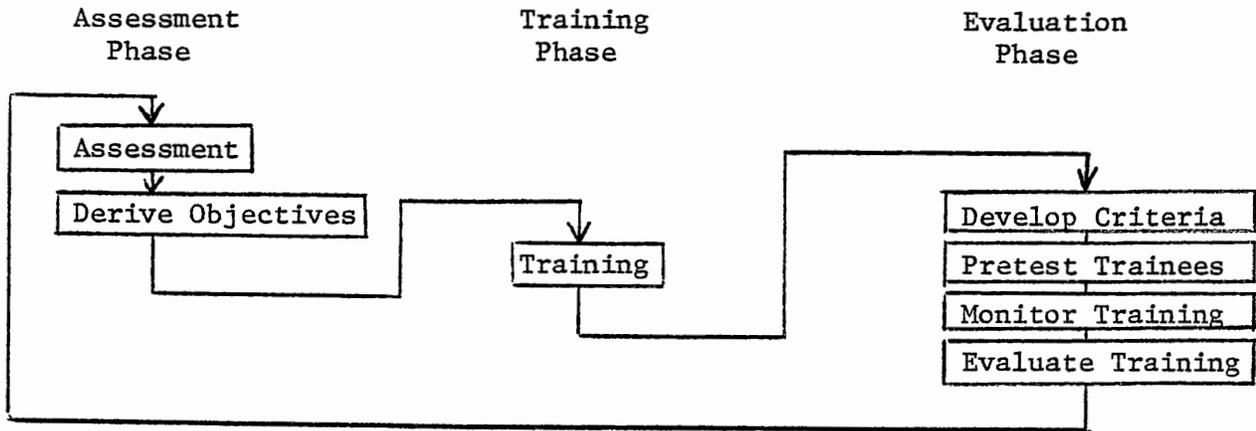
The second possible connection between needs assessment and evaluation is that they become one and the same process. This concept is developed by at least two theorists, who show similar models of the entire training process. Hamblin sees the entire training process as follows:



The objectives initially dictate what form the training will take. After the training, evaluation is conducted against the objectives specified earlier. This data then becomes the input (or needs assessment) which determines both the next cycle of training, and the criteria for evaluation. From then on, evaluation and assessment become one function, the output (evaluation) acting as the input (objectives). In this model, Hamblin allows for outside influences on the system, both at the evaluation and objectives level.⁶

Goldstein derived a similar model for training (following page). Although the appearances of the two models are somewhat dissimilar, the concepts are the same. First, the objectives in assessment find their

way into evaluation as criteria, and second, the data gathered in evaluation is recycled into the system as needs assessment data.⁷



Indeed, once the system is initiated, the evaluation and assessment phases actually become the same process (or at least, "flip sides of the same coin"). As long as the criteria remain fairly stable, this system ensures that training is evaluated against what it was designed for, and that an accurate picture of existing needs is available.

Level of Evaluation

An issue that follows logically after the issue of needs assessment criteria is that which seeks to select the appropriate criteria to use in evaluation. The issue of selecting criteria is discussed widely in the literature, and is one of the biggest problems in formulating and implementing training evaluation. In 1959, Kirkpatrick proposed a way of conceptualizing criteria into four levels of complexity and/or abstractness.⁸ This concept is cited widely in the literature and seems to be broadly accepted. Although other schema have been proposed, they generally follow Kirkpatrick's concept.⁹

Basically Kirkpatrick separates the possible criteria for evaluation

into four areas: reaction, learning, behavior, and results. Each of these categories is farther removed from the actual immediate results of training than the previous one. But each more closely approximates the goals of the organization than the previous one.

The first category, reaction, involves criteria which evaluate training on the simplest level. This criterion usually asks the trainees to rate the effectiveness of the training; it is a purely subjective rating on the part of participants. Commonly known as "smile sheets" or "happiness scales," this type of criterion tells nothing directly about the performance effects of training, except a best estimate on the part of untrained raters. About this type of criterion, Blumenfeld and Holland comment: "Unless the purpose, the pre-stated goal of the training experience was to modify training attitudes toward the course, so what?"¹⁰ On the other hand, this type of data does give some key information as to how closely the employees perceive that training is meeting their goals. In addition, if the ratings are taken seriously by the trainers, this type of evaluation can help enhance the credibility of the program.

The second level of criteria is learning. This type of criteria tries to determine how much cognitive knowledge was gained as a result of training. Various types of tests usually typify this level of criteria including essay, objective, job-simulation, and programmed instruction tests. While simply gaining knowledge is usually not the ultimate objective of industrial training, the cognitive element is often necessary. This type of criterion is usually fairly measurable, and if the connection between job knowledge and performance is clear, learning criteria are well adapted to industrial training evaluation.

Evaluating behavior is the next level of criterion. Obviously, this more closely approximates the ultimate goal of training, which is usually to effect changes in job performance, or behavior. This type of criterion attempts to measure actual changes in on-the-job behavior which have occurred as a result of training. While behavior can easily be traced to the stated goals of training, it is usually difficult to measure, and it is difficult to isolate the effects training alone has on behavior. With this type of criterion, again subjective and sometimes vague measures have to be employed to collect data.

The fourth criterion is that of "results" or "impact on the organization." This type of criterion is sometimes called "ultimate" criterion; it involves the extent to which ultimate organizational goals were affected by training. Such ultimate organizational goals include earnings, gross sales, company growth, organizational reputation, and quality of service offered by an organization (for non-profit organizations). Although this is clearly the criterion that every company activity attempts to impact, it involves such a macro scale, that it is very difficult to gauge the effect of one unit's activities on, for example, the bottom-line profit.

Clearly, as one moves from the end of the spectrum at reaction, to the end of the spectrum at results, there is a trade-off between measurability and applicability in the criteria. The criteria nearer to reaction are more easily measured, but their connection to ultimate impact on the organization is unclear. The converse is true at the end nearest results; the applicability of the criteria to desired goals is great, but the accuracy of measurement is very much in doubt. Thus it is important at the formulation stage of evaluation to decide what

criteria are important, carefully measure the trade-off between measurability and applicability, and be aware of the shortcomings of any criterion used.

Although the Kirkpatrick scheme is widely used, Newstrom points out several insidious assumptions that have appeared along with the popularity of the model.¹¹ Two are germane to this discussion. The first assumption is that the criteria are arranged in increasing value of the information they provide, from reactions to results. It was pointed out above that there is a trade-off in using each of these criteria. While results criteria may appear to be most valuable, because of the imprecision of the measurement, they may in practice be almost completely useless. Thus each situation must dictate which type of criterion is most valuable.

The second assumption is that "there is a high sequential inter-correlation among the criteria." If reaction is favorable, learning will have resulted, and if learning has resulted, behavior will change and organizational results will follow. Newstrom points out that there is no theoretical or empirical evidence to back up the universality of this assumption, and indeed there may not even be any common sense to it. He indicates that it is possible for all four criterion types to be used, and for conflicting data to be obtained in all four categories. In this way, he points out the importance of establishing and justifying the criteria before an evaluation, and the need to use caution in drawing conclusions using the Kirkpatrick scheme.

Degree of Rigor

One of the most hotly debated issues relating to the formulation of

training evaluation is the degree of rigor demanded in the evaluation. In other words, how carefully controlled, precise, and insulated from outside effects does evaluation have to be to be considered effective? This question needs to be fairly well decided in the formulation stage. Although circumstances in implementation necessarily dictate, to a degree, the precision possible, a limit should be set as each evaluation is being formulated, to establish the absolute minimum standards of rigor. As is mentioned earlier, the discussion concerning evaluation rigor tends to revolve around two schools of thought: the experimental school and the practical school.

i. Experimental School

The experimental school is typified by Hinrichs' statement: "We take the strong position that attempts at training evaluation which are not experimental, do not have pre-post measures, and do not have adequate control groups, are not really evaluations."¹² This view is shared by a number of authors writing conceptual or theoretical articles on training evaluation, and by only a few authors who present the design and results of such evaluations actually conducted in industrial settings. The advantages of this approach are clear: intervening variables, or contaminants, are eliminated or severely restricted, the criteria are extremely valid, and the results are extremely reliable. Because of these, the evaluation is bound to have increased credibility, and a very accurate picture of certain effects of training is obtained.

However, there are at least three disadvantages to this approach. First, obviously this type of evaluation is very expensive in terms of time, money, resources, and manpower. If resources are plentiful and/or the evaluation is extremely crucial, this may not be a concern. In most

cases though, time and money are influential constraints. Second, the finer and more microscopic the design, usually the less the resulting data can tell you. In other words, it is difficult to perform an experimental evaluation with other than immediate criteria. Although the data may be very reliable, their applicability may be limited. Finally, rigorous experimental evaluation does not lend itself well to the types of evaluation trainers find most valuable--refinement evaluation. It is unlikely that experiments could be performed in each of the small areas that trainers need information on to improve or refine their training. Also, the results of experimental evaluation tend only to indicate whether a change has occurred, not why, how, or sometimes even how much. These latter questions are the crucial ones in providing data on how to improve training programs.

ii. Practical School

The second school is the "practical school." They are extremely aware of the situational constraints which exist, but yet feel a need to get some evaluation done. Contrary to Hinrichs, Hamblin asserts that evaluation is conducted with or without experimental design: "the trainers, the trainees, and everyone else connected with training are continually evaluating it in terms of criteria which may be explicitly stated, or may be only vaguely formulated in their minds."¹³ He goes on to argue that the goal of evaluation is to improve, as best we can, the evaluation that is always taking place on some level.

It is interesting to note that the "practical" type of evaluations reported in the literature are by far more numerous than "experimental" evaluations, despite the prescriptions. This seems to indicate several things. Probably, training professionals are far less schooled in

experimental design and statistical analysis than academicians and researchers. Trainers are operating in an environment where results, rather than scientific "truth," are important. The environment probably also limits the latitude trainers have in manipulating situations to conform to a true experimental design. Finally, trainers seem to feel a need to do something in the way of evaluation, even if it would be rejected for lack of validity and precision in an academic environment.

The decision on the degree of rigor must be made on the basis of several factors. First, an evaluation can only be made within the existing constraints of time, money, and organizational authority. However desirable a true experimental design may be, it simply cannot be done in many instances because of the factors stated above. This does not mean that trainers must settle for useless or nonexistent designs; the goal should be to develop the most useful evaluation within the existing constraints. Second, the objective of the evaluation indicates the degree of rigor that is important. If evaluations, for example, are mainly for image purposes or to increase cooperation with trainees, rigor may not be of paramount importance. If an entire training program is being investigated to see whether it actually made any difference on a specified criterion, a more rigorous design would be appropriate. The degree of rigor must correspond to the goals of evaluation, and to the practical factors which impinge on the situation.

Financial Considerations

An area that would seem to lend itself well to training evaluation

is financial evaluation. Especially in for-profit organizations, a salient question is: in what ways did the training impact us financially? As a training evaluation is being formulated, one decision that has to be made is the extent to which financial data will figure into the evaluation. Surprisingly, an emphasis on financial considerations is not particularly widespread, although some authors take note of its importance.

Odiorne, for one, advocates the evaluation of training in economic terms.¹⁴ Horrigan suggests using turnover as a criterion in training evaluation mainly because it provides cost justification data for training.¹⁵ Mindak and Anderson have devised a method for approximating the financial benefits of training of key people, and capitalizing those benefits as assets.¹⁶ However, the majority of authors do not consider financial considerations of prime importance for several reasons.

The first reason is that training is often considered a necessary expense, not an investment from which a return is expected. Bunker and Cohen point out that this is often an excuse for not using evaluation at all; training is simply a human resource cost item, a necessary evil.¹⁷ Thus, the cost of training is taken for granted, and it is assumed that the financial return on training could not possibly be estimated.

Another reason financial data is not considered is that it is too difficult to gather. This data needs to be gathered in two areas, cost and results. The costs are not too difficult to measure. Any good accounting system could keep and compile records on the direct costs of training. This could be a problem however because training costs sometimes are charged to departments or users of training, rather

than to the training departments. Also the "opportunity costs," or monetary sacrifices that a unit has to make for training, should be considered. Garbutt discusses this problem at length;¹⁸ these types of costs are usually attributed to loss of productive time, or lower productivity while employees apply their learning.

The benefit side of financial considerations is by far the more difficult to measure, and herein lies the main reason cost/benefit analysis is not advocated more widely. The financial effects of training, especially management development training, are so diffused and delayed, that it is very difficult to conceive of a way to measure them. Financial results of training are almost invariably included in ultimate criteria. These type of criteria, as was discussed earlier, include so many intervening factors that it seems impossible to trace financial results directly to training. One notable and novel exception is the suggestion by Horrigan cited earlier, that turnover figures for trainees be compared to non-trainees. Using an established industry dollar figure per turnover, one can estimate both the savings in turnover dollars and the training effort retained in the firm. Although somewhat of a backward way of approaching the problem, it does offer a unique application of financial data to training evaluation.

ISSUES IN TRAINING EVALUATION: IMPLEMENTATION

Unquestionably the most difficult decisions and the knottiest problems in training evaluation are confronted in the formulation phase. It is in that phase that the direction, purpose and rationale for the evaluation are established. However, there are a number of concerns

remaining in the design and implementation of the evaluation, and in the resultant use of the data. The implementation issues are discussed in this section.

Danger of Self-Monitoring

One of the most curious features of the overwhelming majority of training evaluations is that they are done by training departments themselves. The integrity or the good intentions of training departments are not being called into question here. Rather the organizational control for monitoring a unit does not typically lie with that unit. Of course, if the goal of evaluation is purely to gain information on training in order to improve, this is not so much of a problem. However, in the case of justification evaluation, at attempts to measure effectiveness, there are implications for this type of self-monitoring.

Randall indicates that evaluation done by the training department suffers from "survival bias" or the inherent desire to present oneself in as favorable a light as possible, especially when one's value is being questioned.¹⁹ Depending on one's point of view this may or may not be a problem. From the point of view of the training department, this is simply what the evaluation is for: survival. Furthermore, any evidence that can be gathered is good ammunition for defense, and no one is expected to report how poorly they have done. Also, if top management asks the training department to evaluate their own effort themselves, then top management must understand the implications of such a request.

All these arguments have validity. However, in interpreting the results of such a self-evaluation, these conditions must be taken into

account. And further, as such an evaluation is administered, those involved with designing it must be aware of how those results will be viewed. If a high degree of accuracy and objectivity is desired, a self-evaluation is probably not the route to go. In addition to the possible slant of the resultant data, the credibility of the entire effort may suffer.

Choice of Measure

The choice of criterion in the formulation phase of evaluation is followed by another choice in the implementation phase. How will that criterion be operationalized? What will be used to measure it? This issue is of considerable concern, as evidenced by its presence in the literature. Primarily the validity and reliability of the measure need to be assured, and any effect the measure will have on the training or evaluation needs to be identified.

Once a desired criterion has been chosen, there are any number of measures available to gauge the effect of training on that criterion. Examples at each of the four levels of criteria are: reaction, measures of trainees' satisfaction with training; learning, paper and pencil objective tests; behavior, on-the-job rating of specific behaviors (usually done by trainees' supervisors); results, unit or organization's periodic earnings figures. Each of these measures, to be useful, must have established internal and external validity.

Internal validity is concerned with the extent to which a measure is able to identify the effects of training on specified criteria. In other words, is the measure measuring what it says it does, with respect to the training experience? For example, if the criterion was learning

in an electrical engineering training class, does the test actually measure the concepts that are covered in the course? If the test measured electrical engineering concepts that were not covered in the course, the measure would be invalid. Thus an internally valid measure would be able to measure the effects of training on desired criteria, that are directly attributable to employee training.²⁰

External validity asks a much broader question: does the measure detect changes that are the desired ultimate changes of training? This issue involves assumptions regarding the goals and efficacy of training. While the rationale behind training might be to increase the trainees' knowledge, which in turn will make him or her more skillful, which will enable him or her to perform better on the job, a simple measure of knowledge will not necessarily indicate the desired change in job performance. This is the connection that external validity seeks to make. External validity may be obtained by either accurately measuring the ultimate objective (e.g., job performance) or by providing evidence that the measure (knowledge test) and the ultimate objective (job performance) have a high correlation. Thus, with an externally valid measure, "one can generalize an employee's training performance to actual on-the-job performance."²¹

A measure can be neither internally nor externally valid unless it is reliable. A measure is reliable if it will measure the same thing the same way across time and situation. For example, an objective test should give basically the same results if it is given to the same person two separate times, if all other things remain equal. Or, an objective rating scale should yield the same results if the same person's behavior is rated on that scale by two different raters. The concept behind

reliability is consistency, and it takes mainly a diligent and conscientious effort to ensure reliability. For instance, with objective test, reliability requires the same testing situation, time allowances, reference sources, and testing instructions.

The other effects of the measure on evaluation also need to be considered. The well-known "Hawthorne effect" can tend to exaggerate experimental changes detected on a measure because of the trainees' perception of being involved in a "special activity." Rather than the satisfaction which resulted in the Hawthorne studies, it would seem more likely that in evaluation situations, the trainees would feel on the spot, or under scrutiny. This will be discussed in more detail in a later section, as will the effect of the pre-evaluation measure on the training process.

Experimental Design

This section involves putting the decisions about the acceptable degree of rigor, into effect as actual experimental designs. Campbell and Stanley identify three types of research designs: experimental design, quasi-experimental design, and pre-experimental design.²²

Obviously the most desirable type of design--all other things being equal--is a legitimate "experimental" design. In this type of design, internal and external validity, reliability, and absence of contaminants are assured, and the impact of experimental manipulation can be clearly and accurately identified. There are two outstanding features to a true experimental design, pre-post measures and control groups. The simplest type of experimental design uses two groups of subjects, which are matched and are presumed to be equivalent. The selection of the groups is either through purely random selection of the members of

each group, or by the "stratified" method in which the makeup of the population is known and types of subjects are selected for the groups in proportion to their occurrence in the population. Then, each group is pre-tested, and one group is given the experimental treatment (training, in this case) while the other is not. Finally, both groups are post-tested, and the disparity between the pre- and post-test difference scores is presumed to be the result of experimental treatment. In this design, the passage of time, random fluctuations in scores, and other extraneous factors are controlled.

However, the actual taking of the pre-test may have increased the impact of the training, or increased scores on the post-test, or both, and this factor is not controlled in the design described above. Thus, the ideal experimental design involves the use of four groups, two of which are pre-tested and one of which is trained, and two of which are not pre-tested and one of which is trained. The following chart outlines this design:

| <u>Group</u> | <u>Pre-test</u> | <u>Training</u> | <u>Post-test</u> |
|--------------|-----------------|-----------------|------------------|
| 1 | yes | yes | yes |
| 2 | yes | no | yes |
| 3 | no | yes | yes |
| 4 | no | no | yes |

All four groups in the Solomon four-group design, as it is called, are post-tested. Thus, the effect of training, the effect of pre-testing, and the effect of the interaction between training and testing, can all be ascertained. This is the most precise and most desirable experimental design.²³

The second group of research design is the quasi-experimental

designs, which are slightly less rigorous than experimental designs but "are useful in many social-science settings where investigators lack the opportunity to exert full control over the environment."²⁴ In this type of experimental situation, since less control exists, more inferences must be made in drawing conclusions. One example of this type of research design is the time-series design, in which measurements of one group are taken periodically over a length of time; somewhere in the middle of that length of time, the training is administered. Thus we not only have data on the changes that occurred immediately after training, but also the changes that occurred for a length of time before and after training. In addition to gaining information on the effect time alone has on the criterion variable, we can also see the long-term effect of training.

Another type of quasi-experimental design that is considered particularly desirable when experimental designs are impractical, is the un-pretested two group comparison.²⁵ In this design, two groups are selected to be as equivalent as possible, possibly through random selection. Then, one group is trained while the other is not, or the two groups are trained with different programs. Finally, the two groups are measured and compared on some criterion measure. This design eliminates the pre-test sensitization bias and is fairly easy to design and administer.

There are a number of other types of quasi-experimental research designs. The feature common to all of them is that there is some causal element of variability that is not controlled. For example, in the time-series design, we do not know of additional circumstances occurring simultaneously with training, that may have affected the measurement

immediately after training. However, quasi-experimental designs are considered to have enough control to make their results useful and interpretable.

The final type of research design is called pre-experimental design. In this type of experimental design, very little control exists, data may be collected as few as one time, and drawing conclusions is tenuous, if not impossible. This type of design includes the "one-shot case study," in which one group is trained and post-tested. One of the flaws in this design is that there is no pre-measurement with which to compare the post-test; furthermore, even if there was a pre-test, we have not eliminated any extraneous factors occurring between the pre- and post-tests which may have affected the scores. Finally, we have not controlled for the effect of the testing instrument itself. As Goldstein points out regarding this type of research design, "the only bases for comparisons are intuitions and impressions."²⁶

Both the dangers of using pre-experimental designs and the advantages of using true experimental designs are clear. If the premise of a training evaluation is scientific study, then the rationale for using any of the pre-experimental designs is shot through with holes. No evidence for their inadequacy is necessary; the logical arguments raising doubts about their validity are strong enough to stand on their own. The efficacy of using the sophisticated types of experimental design in evaluation is demonstrated by the results of Belasco and Trice. Using the Solomon four-group design to evaluate the impact of supervisory training, they discovered that "in contrast to training, our information indicated that the completion of the questionnaire alone, was a far more effective change agent."²⁷ In a setting where effecting

the desired change most economically by whatever means is the goal, Belasco and Trice's results offered valuable information. It is doubtful that Belasco and Trice could have obtained evidence for their conclusions without using the experimental design that they used.

There is one strong qualification to the entire preceding discussion however. The discussion above is predicated on the condition that training evaluation is conducted as scientific research; in practice, this is quite often not the case. The goal of industrial training departments is typically training, not research. Evaluation of training is a relatively recent phenomenon and evaluation of training as research is even more recent. Consequently, the freedom to design and conduct rigorous experimental research is seriously hampered in many evaluation efforts. Furthermore the goals of the evaluation may not be entirely congruent with those of scientific study. It is in the light of these two considerations that the choice of experimental design for training evaluation should be seen; these considerations are discussed in more detail in the following section.

Organizational/Environmental Constraints

Experimental research is defined by Campbell and Stanley as "that portion of research in which variables are manipulated and their effects upon other variables observed."²⁸ This type of research which involves pre- and post-testing, control groups, and training as an experimental treatment, is advocated frequently in the literature on training evaluation. Bunker and Cohen describe the application of the Solomon four-group design to training evaluation:

Although logic might indicate that a half-hearted evaluation effort is better than none at all, it is our

contention that 'going the extra mile' and conducting a comprehensive and well-controlled investigation may be more mandatory than optional, more investment than expense, more income than cost.²⁹

The manipulation of variables, especially ultimate criteria variables, is impossible in many training situations, which obviates the choice of experimental design. The ultimate criterion of results, even if it could be controlled, could not freely be manipulated simply because the environment is a profit-making organization, not a social science laboratory. Quite often, there may not be an opportunity to have a control group because management wants everyone trained. It was mentioned earlier that evaluation is often not done because management is convinced of training's effectiveness; certainly this attitude would engender even stronger resistance to using experimentally rigorous research designs. Finally, management may not understand the argument for scientific evaluation, or may understand but not agree with it. In short, the applicability of true experimental designs is limited in industrial training evaluation by the environment in which the evaluation is done.

Additionally, the goals of scientific investigation do not always coincide with those of a training evaluation. Campbell et al. point out a "narrowness" in the discussions of experimental design. They indicate the difference between "statistical significance" and "behavioral significance": while training may have a statistically significant effect, the effect in bottom-line terms may be so small as to be of no consequence to the organization.³⁰ Conversely, the training may have resulted in changes considered significant to the organization, but which were not detected as statistically significant. While the

nature of the scientist is to be cautious, the businessman is by nature a gambler. Thus they have different requirements for precision and certainty.

In spite of the constraints limiting certain situations from using strict experimental designs, the design of evaluation research can be improved. Although all extraneous factors cannot be controlled, some can. Goldstein comments that, while strict experimental designs often cannot be employed, with little extra effort the quasi-experimental designs could be employed.³¹ Giving closer attention to the most important goals of the evaluation and recognizing factors intrinsic to the situation, can help to identify those factors which can and should be controlled. Each situation is unique, and a thorough knowledge of the situation will give clues as to which variables are most likely to be explainable if they are not controlled.

Finally, the industrial training environment has a bias toward efficiency. All things being equal, the evaluation that is most economical is most strongly preferred. Combined with the frequent lack of demand for rigor, this factor suggests that short-cuts, patch-ups, and other less-than-perfect designs are acceptable in industrial training evaluations. This does not mean that meaningful results are not desired; it means simply that economy is a high priority. For example, Brethower and Rummler suggest taking advantage of "naturally occurring research opportunities" in implementing training evaluation.³² This type of opportunity usually lends itself well to quasi-experimental designs, such as a time-series design when program management or program content is altered. Also, to save resources, several functions in the training process can be combined, if the situation allows. It was

suggested earlier that needs assessment and training evaluation can both be achieved in one step, both in theory and practice. Further, research cited earlier has indicated that testing alone can have stronger positive effects than training. In certain situations, testing could possibly take the place of some training sessions, thereby streamlining the entire process.

Finally, the utility of formal training evaluations must be considered in light of the entire situation. Hinrichs, for one, indicates that "there may in fact be more benefit from a careful pre-determination of training needs and task analysis prior to the design of training programs than there will be from elaborate broad designs attempting to evaluate results."³³

Communication Problems in Implementation

Concomitant with the technical and administrative problems involved in the implementation of training evaluation, are communication and perception problems. Evaluation, by its very nature, is questioning the value of someone's effort. Thus, someone has something at stake when training is evaluated. While the threatening nature of evaluation can never be entirely removed, through wise administration of the evaluation, it can be reduced. The consequences of not addressing the perception of evaluation as negative, are reduced cooperation, perhaps inaccurate data and increased resistance to the evaluation effort.

Bunker and Cohen observe that tests can arouse a fear of failure in industrial trainees, who perceive that their career depends on success or failure in training.³⁴ While the trainees' perception may or may not be true, the perception exists. The answer to this is to communicate

clearly the purpose of the testing; if it is not for individual appraisal, it should be clearly packaged and communicated as such. This step, if administered consistently, should alleviate much of the anxiety on the part of the trainees and improve the image of the testing process. The same reaction to evaluation is likely to occur with the trainers. In this case, however, the perception is more likely to be accurate; there is a good possibility they are being evaluated. However, this is not a communication problem primarily because the trainers' appraisal is no doubt part of their job.

The subsequent use of the data causes counter-productive perception problems also. This type of problem may occur in any type of data-collection: "If data collection in the past have been used only to expose, probe, and manipulate, enthusiasm for future involvement is likely to be moderate or non-existent."³⁵ In addition, if training departments, through the evaluations, tend to take credit for success, and lay blame for failure, a serious communication problem is likely to arise in the future. Management will resist and discredit both training and evaluation.

Thus the promotion and packaging of a training evaluation needs to be implemented with sensitivity to the way the program is perceived. A clear statement of the purpose of the evaluation and the intended use of the data should be presented, and consistently followed. Especially if evaluation specialists are dispatched from a distant home office to field locations, the goals of the evaluation, its consequences, and who will be affected, should be clearly communicated. Even the most technically flawless training evaluation can wreak havoc in an organization if attention to communication needs involved in the program

is neglected.

ISSUES IN TRAINING EVALUATION: USE OF DATA

As was briefly touched on immediately above, the use of training evaluation data has important implications in the entire evaluation process. Although these implications are not nearly as extensive as the issues involved in formulation and implementation, they can dictate to a great extent the worthwhileness of the the evaluation effort. There are two basic considerations involved in the use of training evaluation results, which are discussed below. While neither is particularly difficult to understand or to put into practice, the success of an evaluation effort hinges on these two factors. Neglect of either or both can effectively eliminate any positive outcomes of training evaluation.

Prompt Feedback of Results

So much emphasis seems to be placed on the issues involved in setting up a training evaluation that often the goal is lost once the data is gathered. The result of such a situation can be that after the data is obtained, the evaluation is considered complete, and the results are not disseminated to those to whom the information is most valuable. Or, the information may be communicated, but if a lengthy period of time has passed, the feedback becomes valueless.

If evaluation is considered to be part of the ongoing training process, as is advocated by Hamblin, then the results of the evaluation must be fed back into the system, to the persons for whom the data was

gathered and to the persons whom the data impacts. A prime requisite for two way communication (that which has a feedback loop) is that the feedback information is timely and current. If evaluation information is to be used to correct and/or refine training, it should be relayed quickly, while the conditions the information describes still exist. Delays in getting this information to the decision-makers can result in distortions of the information itself, or distortions in the situations which the information is designed to correct.

In addition, the various groups involved in the evaluations all have expectations about the evaluation data. The group that the evaluation is targeted at will want to know the results of the evaluation as soon as possible. For example, trainers will almost always have a stake in the outcome of a training evaluation. To a large degree, they are responsible for the effectiveness of training and they will thus want to know as soon as possible what the evaluation can tell them. Furthermore, the sooner the trainers have the information on training's effectiveness, the sooner they can make productive adjustments or changes. In the case of a refinement evaluation, the sooner that the developers of training programs have the evaluation information, the more useful is that evaluation information. Thompson recognizes the factor of quick feedback of evaluation information to trainers as one of the main differences between scientific-oriented evaluation research and practical-oriented research.³⁶

The trainees also often have a stake in the outcome of training evaluation. If any type of tests are used as measuring instruments, there is a natural desire for them to know how they have done even if the purpose of the evaluation was not to measure individuals' performances

as such. In many training situations, too, the evaluation can be incorporated as part of the learning process. In this case, prompt feedback of results is crucial to the effectiveness of the training effort. Immediate feedback to trainees on their performance, then, is necessary to achieve the maximum value of the training program, as well as to meet the expectations implicit in the evaluation process.

Use of Results in Decision-Making

The final and perhaps most important criterion for success of a training evaluation is whether the resultant information is actually used in decision-making. This discussion could just as easily have been the first issue discussed in the formulation stage. Before an evaluation is even conceived, it needs to be determined whether those responsible for decision-making will be willing to change anything based on the results of an evaluation. If not, the evaluation should probably not be pursued. Jay Alden, Manager of Training Evaluation for Xerox, believes that evaluators should take the position, "If they won't change it, don't evaluate it."³⁷

The reasons for this are obvious. First, an expenditure of time, money, and resources will have been made for information that is not used. Unless the purpose of an evaluation was purely for show, or was a symbolic gesture to placate some organizational constituency, the rationale for spending money on evaluation is that the information would be an input in decisions. If the information is available but not used, then the effort to gather that information will have been a waste of money.

Second, whoever has gathered the evaluation information will be very frustrated. Assuming the effort has been conscientiously and

competently carried out, the evaluators will be distressed that their voices were not heeded. Aside from the personal frustrations involved, this will have ramifications for future requests to perform evaluation. If decision-makers ignore evaluators' information, future requests to the evaluators will be met with at least reluctance and at worst refusal.

Finally, if evaluation information is not actually used in decision-making the credibility of the training and of the evaluation is certain to suffer. The perceptions of the training department will be of hypocrisy and/or ineffectiveness. This would produce a worse result than if evaluation were never attempted: "the trainers are asking for our input, but never intended to use it."

If management is serious about the use of evaluation information in making training decisions, and if the evaluation is carried out well, the resultant information should be acted upon. This does not mean anything will necessarily be changed; the evaluation may indicate no change is necessary. However, the information should be primary in making such decisions. This points out clearly how important careful consideration of all factors is in the formulation and implementation stages. If there are flaws in these two stages, management would have good reason not to use the resultant data. But if the initial stages of formulation and implementation are sound, the primary benefits to the organization of an effective training evaluation, hinge on the degree to which that information is actually used when decisions are being made.

CONCLUSION

The literature describes a number of reasons why evaluation of

training is done, and why it is not done. And in spite of ambivalent attitudes and incomplete research on training evaluation, it seems to be a topic that is becoming increasingly important in organizations. It is, however, an area in which academic theory and pragmatic implementation are at distinct odds with one another. Issues in the evaluation process from conception to completion involve the demands of both theory and practical constraints, but most authors tend to favor one perspective or the other. Clearly, a model for analyzing industrial training evaluation has to combine and integrate the perspectives of both theory and practice.

NOTES: CHAPTER THREE

- ¹Jay Alden, "Evaluation in Focus," Training and Development Journal, 32, No. 10 (1978), p. 46.
- ²Kent J. Chabotar, "The Logic of Training Evaluation," Personnel, 54, No. 4 (1977), p. 25.
- ³John T. Thompson, "How to Develop a More Systematic Evaluation Strategy," Training and Development Journal, 32, No. 7 (1978), p. 88.
- ⁴J.W. Lawrie and Clayton W. Boringer, "Training Needs Assessment and Training Program Evaluation," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 294.
- ⁵Warren S. Blumenfeld and Max G. Holland, "A Model for the Empirical Evaluation of Training Effectiveness," Personnel Journal, (1971), p. 638.
- ⁶A.C. Hamblin, Evaluation and Control of Training (London: McGraw-Hill, 1974), p. 17.
- ⁷Irwin L. Goldstein, Training: Program Development and Evaluation (Monterey: Brooks/Cole, 1974), p. 18.
- ⁸Donald L. Kirkpatrick, "Techniques for Evaluating Training Programs," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975).
- ⁹Hamblin, p. 14.
- ¹⁰Blumenfeld and Holland, p. 639.
- ¹¹John W. Newstrom, "Catch-22: The Problems of Incomplete Evaluation of Training," Training and Development Journal, 32, No. 11 (1978), p. 22.
- ¹²John R. Hinrichs, "Personnel Training," Handbook of Industrial Psychology, ed. Marvin D. Dunnette (Chicago: Rand McNally, 1976), p. 845.
- ¹³Hamblin, p. 9.
- ¹⁴John P. Campbell, Marvin D. Dunnette, Edward E. Lawler III, and Karl E. Weick, Jr., Managerial Behavior, Performance, and Effectiveness (New York: McGraw-Hill, 1970), p. 271.
- ¹⁵J. Thomas Horrigan, "The Effects of Training on Turnover: A Cost Justification Model," Training and Development Journal, 35, No. 7 (1979).
- ¹⁶William A. Mindak and Robert A. Anderson, "Can We Quantify an Act of Faith?," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 87.

- ¹⁷Kerry A. Bunker and Stephen L. Cohen, "Evaluating Organizational Training Efforts: Is Ignorance Really Bliss?," Training and Development Journal, 32, No. 8 (1978), p. 4.
- ¹⁸Douglas Garbutt, Training Costs (London: Gee and Co., 1969), p. 20.
- ¹⁹Lyman K. Randall, "Evaluation: A Training Dilemma," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 187.
- ²⁰Bunker and Cohen, p. 6.
- ²¹Bunker and Cohen, p. 5.
- ²²Donald T. Campbell and Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research (Chicago: Rand McNally, 1963), pp. 2-6.
- ²³Campbell and Stanley, p. 24.
- ²⁴Goldstein, p. 78.
- ²⁵James A. Belasco and Harrison M. Trice, The Assessment of Change in Training and Therapy (New York: McGraw-Hill, 1969), p. 30.
- ²⁶Goldstein, p. 79.
- ²⁷James A. Belasco and Harrison M. Trice, "Unanticipated Returns of Training," Evaluating Training Programs, ed. Donald L. Kirkpatrick (Madison, Wisconsin: American Society for Training and Development, 1975), p. 69.
- ²⁸Campbell and Stanley, p. 1.
- ²⁹Bunker and Cohen, p. 11.
- ³⁰Campbell, Dunnette, Lawler, and Weick, p. 283.
- ³¹Goldstein, p. 87.
- ³²Karen S. Brethower and Geary A. Rummel, "Evaluating Training," Training and Development Journal, 33, No. 5 (1979), p. 21.
- ³³Hinrichs, p. 846.
- ³⁴Bunker and Cohen, p. 10.
- ³⁵Michael L. Moore and Philip Dutton, "Training Needs Analysis: Review and Critique," Academy of Management Review, 3 (1978), p. 544.
- ³⁶Thompson, p. 92.
- ³⁷Alden, p. 46.

CHAPTER FOUR

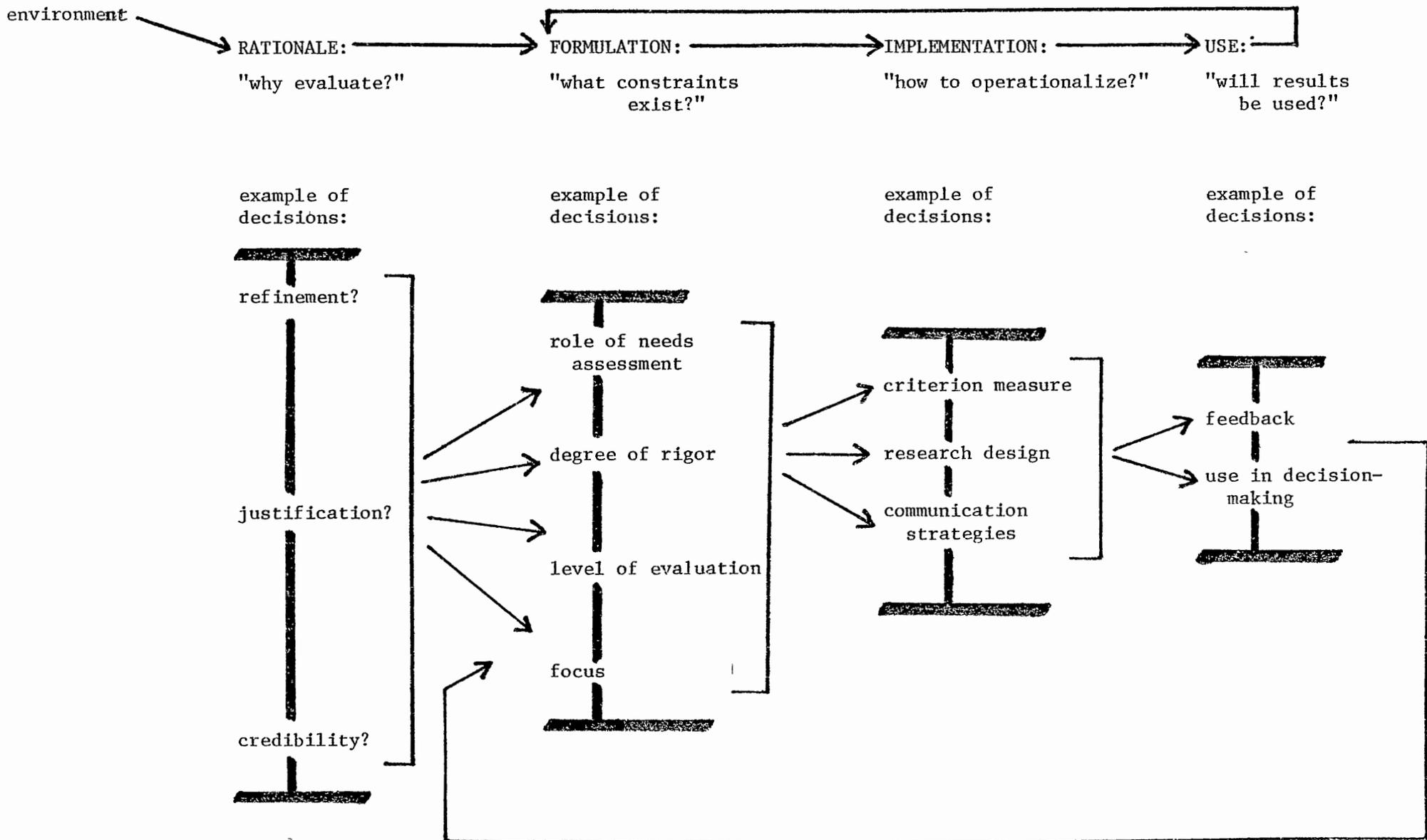
MODEL FOR TRAINING EVALUATION

The discussion of various issues involved in the phases of training evaluation may be summarized into a system, or model, for the training evaluation function. In this model, no prescription is made for the ideal example of training evaluation. Rather it is a description of the processes and factors that need to be considered in making decisions. The choice of training evaluation methods depends to a large degree on the situation under consideration. The model attempts to identify the important questions to ask about the situation, and to identify directions to take with the answers. In addition, there are a number of general characteristics which should be part of the training evaluation system regardless of the situation. These also will be outlined below.

COMPONENTS

A system of training evaluation has basically four sequential stages. The development should move from a) a consideration of the rationale behind the evaluation, to b) formulation decisions, to c) implementation decisions,

MODEL FOR TRAINING EVALUATION



and finally to d) the use of the resulting data. Normally, the decisions made in each stage would set the parameters for the decisions made in the next stage. Each stage is less abstract than the one before it, and each stage more specifically identifies actions that will, and should, be taken. If an evaluation is developed in this way, it will establish direction through a logical progression of thought and will ensure that the actual goal of the evaluation is finding its way into practice.

The rationale stage of evaluation is actually a pre-development stage, and may never have previously been made explicit. In this stage the question "Why evaluate?" is asked. The section in Chapter Two on the rationale for and against evaluation provides clues to possible answers at this stage. Possibly, the rationale for evaluation could include refinement, justification, and credibility factors. Each of these has implications for decisions that are made in later stages of evaluation. While the answer to "Why evaluate?" may seem obvious, there may be hidden, underlying, or implicit reasons for evaluation that are most influential in some situations and which are most difficult to discover.

The formulation stage takes the broad goal and rationale identified in the first stage, and attempts to translate it into a direction which is appropriate in the situation. This stage asks the question "What constraints exist which must be taken into account?" The formulation stage narrows the focus and further defines the parameters of the evaluation. For example, the demands of how rigorous the evaluation must be in this situation are considered. The importance of needs assessment as a part of evaluation is also taken into account. The viewpoint of top management, resource limitations, internal politics,

and the existing needs of the training department all help to define the direction that is established at this stage.

From here, decisions can begin to be operationalized. In the implementation stage, the nuts and bolts decisions about how to carry out the evaluation, within the constraints defined in the formulation stage, are made and implemented. This stage basically involves the important mechanics of training evaluation. It asks, within the organizational demands for rigor what is the most effective experimental design that is feasible in this situation? How will the criterion decided on in stage #2, be translated into a measuring instrument? And how can we most optimally communicate the entire program to our various constituencies? These decisions are all made, based on the parameters which have been defined in the previous stage.

Finally, the decisions about how the results of the program will be used are made in the Use stage. By this time, the decisions are limited to the data which has been made available and there is not a broad range of options available. Nevertheless, this is a crucial element in the effectiveness of the entire program. Unless the decisions made in the previous three stages are operationalized by the actual use of the information, the total effort is likely to fail. Final decisions about how the results are to be fed back to organizational members, and used in decision-making are made at this stage. Optimally, the results will be made available to trainers, trainees, or organizational decision-makers, depending on the situation, and the results will be recycled back into the system as input at the formulation stage.

KEY CHARACTERISTICS

This model implies that there are a variety of strategies that can be used for training evaluation, depending on the purposes and the constraints inherent in the training situation. But this model also implies several characteristics that should be included in any training evaluation system. Based on the review of the literature, the model outlined above should describe a training evaluation system that is planned, unified, flexible, appropriate, and complete.

The evaluation should be planned in that it follows a logical progression of development, especially in the early stages, and it should adhere to the purpose which is identified. The evaluation that is conceived without a well thought-out rationale is likely to miss its mark and waste resources. Furthermore, the evaluation should be unified, that is, each stage of its development should reflect the decisions made in the other stages. Also, training evaluation is part of the entire training system, and as such it should fit with the other components of the training system. Emphasizing training evaluation as needs assessment, and emphasizing training evaluation as an element of training itself, is consistent with this concept.

However, training evaluation systems should also be flexible. Training and evaluation are both done in response to environmental demands and constraints, and the organizational environment is not static. Hence, it is likely that both demands and constraints may change during the development and implementation of an evaluation system, and the evaluation system must also be able to change. But if a system is considered "planned" and "unified," it is often considered unchangeable.

An optimal training evaluation will balance a planned and unified approach, with flexibility and responsiveness to the organizational environment.

Any training evaluation should be appropriate to the circumstances. This is one of the basic concepts of the model presented above. Different circumstances call for different types of evaluation systems. Not only should the techniques match the situation, but the expenditures involved should also match the situation. Especially in an industrial setting, the costs or the effort involved in a training evaluation should match the importance and/or the benefits of that particular evaluation. An expensive, involved and sophisticated system should not be set up when only preliminary, rough or intuitive data is needed. Conversely, when evaluation information is crucial, and reliable and detailed is needed, the costs and care involved in the resulting system should match the needs of the situation.

Finally, an evaluation must be complete, in the sense that all stages must be followed through to completion. An evaluation which is soundly conceived and implemented, but whose results are not used or made available, is not effective. The necessity of completing the process of evaluation has been made clear in previous sections, and this is where many evaluations apparently fail. All the effort and resources that have been used to develop an evaluation system are wasted if the output of that system is not used in some meaningful way. Analogously, the effectiveness of an entire evaluation system may also be reduced if the results of the evaluation are not communicated to the constituencies which are involved.

CHAPTER FIVE

METHODOLOGY FOR EVALUATING THE TNA PROGRAM

The model in Chapter Four serves as the backdrop for the study of a program to evaluate training that has been implemented in a large corporation. The background of that organizational environment, and a description of the program implemented and the methodology used will follow in this chapter.

ORGANIZATIONAL BACKGROUND

The evaluation program that was studied occurred in a large airline corporation. This corporation is a major carrier and rates among the top five domestic carriers in volume of sales. The industry, in spite of regulation, is in a highly volatile environment and is subject to large seasonal and cyclical fluctuations in business. The gradual deregulation of the industry adds to the dynamism of the environment, as market shares, strategies, and competition are rapidly changing. In addition, technology is a major factor in the airline industry, both in the actual aircraft and also in the support functions. Obviously this also adds to the dynamic, rapidly changing nature of the airline industry.

Training plays a fairly visible role in the organization that was studied. A large training facility was developed and built in the Midwest to service some of the training needs of the corporation, although much of the training is done locally throughout the organization. The training function is separated into four areas in the organization: sales and services training, management training and development, in-flight training, and pilot training. The Sales and Services division involves airport desk personnel, cargo and baggage handlers, telephone reservations agents, ramp service personnel, and other airport support personnel. Management training and development is geared to training managers and administrators throughout the organization. In-flight training is given to the service personnel who are actually on a flight, that is, flight attendants. Obviously, pilot training is directed toward the highly specialized professionals who fly or direct a flight, such as pilots and navigators.

This study focused on a program that was instituted in the first of these four areas, the Sales and Services division. This division is included in the Department of Marketing, which is made up of slightly over 12,000 employees. Of these 12,000, about 450 are in the market analysis and planning group. The remainder, almost 12,000 employees, make up the Sales and Services division. These include 1) customer contact and sales people, such as reservations, cargo and baggage personnel, 2) technical support people, such as buildings and facilities maintenance and engineers, 3) in-flight services, and 4) finance and accounting personnel. Of these four areas, Sales and Services Training is responsible for training the first two. In-flight services has their own training operation, and finance and accounting training is not under corporate

organization.

Sales and Services Training

The Sales and Services Training organization is centered at the organization's training facility in the Midwest. There are twelve staff members there who are basically support staff for the actual training force. These twelve staff develop and communicate corporate policy, develop and plan training programs, serve as resource persons for the field training staff, and administer and evaluate training.

The actual sales and services training is done by the field training organization and local managers. There are 35 staff members, located in ten cities and international locations, who coordinate and do training. Because the responsibility for training lies with the local managers, these field trainers are assisted by local line personnel in training. With 35 staff responsible for training up to 12,000 employees, local line employees have to play a significant role in the training effort. The training is very decentralized in this organization and the role of the training organization is to support and facilitate training; the day-to-day decisions about who is to be trained, and how much, are made by the local managers.

The training that is given varies from job to job within Sales and Services, but usually involves a number of different techniques. Lectures and presentations make up one element of training that usually is given. Normally these follow a syllabus or outline that is developed at corporate training headquarters. Other "classroom" techniques used include discussion sessions, problems and assignments, and role-playing. Media are used extensively in training. Video and audio tape presentations,

films, and slide shows are developed by the corporate training staff for use in local training. Self-programmed learning packages are also developed by the central training staff for use in training for a variety of different jobs. Finally, in certain positions which require interaction with the computer, an extensive system of computer-based instruction (CBI) is available.

In addition, some formal training courses are available at the central training facility. These are typically seminar-type courses, including the two-day orientation session for all new employees. However, the bulk of training is done at local sites and the local manager is responsible for the "delivery" of the training, whether it is given by field training staff, or local line personnel.

RES Agents: the Job

The evaluation program that was studied was involved specifically with the training of reservations (RES) agents. These people are the ones to whom you talk when you call the airline telephone number to make a reservation or get information. They are located in eight offices nationwide; these eight offices handle all the telephone calls from anywhere in the United States. They are typically located in each of the eight cities in a facility which is not at the airport, but which is set up specifically to handle telephone operations.

RES agents are not face-to-face customer contact personnel, but almost their entire job involves customer contact over the phone. Each agent has a computer terminal at his or her desk with which he or she makes reservations, cancels reservations, and obtains information. There are three basic customer groups with whom the RES agent may make

contact: travel agents, business travelers, and leisure travelers.

Although any RES agent may come in contact with any of these three groups, a separate desk is set up to handle large corporate customers, charters, large travel agency orders, and other large or frequent customers.

There are three prime requirements for the position of RES agent. First, the agent must have a pleasant telephone manner. This involves not only a pleasant voice and manner, but also the ability to deal smoothly with different types of people in different situations. It involves persuasive and selling abilities, and the skill of controlling a conversation. Because time is money, literally, in this situation, the RES agent has to be efficient and quick without being curt.

The second requirement is the ability to know and manage a great deal of information, much of it very detailed. The RES agent must be familiar with fares, tickets and ticket procedures, airline rules and regulations, geographic areas, auxiliary services (rental car, hotels, etc.), airline schedules, and a host of other subject areas. Although much of this information does not have to be memorized (such as fares and scheduling), much of it (like regulations and geographic information) should be familiar to the agent, whether through reference sources, the computer, or another agent. Thus, the agent must either know the information or know where to get it quickly.

The third requirement is the knowledge of how to manipulate the computer system. Because the entire reservations system and the bulk of the information available to a RES agent have been computerized, skill in working with the computer is essential. This, like the second requirement, demands the knowledge of a broad spectrum of information. Skill in accessing the computer, accessing a certain subject, making

entries correctly, and correcting mistakes are the types of demands made by the RES job. Again, most of the information needed by an agent is contained in the computer, but he or she must have the knowledge of how to get to that information. The more knowledge the RES agent has, the less time and fewer errors will be spend attempting to discover other information.

Because of the nature of the job, the RES agent group is easily and thus highly measured. Each call that is received is timed; there is an optimal time limit (two minutes, 56 seconds) for each call. This "agent occupied time" (AOT) is recorded, compiled and averaged for each agent, unit and office, and becomes an important barometer of productivity. In addition, incorrect entries on the computer are recorded and compiled. These computer "bounces" are important because of the wasted agent time and computer time involved. Because of the high degree of measurement and visibility of productivity, RES agents are a highly competitive group. Each office is very aware of where they stand relative to other offices, with regard to productivity.

Obviously, one of the prime factors in performance as a RES agent is the possession of a good deal of very specific information. Although a highly informed and knowledgable agent would not necessarily be highly productive, he or she is more likely to be productive than an ill-informed agent. Also, while knowledge does not ensure high performance, in this case, low knowledge would ensure low performance.

Furthermore, a small increase in job knowledge could possible lead to a much larger increase in productivity, especially with regard to computer interaction. And since much of the information in the RES agent's environment is constantly changing (fares, etc.), an emphasis on knowledge and information is quite important in the job. Consequently, the

corporation recognizes that job knowledge is the prime requisite of the RES agent, and thus stresses job knowledge in RES training.

RES Agent Training

The training that a RES agent goes through involves a maximum of eight weeks of intensive training at the local site before the agent starts the job. If the agent is a new hire, he or she will have a two-day orientation session at the central training facility before the local training starts; employees who are transferring from other departments move right into RES training. The RES training follows the general pattern of other Sales and Services training outlined earlier. The training is conducted and scheduled locally, although the field training staff plays a large role in actually conducting the training. The maximum term for new agent training is eight weeks; however, the local line manager may decide if a new agent is ready before that time.

The RES agent training is highly technical, focusing on the specific, concrete areas that a RES agent will need to know in his job. As mentioned earlier, the training is very knowledge-centered. That is, the emphasis is primarily on gathering knowledge or learning how to get information. Because the initial training may be the only formal training a RES agent will receive, it is intense and covers a great deal of information in a relatively short period of time.

The training consists initially of five actual courses that a new RES agent will receive: 1)Basic Reservations Training, 2)Transactional Analysis, 3)Reservations "Sell It Like It Is," 4)PARS Automation Training, and 5)Reservations Sales Effectiveness. These are classroom-type courses (from several hours to one day) in which presentations and discussions

are conducted. As with all the Sales and Services training, there is a great deal of media presentation in these courses. Typically, the courses are conducted by field training staff.

Two of these courses (Basic Reservations Training, and PARS Automation) are ongoing training sessions in addition to the initial introduction. The Basic Reservations Training lasts the duration of the new agent's training and involves further classroom sessions, working with experienced RES agents, assigned problems, and other specialized training. The content and duration of this element of RES agent training depends to a large degree on the local manager.

The PARS Automation training is training on the computer system with which RES agents work (Passenger Automated Retrieval System); this is the central element of RES training. After the initial session the new agent goes through a number of programmed computer-based instruction courses, in which the trainee learns from the computer. The new agent is assisted by the instructor or by an experienced RES agent. The CBI system permits the trainee to work at his own pace and gain "hands on" experience with the computer. His or her progress is recorded on the computer and is available for the instructor or manager to monitor. There are 68 different training programs available in CBI for the RES agent, in such areas as "Creating Itineraries," "Pricing and Ticketing" and "Seat Assignment."

Finally the trainee uses a "double-jack" to listen in to experienced RES agents actually working with customers. This occurs throughout the training and enables the trainee to hear the things he has been learning being put into action. When the trainee is ready to receive calls, the double-jack will be used by the manager or instructor to

listen in and correct the trainee.

When the manager considers the trainee ready (sometime within eight weeks) the new agent will start work. The initial training which is decided upon by the local manager and carried out by field training instructors, will usually be the only formal training that a RES agent receives. Before the development of TNA (described below) there was no formal way to measure job knowledge or evaluate the effectiveness of training.

THE TRAINING NEEDS ANALYSIS (TNA) PROGRAM

Description of TNA

The TNA program is a set of tests designed to measure job knowledge in all areas of Sales and Services training. They are grouped by job area (Reservations, Airport Ticket Officer, Baggage Handler, etc.) and then by subject area within a specific job. For example, the Airport Ticket Officer (ATO) would have a group of test modules, which cover all areas in which the ATO is trained. Although TNA modules for all areas of Sales and Services training are projected, only RES and ATO test sets are available at this time.

Each module is designed to cover a specific area in which job knowledge is crucial to job performance, and in which the employee has received training. Modules are made up of usually 15 to 40 multiple choice questions. There are four or five possible answers, with only one correct answer. For each module, a Minimum Acceptable Performance (MAP) score has been identified, which is the level of job knowledge that is considered essential to optimal performance of the job.

For the RES agent job, there are 15 TNA modules which cover all

critical areas of job knowledge. All 15 have been developed at this point; this thesis reports on preliminary administration of these modules to RES agents. Examples of the subject areas of the RES modules are: "Domestic Flight Availability," "Passenger-Name-Record Combined Operations," "Seat Assignments," and "Auxiliary Services."

Development of TNA

In the early 1970's, the corporation adopted a computer system for virtually all reservations, scheduling, fares and other information. Although the computer had been around for a number of years, the early 70's marked a total changeover to a successful computer system. As with all major companies in the airline industry, this action caused a minor revolution in the way the company conducted their business. Especially for the company's Sales and Services employees, some of whom had been with the company for 20 years or more, automation signaled a tremendous shift in job procedures, requirements, and knowledge.

By 1974, a number of factors indicated to management that the transition to automation had not been as successful as had been hoped. Signals such as low productivity, customer complaints, and observations by managers seemed to indicate that there were "significant job knowledge deficiencies." In addition, an excessive number of computer "bounces" meant that personnel were not familiar enough with the computer system to operate it efficiently.

These indications of lower than optimal performance in Sales and Services jobs prompted management to begin a comprehensive effort to determine the nature of the problem and identify corrective action. The effort that was undertaken was a widespread research activity to evaluate

operations in all airport locations and reservations offices. A team of trainers, marketing staff personnel, and managers visited each location, observed the operation, and interviewed both employees and managers.

This was an ongoing information collection effort and took several years to complete. After each visit to a field location, the research team compared notes, prepared an evaluation and fed it back to the field location, in addition to presenting it to corporate management. This overall research project uncovered a low level of proficiency and job knowledge in many of the front-line technical and sales personnel. In addition, they discovered that there was no measuring tool available to evaluate performance in the Sales and Services area, and which could pinpoint the extent of the problem. They suspected that training content, training delivery systems, line management, and individual problems combined to contribute to the situation. Also, it was apparent that in some locations, training in some areas of Sales and Services was not being done, or was being done very minimally.

The initial answer to this problem was a comprehensive system of "proficiency tests" which eventually would become the TNA system. The goal of this system was basically fourfold. First, a comprehensive system of tests would provide an accurate overall picture of the level of proficiency among employees at any given time. It would provide data from which other decisions could proceed. Second, it could give insight into the effectiveness of training in all important subject areas and could point to specific problems in training content and delivery. Third, it could improve the quality of training by giving accountability to the local training operations through measurement system. Finally, this system could provide an accurate assessment of training needs and

indicate the direction that the training organization should take in the future, by identifying areas of low job knowledge and/or proficiency.

The "proficiency tests" that were first conceived were given a preliminary introduction to the field. Several things became clear as a result of that initial exposure. The tests should not be labeled "tests"; this label sabotaged the effort and created resentment and resistance. Furthermore a concerted marketing effort would have to accompany the introduction of any new system to explain its purpose and procedures. Also, it was discovered that great care would have to be taken in developing test questions in order to have valid data, and to reduce resistance on the part of the employees taking the tests.

As a result, the TNA system was developed. Modules were targeted to specific subject areas essential to a particular job. The questions were originally developed by the corporate training staff, in conjunction with line personnel. A large pool of questions was developed, which was then tested on a pilot group. On the basis of the pilot group, many questions were rejected or revised. Finally, when the training staff arrived at an acceptable form for the tests, they were scrutinized by a panel which consisted of training, marketing, computer specialist, and actual line personnel. This panel then had the task of final revision and approval of the TNA modules.

The development of the test modules was accompanied by an extensive marketing effort to the Sales and Services Division, and to corporate management. The name of the program was changed to "Training Needs Analysis." It was part of a move to "Performance Based Training," which would tie training to specific performance objectives which could be measured. Memos describing the program and outlining its rationale were

sent to all managers in the Sales and Services Division. In addition, a brochure and a videotape presentation to "sell" TNA is being developed for distribution to the field.

One important element of the marketing effort for TNA was the repeated emphasis to the field that TNA tests would not be the basis for any individual job action, such as promotion, demotion, transfer or termination. Both the training staff and line managers considered this extremely important in preserving morale on the job and assuring cooperation with TNA.

DATA COLLECTION AND ANALYSIS

The specific data that was collected and analyzed in this study represented a preliminary application of TNA in the RES agent area. The training personnel who were developing the tests were still deciding what role TNA would play in the entire training process and specifically how TNA would be administered.

TNA for the RES area was to accomplish three goals, in this application of the program. They were:

- 1) to validate and/or refine the test questions
- 2) to assess the current state of training needs, and thus provide direction for future training
- 3) to evaluate roughly the effectiveness of past training.

Inherent in the third goal is the corporate training department's desire to provide motivation to local training operations to upgrade their operations. The introduction of accountability and visibility of results to the local training operations would hopefully result in more energy and commitment to training at the local level.

Administration and Sample Selection

The administration of the RES TNA tests was done in the period from December, 1979, through February, 1980. All fifteen modules in the RES area were administered to a sample of RES agents from each of the eight domestic reservations offices. Of approximately 2,000 total RES agents, 118 were involved in this administration of TNA. A proportionately equal number of agents was chosen from each of the eight offices. Six offices were larger, from each of which fifteen or sixteen agents were chosen, and two offices were smaller, from each of which eleven or twelve agents were chosen.

There are three job classifications for a RES agent: general public sales agent, corporate accounts agent, and lead reservations agent. Of the agents chosen from each office, these three classifications were represented in proportion to their percentage of total agents in the office. Thus, for each city, one or two lead agents were included and the remainder was divided approximately equally between general public sales agents and corporate accounts agents. Once it was decided how many of each job classification were to be selected from each city, names were simply picked off an alphabetical listing of all members of the job classification at that city. These names were picked at equal intervals to arrive at the desired numbers.

The fifteen modules were divided into two groups, one with seven modules and one with eight modules, following no particular pattern. Each agent was administered only one of the two groups of modules, but an approximately equal number of agents received each group of modules. Consequently each module had between 49 and 61 agents answering it, which represents about 2.5% of the entire RES population. In the sample,

cities and job classifications were represented in approximate proportion to their number in the population.

The tests were administered locally. Time limits for each test had been established and communicated, as had the allowable reference sources for each test. Local training staff were allowed to read over the tests and eliminate a small number of questions if they were not relevant to that office's operation. As far as was practical, test conditions were to be equivalent for all agents taking the tests. The corporate training staff who supervised the administration report that the tests were administered for the most part under strict adherence to the established guidelines.

Coding

The agents answered the tests on "self-scoring" answer cards, on which there were four possible options. The agent would erase a gray covering off what he considered the correct answer. Under the gray covering would be a mark that would tell the agent whether or not he or she had chosen the correct answer. If not, the agent kept erasing until he or she discovered the correct answer. With these cards, the agent would receive instant feedback on the answer to the question. However on questions where an agent erased more than two spaces, information on what his first choice was, was lost.

In addition this answer format caused some complications in coding the answers for analysis. If there was only one space erased, it was clear what the agent's first choice was. If there were two spaces erased, it was also clear what the agent's first choice was--the incorrect option of the two. In these cases, the answer was simply coded 1, 2,

3, or 4 depending on which option the agent chose his or her first time. If the agent had erased three spaces, the answer was coded "5," meaning "two incorrect choices," with the assumption that it is unclear what the first choice was. If the agent erased all four spaces, it was coded "6," meaning "three incorrect choices," following the same logic. If the agent made no attempt to answer the question, it was coded "7." And finally if the question was thrown out by local administration, it was coded "8."

All eight coding options are mutually exclusive and comprehensive. Furthermore, an accurate count can be obtained on how many agents answered correctly the first time, which is the key information. The only information that is lost on some questions is which option was an agent's first choice on a question which he or she answered incorrectly. On the other hand, this coding scheme does give information on which questions the agents found particularly tough, i.e., the questions that have high concentrations of 5's and 6's.

Data Analysis

The data that was gathered was analyzed in three ways, as described below.

i. Test Scoring and Item Analysis

The test scoring and item analysis was performed using the TESTAT program developed by Donald J. Veldeman¹ and implemented at the University Counseling Center at the University of Kansas. The analysis was done on the Honeywell system 6660 computer at the Academic Computer Center, University of Kansas.

All answers to each question were compiled and distributed by

percentages of each of the eight coding options. The correct option for each question was identified and keyed in. Thus the number of correct answers for each question (difficulty index) was obtained, in addition to the distribution of all eight options. This data indicated the relative difficulty of each item, along with possible indications to explain that difficulty. For example, a low difficulty (low number of correct answers) along with a high concentration on one other option might indicate a widespread misconception in training or a poorly worded question.

In addition, the item's "discrimination index" was calculated using the point biserial correlation coefficient method. The formula for this calculation is:

$$r_{pbis} = \frac{M_r - M_w}{S_t} \cdot p(1-p)$$

where

M_r = mean score for persons choosing the right answer

M_w = mean score for persons choosing the wrong answer

S_t = standard deviation for all scores

p = proportion choosing the right answer.²

Basically this index computes the correlation between a set of scores on the one item (a dichotomous variable) and a set of scores on the entire test (a continuous variable).

On the assumption that the score on the entire test is a valid indicator of competence, the discrimination index estimates the degree to which competent performers performed well on any given question. A low discrimination (below .20) would normally indicate that the question does not discriminate adequately between good and bad performers. A significant negative discrimination index would indicate that the question discriminates against good performers. This index can help

to provide evidence for the internal validity of a certain test question.

However, it must be kept in mind that by nature, the TNA tests are different from tests in an educational setting. The goal of the TNA tests is not to obtain a normal distribution with a mean of 60 and to discriminate within the full range of performers. The goal of TNA is to provide a set of questions of minimal requirements, in which virtually everyone will score over 80%. The desired result is not a normal distribution with an ideal mean of 60, but rather a distribution where the lower end is above 80 (a typical MAP level). Obviously this will ultimately result in a distribution that is normal neither at the center nor at the ends of the distribution. Consequently, as the point biserial correlation coefficient is least powerful at the extreme ends of the continuous variable,³ some power is lost as the difficulty index for a question approaches 1.0 or zero. This must be kept in mind in interpreting the discrimination index scores.

ii. Comparison to Minimum Acceptable Performance (MAP)

The second area of analysis done with the TNA data is a comparison of the test score distributions for each module with the MAP score for that module. This analysis is not a statistical manipulation, but rather a simple comparison between the desired and the actual situation. For each module the percentage of agents who scored above the MAP level was calculated and compared to the desired level of 100%.

This analysis gives key information as to the effectiveness of past training against corporate goals, systemwide. It gives a graphic picture of where the organization currently stands with regard to training goals, and indicates what needs to be done to change the situation, if change is desired.

iii. Analysis of Variance Between Locations

For each module, one-way analysis of variance was performed on the module scores, using the eight office locations as groups. The BMDP 7D program* was used for this analysis, using the Honeywell system 6660 computer at the Academic Computer Center, University of Kansas.

The object of this analysis is to determine if there is any significant difference in performance on any of the modules between the different locations. Although this does not represent an experimental manipulation, the data that is obtained nonetheless would be helpful in pointing to discrepancies that exist between locations, that could not have occurred by chance (at the .05 level of significance). Consistent differences that occur between locations will provide an indication of possible ineffectiveness of past training at certain locations. As all locations receive the same materials and advising from the corporate training staff, discrepancies in TNA performance between locations can probably be traced to the training delivery system or local management.

However, there are virtually no experimental controls on this research. Consequently intervening variables and contaminants will be present, and this fact should be taken into account in interpreting the resultant data.

SUMMARY

There were actually two different methodologies used in this study,

*The Biomedical Computer Programs (BMDP) were developed at the Health Sciences Computing Facility, UCLA, which was sponsored by NIH Special Research Resources Grant RR-3. The 7D program was revised in November, 1979.

and the second is embedded in the first. The primary goal of this study was to analyze and critique a training evaluation program which was implemented in a large airline corporation. The methodology for this goal was to describe in detail the program under analysis and to describe the organizational history and environment in which it occurred; this was done in the first part of this chapter. However, part of the description of that training evaluation program was a description of the methodology the company used in collecting and analyzing their evaluation data. This evaluation methodology was set up by the company, and part of it (item analysis and analysis of variance) was actually carried out as part of this study. This "second" methodology is described in the latter part of this chapter.

NOTES: CHAPTER FIVE

¹Donald J. Veldeman, Fortran Programming for the Behavioral Sciences (New York: Holt Rinehart, 1967), pp. 170-181.

²Robert L. Thorndike, Educational Measurement (Washington, D.C.: American Council on Education, 1971), p. 142.

³Thorndike, p. 142.

CHAPTER SIX

RESULTS OF TNA

TNA (Training Needs Analysis) was administered for the second time during the period from December 1979 to February 1980. The results of that administration are reported below. The results fall into three categories: item analysis, comparison to MAP, and analysis of variance between locations.

ITEM ANALYSIS

Difficulty Index

The difficulty index indicates the percentage of respondents who answered any one question correctly. In this sense, "difficulty index" is a misnomer because it increases as the real difficulty of a question decreases. For example, a question which 95% of respondents answered correctly would have a higher difficulty index than a question which only 75% of respondents answered correctly. Thus, a high "difficulty index" would actually indicate a high number of correct answers, and would indicate a low actual difficulty. Why the majority of textbooks persist

in using this confusing phrase is unclear. However, since it seems to be in standard use, it is used in this study, despite its unfortunate awkwardness. The value of this statistic is simply to indicate the relative difficulty of each test item.

Discrimination Index

The discrimination index obtained a correlation between subjects' scores on one question and their score on the entire test. In other words, if most of the people who scored high on the entire test, got one item correct, and most of the people who scored poorly on the entire test got the item incorrect, then the item would have a high discrimination index. In this case the test item is able to "discriminate" between those individuals who score high on the overall test and those who score low on the overall test. If half of the high scorers and half of the low scorers got an item correct, then the item would have low or zero discrimination; the item did not differentiate between low and high scorers. Finally, if the high scorers got an item incorrect, while the low scorers got an item correct, then the item would have negative discrimination. In other words, the item discriminates against high test performers.

The discrimination index, as calculated in this study, is least powerful at both ends of the difficulty scale. That is, as difficulty approaches 1.0 or zero, the discrimination figure is less interpretable. This is logical at its face; if 90% of respondents answer a question correctly, it cannot discriminate effectively because the vast majority of respondents answer the same way. Put another way, in the case of a high difficulty (>90%), the statistic becomes too sensitive to the

to the small number ($<10\%$) of incorrect respondents.

On the assumption that the entire test is sound, this statistic is one measure of internal validity; a high score on the test should correspond with a correct answer on the particular item. The goal of this analysis was to verify the validity of questions on the TNA tests and to identify questions with weak or negative discrimination.

A somewhat arbitrary choice was made with respect to these statistics. A discrimination index was considered acceptable if it were over .20. However, those questions with a discrimination index of less than .20 and a difficulty index of .90 or higher were also considered acceptable for the reasons explained above. Those questions with extremely low difficulties ($<.50$) were pointed out as possible areas of training needs.

The results in Table I indicate an overall good discrimination ability across the fifteen tests. Although most modules had at least one question which had a low discrimination index, many of these numbers are uninterpretable because 90% or more of the respondents scored correctly on that particular question. Of the low discrimination questions, only 25 (out of 304 total questions) also had acceptable difficulty figures (that is, less than 90%). Hence, only 25 of 304 total questions, or 8.2%, indicated a problem with discrimination power. This would give a fairly good evidence of the overall internal validity of the tests. Furthermore, it is evidence that the validity strived for in the creation of the tests was, to a large degree, accomplished in the administration of the tests.

The 25 questions which exhibited poor discrimination were examined by the corporate training staff for ambiguity, error, unclearness, or

TABLE I: RESULTS OF ITEM ANALYSIS

| Module | no. of questions on test | no. of questions w/ discrim. < .20 | no. of questions w/ discrim. < .20 and diffic. < .90 | list of questions w/ discrim. < .20 and diffic. < .90 | no. of questions w/ diffic. < .50 | list of questions w/ diffic. < .50 |
|--------|-----------------------------|---------------------------------------|--|---|--------------------------------------|---------------------------------------|
| 501 | 38 | 13 | 4 | #6,22,27,28 | 4 | #10,15,18,30 |
| 502 | 17 | 0 | 0 | - | 1 | #17 |
| 503 | 19 | 3 | 1 | #2 | 2 | #7,16 |
| 506 | 34 | 9 | 6 | #1,2,18,22,24,34 | 3 | #18,22,31 |
| 507 | 18 | 2 | 2 | #6,9 | 3 | #6,12,15 |
| 508 | 13 | 2 | 1 | #10 | 2 | #4,11 |
| 509 | 21 | 5 | 3 | #2,4,6 | 5 | #6,11,14,16,17,21 |
| 516 | 22 | 4 | 1 | #18 | 1 | #15 |
| 521 | 18 | 1 | 0 | - | 1 | #11 |
| 522 | 12 | 3 | 1 | #6 | 4 | #4,5,7,12 |
| 526 | 15 | 3 | 0 | - | 3 | #6,10,11 |
| 529 | 15 | 2 | 0 | - | 2 | #6,10 |
| 541 | 13 | 0 | 0 | - | 2 | #9,11 |
| 551 | 26 | 1 | 1 | #15 | 3 | #2,5,11 |
| 700 | 23 | 7 | 5 | #2,4,11,15,17 | 3 | #7,15,21 |

irrelevancy, to account for the low discrimination. That is, the statistical analysis was to help point out deficiencies in the questions or answers which may not have stood out initially. For some of the questions, the questions and answers appeared valid and legitimate and the item was allowed to stand. In these cases, especially when the question also showed a low difficulty index, the item was flagged as an indicator of a possible training need.

However, in many cases the question or answer was revised after an examination of the statistical data. In addition to looking at questions with low discrimination, the staff also examined questions which had a low difficulty index. These questions, on which fewer people scored correctly, were examined for clarity and relevance. Between questions with low discrimination and those with low difficulty, 19 were revised or omitted on the basis of the statistical analysis.

In addition, the corporate training staff's scrutiny of the test results by discrimination and difficulty helped to uncover training needs. The computer printout which included the choice distribution for all the answers to each question, tended to emphasize areas which could be addressed by training. The staff used the choice distribution to explain why some questions posed problems to employees. Comments such as "bad habits," "agent carelessness," "new flight availability rules," and "too similar computer entries" described areas of training needs indicated by particular questions.

COMPARISON TO MAP

Each of the fifteen TNA modules has a "minimum acceptable

performance," or MAP score (except two for which the MAP has not yet been established). This is considered the level of knowledge that each employee should have in each area described by a TNA module. Thus the corporate goal or standard is to have 100% of employees scoring at MAP or above. Part of the goal of the present administration of TNA was to measure the performance of RES employees against the MAP standard.

The results of the module scores, as against the MAP scores, are shown in Table II. The modules are placed in increasing order of percentage of respondents who scored above MAP. On none of the modules did performance meet the corporate standard. On two, modules 502 and 516, over 50% of the respondents scored above MAP. Also, these two modules were the only two for which the mean score of all respondents on this administration of TNA, was at or above MAP. These two areas, "Determining Domestic Flight Availability Manually," and "Two-City Itinerary Pricing," are the current strong areas with respect to reservations training.

On the other hand, four modules--506, 522, 526, and 541--showed the poorest performance as compared to MAP levels. On module 506, "Entering Passenger Ticket Reservations," only 11.1% of the respondents scored at MAP or above. The mean of all respondents on the module was 76.8% or 13.2 points lower than MAP. Module 526, "Hotel Services," only had 15.1% of respondents scoring at MAP. Module 541, "Flight Service Information," had 20.0% over MAP while module 522, "Seat Assignment," had 21.2% at MAP or above. On none of these three modules did the mean score approach the MAP level; the means were from 10.4 to 15.3 points lower than the corresponding MAP.

The scores for the four modules above indicate clearly that a training need exists. Furthermore, it would appear that past training

TABLE II: COMPARISON OF RESULTS TO MAP SCORES

| <u>Module</u> | <u>No. in sample</u> | <u>Minumum Acceptable Performance (MAP)</u> | <u>% above MAP</u> | <u>Mean</u> | <u>Standard deviation</u> |
|---------------|----------------------|---|--------------------|-------------|---------------------------|
| 506 | 54 | 90 | 11.1 | 76.8 | 11.2 |
| 526 | 53 | 85 | 15.1 | 70.1 | 16.8 |
| 541 | 60 | 85 | 20.0 | 69.7 | 17.6 |
| 522 | 52 | 85 | 21.2 | 74.6 | 13.3 |
| 508 | 54 | 85 | 25.9 | 74.6 | 14.3 |
| 503 | 49 | 85 | 28.6 | 76.4 | 14.1 |
| 529 | 60 | 80 | 33.3 | 69.4 | 18.8 |
| 501 | 61 | 85 | 39.3 | 79.8 | 11.9 |
| 551 | 60 | 80 | 40.0 | 70.1 | 20.1 |
| 507 | 54 | 85 | 40.1 | 77.2 | 14.0 |
| 521 | 53 | 85 | 45.3 | 78.3 | 14.7 |
| 502 | 61 | 85 | 62.3 | 85.0 | 11.2 |
| 516 | 60 | 80 | 81.7 | 85.6 | 11.0 |
| 509 | 54 | - | - | 66.3 | 13.3 |
| 700 | 57 | - | - | 79.4 | 9.2 |

had been deficient when comparing performance to corporate goals. However, the TNA scores by themselves cannot indicate definitively whether training has been ineffective, or whether some intervening factor (such as differences in personnel selection, management, or seniority) has been responsible for the low test scores. The test scores in themselves also cannot prescribe what specifically should be done to improve the effectiveness of the training. What can be determined from TNA is that the results of training in four specific areas need to be improved. Based on previous observation and analysis by the corporate training staff, the possibility exists that training is at times simply not being done in the four areas cited above.

ANALYSIS OF VARIANCE BETWEEN LOCATIONS

The scores for each module were grouped by location in order to perform analysis of variance between locations for each module. The results of this analysis are given in Table III.

As can be seen, a significant difference in score between location was discovered on only one module. On module 529, "Car Services," the variances between location exceeded those attributable to chance at the .05 level. This would seem to indicate that some type of systematic influence is causing the San Francisco office to score a mean of 45.5%, for example, while the St. Louis office scores a mean of 80.8%. Some differences in personnel, selection, management, or training apparently is resulting in differences in performance on one particular TNA module.

The interpreter of these location scores must keep in mind that the sample sizes were extremely small. Consequently differences between

TABLE III: RESULTS OF ANALYSIS OF VARIANCE

| <u>Module</u> | <u>Lowest score</u> | <u>Highest score</u> | <u>Degrees of freedom</u> | | <u>F value</u> | <u>significant</u> <u>(.05 level)</u> |
|---------------|---------------------|----------------------|---------------------------|----------------------|----------------|--|
| | | | <u>between groups</u> | <u>within groups</u> | | |
| 501 | 41.0 | 100.0 | 7 | 53 | 1.60 | no |
| 502 | 43.7 | 100.0 | 7 | 53 | 1.10 | no |
| 503 | 31.5 | 100.0 | 6 | 42 | 1.43 | no |
| 506 | 44.1 | 94.0 | 7 | 46 | 1.36 | no |
| 507 | 35.2 | 100.0 | 7 | 46 | 1.06 | no |
| 508 | 30.7 | 100.0 | 7 | 46 | 1.77 | no |
| 509 | 38.0 | 93.3 | 7 | 46 | 1.62 | no |
| 516 | 54.5 | 100.0 | 7 | 52 | .96 | no |
| 521 | 33.3 | 100.0 | 7 | 45 | .62 | no |
| 522 | 40.0 | 100.0 | 7 | 44 | 1.70 | no |
| 526 | 28.5 | 100.0 | 7 | 45 | .51 | no |
| 529 | 26.6 | 100.0 | 7 | 52 | 4.31 | yes |
| 541 | 7.6 | 100.0 | 7 | 52 | 1.29 | no |
| 551 | 20.0 | 100.0 | 7 | 52 | 1.19 | no |
| 700 | 52.1 | 95.2 | 7 | 49 | 1.67 | no |

location were very difficult to detect statistically. While this would give power to the one significant difference score found, it would also indicate that no conclusion should be drawn quickly about the absence of significant differences on the other modules. If sample sizes greater than four to nine per location per module were obtained, possible significant differences would be discovered that would not be discovered with the smaller sample sizes.

A cruder examination of the module by location scores seems to indicate a pattern of performance by location. That is, some locations tend to score higher or lower than the overall mean consistently over modules. For example, the Philadelphia and Los Angeles offices scored above the overall mean on twelve of the fifteen modules. Conversely, Chicago scored above the overall mean on only five of the fifteen modules. This kind of "eyeballing" does not give technically significant differences, but it does point to patterns which may be reinforced over different administrations. The pattern pointed out above, for example, would merit a closer examination of the training operation in Chicago to see if any obvious training deficiencies exist.

CHAPTER SEVEN

EVALUATION OF TNA

In addition to analyzing the numerical data obtained from this administration of TNA, the entire TNA program was evaluated against the criteria for training evaluation programs developed in Chapter Four. The TNA program was not a "pure" training evaluation in that evaluation was not its only purpose. However, training evaluation was central enough to the purpose of TNA to evaluate the program against the training evaluation model. Like many programs occurring in a real-life setting, the TNA program was subject to compromises and constraints which made it a combination of training functions. This factor, far from making TNA ineligible for analysis as a training evaluation program, makes it a realistic case to study.

The model presented earlier identified four stages of evaluation--rationale, formulation, implementation, and use--and five key adjectives describing a good evaluation system--planned, unified, flexible, appropriate, and complete. Each of the four stages of TNA will be described and evaluated, and then comments will be made on the extent to which TNA can be described by the five adjectives above.

RATIONALE

As was pointed out in Chapter Five, the TNA program was initially a response to an urgent problem which existed in this particular corporation. It had been discovered that performance in various Sales and Services areas was below par. Corporate management eventually came to the conclusion that training, or lack of it, was largely responsible for this situation. Consequently a plan was conceived to rectify this situation with respect to training. This plan included the development of a performance measurement system which could give a rough idea of the success of past training while pointing to directions for future training.

Because many employees did not have much of the basic knowledge required to do their jobs, the conclusion was drawn that training was the culprit leading to low productivity. Considering the importance of knowledge to most of the Sales and Services jobs, this was a logical conclusion to make. And certainly focusing on training should be at least part of the solution to the problem. However, there may be other contributing factors to a low productivity problem. Although the original data was not accessible, other contributing factors could include personnel selection, job design, morale, or management policy. Even if training were to be improved, there is no guarantee productivity would improve if these other factors existed.

Even if other action needs to be taken in addition to addressing the training problem, the rationale behind developing a training evaluation program in this situation is sound. Job knowledge deficiencies do exist and can be corrected through improved training, and thus

training should be at least part of the focus of solving the productivity problem. Other action may also be necessary if closer examination indicates that management, selection, or job design may also be contributing to the problem; this is, however, outside the scope of this study.

In general, the problem of low productivity detected before TNA was developed, produced the rationale for improving training. In particular, the response to the problem was to include four functions. These were 1) to develop a way to measure an indication of performance, 2) to evaluate the effect of past training, 3) to assess needs for future training, and 4) to motivate line management to improve and expand training. The rationale behind each of these four functions is sound, and is a logical response to the problem. First, developing a measure of performance was necessary to obtain data in order to know the precise parameters of the performance situation. Management had informal, imprecise information on the extent of job knowledge among employees; they needed a more precise barometer, though.

Second, the rationale for evaluating past training is clear. Once information on the current state of competence among employees is available the next step is to determine how effective past training has been in developing that competence. Hopefully, this training evaluation will also show in what specific areas training has been effective, and indicate why training has or has not been effective.

A corollary to the measurement and evaluation functions is an assessment of current training needs. This third function has a clear rationale of outlining the needs that future training will meet. If training is to improve, it must be responsive to actual and timely needs that exist and that can be met through training.

Finally, the function of motivation is largely a product of the organizational set-up of the corporation. The line management functions actually conduct training, but the corporate training staff is ultimately responsible for its effectiveness. Thus, line management previously was not directly accountable for the effectiveness of training and consequently had less incentive to be overly concerned with training. This was especially true since there was no measuring system which identified specifically the effects of training. The introduction of a measuring system designed solely to key in on training effects suddenly made very visible the effect of line management's training efforts. Thus line management now becomes accountable for the outcome of the training for which they are responsible.

In the early stages of development of TNA, the above are the four basic rationales which were behind the program. Each was a response to the problem as it was perceived and each was a coherent and logical function.

FORMULATION

In the formulation stage, the constraints acting upon the goals stated in the first stage are outlined. In addition, a number of decisions and assumptions are made regarding how the goals of the program will be set into motion.

The first of these decisions is the level of evaluation to be used. Will reaction, learning, behavior, or results be the focus of the measuring instrument and thus the evaluation? In this case, the level of learning was selected as the criterion on which information would be

gathered. This selection is wise; knowledge is crucial to most Sales and Services jobs, especially those of RES specialists. Knowledge is also fairly easy to measure accurately, which is the biggest disadvantage of the level of behavior.

Intuitively, the connection between knowledge and employee productivity seems clear. One corporate training manager makes this assumption explicit as he states in a memo about TNA, "I believe there exists a direct correlation between job knowledge and performance." Evidence for this assumption may have been collected, but earlier information-gathering efforts on the productivity problem were not available. However, an external validity test with the TNA modules is still recommended.

A selected group of employees could be rated on some external measure of productivity and then these scores could be compared to scores on the various TNA modules. Obviously the biggest problem in this procedure is the selection of an appropriate external measure. Such a selection must be made by corporate management as they decide what is ultimately important for a RES agent. Once the selection of an external measure has been made, a comparison of it and performance on TNA would provide invaluable information.

The second major decision made in the formulation stage refers to the "packaging" or communication of the new program. The corporate staff wisely realized that the acceptance of the program by line management and employees was crucial to its success. Thus the communication of the new program had to be done in a manner that was perceived as non-threatening and positive by the field. This constraint would prove to have an effect on the substance of the program as well as the appearance of it, in that the fairness and validity of test questions would influence

greatly whether the program was accepted by employees.

Third, the decision about the focus of the program had to be made in the formulation stage. In Chapter Three the importance of limiting and directing the focus of any training evaluation was emphasized. It was also pointed out that in many situations the existing constraints make compromises and less than "pure" solutions necessary in the name of efficacy. In the case of TNA, the focus was not sufficiently limited; the program was trying to do four things at once. However, this is both a shortcoming and a benefit. Of the four functions of TNA described several pages earlier, some are accomplished more effectively than others. The training evaluation function of TNA suffers from being combined with other functions, but measurement and motivation functions work well in TNA. The assessment function is more effective than the evaluation function, but also suffers from not having clear, accurate information as to the effectiveness of past training. While combining a number of functions may water down the effectiveness of some of those functions, it also is enormously practical. Doing "four jobs for the price of one" may be worth making some compromises in some of the functions.

Finally, the decision on the necessary degree of experimental rigor was made at this stage. A good decision was made: the exclusion of any contaminating variables was not necessary with regard to the evaluation function, and somewhat speculative data would be acceptable. Clearly, the role of experimentally valid evaluation information was not as important as, for example, clear information on the current state of training needs. Whatever information was needed on the effectiveness of past training could be extrapolated from the assessment conclusions.

This decision was probably made as a result of several existing constraints. The improvement of training was of paramount importance at this point, regardless of the extent to which past training had been effective. Thus assessment of the current situation held the most urgency. Also, the corporate training office was operating with a limited staff. Even if a rigorous evaluation was desired in addition to the other functions, the corporate staff may not have had the manpower to accomplish it. Thus a multipurpose program could look quite attractive. Finally, an experimentally precise evaluation may not have been necessary because the staff already believed they knew how effective past training had been; their conclusions probably were actually quite reliable.

IMPLEMENTATION

The decisions made in the formulation stage are operationalized in the implementation stage. With regard to implementing the communication strategy, the staff appeared to engender the cooperation and acceptance by the field at the same time an atmosphere of serious accountability was developed. Several efforts were made to reduce the threat of the TNA program. First, although the program was made up of tests, no mention was made of tests. The name "Training Needs Analysis" emphasized a positive perspective without hinting at punishment. The answer sheets provided a self-correcting, instant feedback system which discounted the feeling of being "graded." The introduction of TNA was accompanied by a public relations effort which fully explained the role of TNA and its function. As outlined previously, the staff went through a great deal of

effort to assure that the tests were fair, valid, relevant, and positively related to the job.

At the same time, the staff made it clear that the TNA program was to be taken seriously. Field personnel were aware that corporate management was behind this effort to address the productivity problem in Sales and Services. The training staff also made it clear that performance on TNA across the system would be published. Thus while no outright punishment was indicated, each field office knew that their performance on TNA, as well as that of the other offices, would be available to all offices. This provided an effective incentive to take the program seriously.

The decision of learning as the criterion measure was translated in this stage into the actual job knowledge tests. The major concerns in implementation at this point were with validity and reliability. Validity was emphasized through the careful development, testing, and refinement of the tests. The item analysis performed as part of this study also indicated a high degree of internal validity, with further refinement possible in the future. However, as was pointed out in the last section, evidence of external validity (comparison with an external measure of performance) is also called for. Concern with reliability for these tests was clear. Rules for consistent administration of the TNA tests were established, communicated, and followed. As long as enforcement of these rules is accomplished, the reliability of the tests should be assured.

The implementation of the research design for the TNA program was the weakest part of the program. This administration of TNA basically followed the "one-shot case study" design, which it was noted, has

virtually no experimental controls. The data from this administration can give an accurate measure of the current state of job knowledge among RES employees. The information indicates the training needs that currently exist and thus suggests areas which most need training.

Any conclusions on the effectiveness of past training from this information are entirely speculative. Any number of extraneous variables, in addition to training, could be responsible for performance on the TNA tests. Performance against the corporate standard (MAP) can indicate, to a degree, the effectiveness of past training, but this information is not helpful in improving training in the future. There are too many extraneous variables to draw any conclusions about the causes of past training effectiveness.

It should be remembered that the decision against experimental rigor was made at an earlier stage, and was made in light of all the circumstances. The value of experimentally precise evaluation information was judged too low, as compared to the costs of obtaining that information. Speculative evaluation information which was incidental to the other functions of TNA was deemed adequate.

However, a simple pre-post test, control group design should be used in future administrations of TNA, if practical constraints allow. If a number of new reservations agents are hired or transferred to any location, they could be administered TNA before they start their training, and then again after the eight-week training is over. A control group of non-RES employees could be chosen for each module of TNA, to be given the module at the time the new employees are pre- and post-tested. Thus information could be obtained on the experimental effect of training on TNA performance. The weakness of this design, a possible sensitization

to training by pre-testing, would actually be an advantage from the perspective of trainers, in that it would facilitate learning on the important points in training.

USE OF DATA

The use stage of the TNA program was successfully implemented.

In this administration of TNA, the location results were compiled, published, and distributed to the field within about four to six weeks after the tests were administered. This feedback was accomplished with a memo which described performance on TNA in detail, including information on scores by module and location, performance across the entire system, and comparison of scores to MAP. In addition, this information was made available to all members of the corporate training staff and was forwarded to the regional Vice Presidents in charge of each RES office.

Whether or not this TNA performance information was made available to most RES employees is unclear. This decision was left up to the line managers at each location. Because of the decentralized nature of the training organization, the central training staff simply sent the information to the location managers, who could each decide how widely the performance information would be disseminated. Although the most desirable situation would probably be for each employee to know the status of their office's performance, that is not within the corporate training staff's area of responsibility. A positive feature of TNA with respect to individual feedback is the nature of the test score cards. Each person who took a test knew at the time of taking the test what their score was, because of the "self-scoring" nature of the answer

cards.

Was the TNA performance information actually used in decision-making? There were four goals of this program and the use of data against each of these goals must be analyzed. The first goal was to develop a way to measure performance. The information was used both in the development and measurement aspects of this goal. The item analysis (discrimination and difficulty) was used to identify, and reject or revise nineteen possibly invalid questions. Also, performance on TNA was accepted by the training staff as being a satisfactory indicator of job performance, and was communicated as such. Thus, the information derived from TNA was used in meeting the goal of developing a measuring instrument for Sales and Services performance.

The second goal, evaluating past training, used the results of TNA only in a broad and imprecise fashion. In the report to the field, the comparison of MAP to the TNA scores indicated clearly that most locations were below MAP. The implication was that past training had been ineffective to the extent that employee scores fell below MAP. However, as pointed out earlier, drawing conclusions about past training based on a one-shot test administration, is tenuous.

The third goal was to use the TNA results to assess needs for future training. The resultant data was used largely for this purpose. Three modules were identified as areas which were the most urgent with respect to current training needs. These three modules are being emphasized to the field as the primary areas for which remedial training is to be given. Although the corporate staff does not have ultimate authority over training at field locations, they made the strong suggestion to the field that "a very important aspect of the TNA approach is follow-up in the

form of re-training to strengthen identified skills weaknesses by subject matter."

In addition to the three modules identified as system-wide needs, field locations were to look at TNA areas in which they performed less well than the system, and tailor their training to address those areas. Furthermore, through looking at the commonly chosen answers to selected questions, the corporate training staff identified specific areas in which the training materials they sent to the field could be improved. Thus the information resulting from TNA was actually used in making decisions about the directions for future training.

Finally, the TNA information was also used in reaching the fourth goal, motivating the field to have accountability for their training. It was pointed out in an earlier chapter that the people making up the RES function are very competitive. Consequently a measuring, or accountability program should stimulate performance among that group. This result was promptly achieved.

The TNA results of each location on all modules were included in the reports sent to all locations, but the locations were identified only by number. In an interview with a member of the corporate training staff, I was told that on the same day the report was distributed, nearly all of the location managers were on the phone trying to find out which numbers corresponded to which locations. This staff member indicated that no RES location manager wanted to be at or near the bottom of the performance scale, especially with their regional Vice President getting the information. Consequently this informal indicator would appear to show that this information is actually motivating the field training function. The ultimate result however, will be seen further down the

line if TNA scores improve.

EVALUATION AGAINST KEY CHARACTERISTICS

The TNA program performs well against most of the five key descriptive adjectives of a good training evaluation program--planned, unified, flexible, appropriate, and complete. The program was soundly planned and conceived, in that its ultimate objectives were valid and were made clear. The goals which were identified were adhered to throughout the program. The only possible weakness against this point is that the planning came up with too many goals for one program. Consequently the program did not have a well-defined focus.

The TNA program was unified, as the decisions made earlier were carried out throughout the program. Whether or not the early decisions were correct ones, they did not get lost along the way. Thus, for example, the decision about experimental rigor was consistent with both the goals of the program and the constraints inherent in carrying out those goals. The TNA program was also a unified part of the larger system, the entire training system. TNA was designed to be a learning process in itself, in addition to being a measuring tool. Furthermore the output of the TNA program is easily and readily applicable as input into the training process.

TNA also meets the criterion of flexibility; this is perhaps one of the stronger points of this program. The tests are designed so that modules can be changed, added, or removed without disturbing the integrity of the program as a whole. Also each module corresponds to an identifiable area of job knowledge. Thus as information and conditions change within

one job area, questions can be revised, added or deleted. Finally, the administration of the tests is flexible. At this point, the measuring instrument itself is sound, although the research design using it may have been flawed. This administration was largely a trial one and thus, in future administrations the design can be changed while using the same TNA modules. The TNA concept should lend itself well to the changes that will no doubt occur in the training function in the future.

The TNA program also met fairly well the characteristic of appropriateness. Choosing the criterion of knowledge as the focus of measurement was appropriate in the case of RES agents. The entire test concept was a good one under the circumstances, as it was a quick, accurate and efficient way of measuring a close indication of job proficiency. The crucial trade-off came between combining four functions in TNA and doing a precise, complete job of training evaluation. Obviously the choice was made to combine a number of functions, save time, money, and resources, and sacrifice precise evaluation information. This decision was probably appropriate to the circumstances in light of the corporate training staff's limited manpower, and the relative lack of importance of detailed evaluation data.

The TNA program was complete. On this important criterion, the TNA program carried out its mission through to completion, including the dissemination and use of the information that was gathered. The completion of the process and the use of TNA information as input into the next stage of the training process will undoubtedly inject the program with credibility, and enhance its effectiveness. Despite the shaky nature of the research design and the somewhat preliminary nature of this administration of the tests, the TNA program has to be considered an overall success, especially with an eye toward potentially improved

future administrations of the program.

CHAPTER EIGHT

DISCUSSION AND RECOMMENDATIONS

This study set out a model for training and then examined one training evaluation program occurring in an actual corporate setting. The mechanics and the outcome of the TNA program have been examined, and the program has been analyzed in the context of an ideal training evaluation model. As a result of this investigation, several conclusions and some recommendations for future research will be offered. The conclusions range from those dealing with the specific methodology used in this study to those dealing with investigation in organizations in general. The recommendations for future research concern further examination into the TNA situation and further examination into the area of training evaluation.

CONCLUSIONS

(1) The model developed in Chapter Four was successful in meeting its goals, while there are still shortcomings in the "model idea" itself.

In Chapter Four, after an examination of the existing literature on training evaluation, the model that was presented synthesized the important

aspects of training evaluation found in the literature. The purpose of that synthesis was to provide a framework against which operative training evaluation programs could be judged. The goals of the model were to make it broad enough to be used with a wide spectrum of evaluation systems, specific enough to critique evaluation systems on tangible points, and realistic enough to take into account the many pragmatic concerns that seem to be prevalent in corporate programs.

The model provided an effective outline from which the TNA system could be critiqued. It provided a ready-made progression of thought for the analysis, and the organization of the model helped to assure that no major point would be overlooked. The fact that TNA turned out not to be primarily a training evaluation system, did not ruin the applicability of the model. On the contrary, it helped to clarify that TNA did not meet many of the training evaluation criteria, particularly in the rationale and formulation stages.

In spite of the success of the model, this concept had two shortcomings. The first is that there was no "formula" or prescription derived for effective training evaluation. Although components and characteristics were identified, the formula for the right combination of them could not be specified, unless the formula only applies to a very narrow type of training evaluation situation. If a training evaluation model is to apply to the broad area of training evaluation, there are too many contingencies involved to be very specific in prescribing "the" right evaluation program. Possibly, future research could identify the key contingencies along with the proper response in many different situations. This model attempted to identify the key contingencies, but without prescribing a specific response.

The second shortcoming to this model is that, in spite of the attempt to make it apply broadly, the model may have too limited a focus when applied to corporate situations. In other words, programs in organizations may be so situation-specific that they cannot be narrowly defined as "training evaluation" or "needs assessment." This study illustrated one example of a corporated program that was neither exclusively training evaluation, needs assessment, nor performance motivation. The fact that probably many corporate programs in practice have multiple purposes may confound an attempt to pattern an ideal training evaluation. While the occurrence of multi-purpose programs may make the creation of a training evaluation model difficult, it does point out the systemic nature of units and activities in an organization. The various functions in the training process are all interconnected, and thus it is difficult to isolate one part for examination.

(2) The discrimination index should not necessarily be used as a final or "acid" test in determining whether an item is acceptable.

The discrimination index was the statistic used to determine how well employees' responses to a test question related to their performance on the test as a whole. This statistic was used to verify the validity of test questions and to provide information which would help improve the test questions.

The discrimination index should be a first indicator of whether there may be a problem with an item, but ultimately an examination of the face validity of the test question should determine whether an item with low discrimination should be rejected. In this study, several of the items with low, or even negative, discrimination were examined by the training staff and considered acceptable. This was an appropriate reaction in

view of the care with which the test items were written, and the trainers' familiarity with the subject matter. In addition, the newness of this entire testing process would make the interpretation of the overall test scores somewhat tentative.

In spite of the statistical and objective nature of the discrimination index, this statistic still must be interpreted somewhat subjectively. As mentioned above, it is not a final indicator of validity. Furthermore, in interpreting the discrimination scores, one must have a close familiarity with the subject matter and the training process in order to judge why an item does or does not discriminate. Since I did not have that familiarity in this case, the discrimination scores could give me relatively limited data. Fortunately, the trainers in this study made the finer interpretations on why items discriminated the way they did, and what should be done in response to those scores.

(3) The case study approach appropriately provided useful information, and pointed the way to further research.

There are a variety of approaches to training evaluation. Training varies, organizational environments vary, and thus appropriate training evaluation systems vary. For this reason, it is appropriate to examine in detail a number of unique situations in order to gain insight into the principles that apply across the board to all training evaluation situations. A case study, like the one developed here, is useful for examining one situation closely in order to identify the important elements of effectiveness. In addition, the case study approach tests knowledge or theory which has been developed. In this case the model, which attempted to integrate a number of views on training evaluation, was tested by the actual TNA system. The concepts in the model were

examined in detail in the framework of a case study.

The current state of the training evaluation literature also makes a case study approach feasible. Not a great deal of theory has been developed with regard to industrial or corporate training evaluation. The one piece by Kirkpatrick on the levels of evaluation is the most significant theoretical development which encompasses the entire training evaluation area. There are numerous good articles which stress either one concept or one type of training evaluation, but few successfully integrate the entire area. If the area were narrowed to "Evaluation of Human Relations Training for Managers," for example, experimental studies would be much more appropriate, because the results could be generalized within the area, for which a cohesive theoretical framework has been developed. But for the broader area of training evaluation, the case study approach encourages theory development and testing.

At the same time, the case study approach necessitates a trade-off. Because each situation is unique, there may be limited generalizability of the results. There is no hard statistical proof that a certain technique works in a certain type of situation. The lack of experimental controls limits any assurance that the same study and results could be duplicated. In the case of this study, that trade-off, though, was a worthwhile one. The experimental design was poorly constructed, and so any "experimental" study would not have had valid conclusions. Both the TNA program and the theories in the model were "in process" trials rather than final solutions. Thus, this case was an opportunity to test a new theory, identify key elements in training evaluation, and recognize specific shortcomings in one evaluation system.

(4) Informal or "non-official" factors play an important role in official organization activities, but are difficult to discover and document.

In order to critique the TNA program against the training evaluation model, a clear and complete picture of the TNA program had to be obtained. Information was collected through the organization's official documents such as memos, reports, and project outlines, and through interviews with training staff, actual test results, and my own observations.

The official literature describing TNA gave a detailed rundown on the program, including the rationale and stated purpose. However, it soon became clear through discussions with the staff that there were some other underlying motivations that were less apparent, and it was somewhat difficult to gain information on these informal aspects of the program. Being an outsider coming in to an organization makes it particularly difficult to collect this type of information.

As an outsider, one does not have familiarity with the subtle and often political, informal relationships which exist in any organization and which have a great deal to do with why and how a program will be implemented. Knowing these relationships, such as the units which have more power than other units, the historical relationship among departments, and the personality of the leadership in different units, can give key insight into the "real" reason programs or systems are implemented. Second, as an outsider collecting information, one has to establish trust before organizational members are willing to freely supply information. This involves both the integrity of the investigator, and the assurance that the organization stands to gain something through the process.

In this case, there were no great revelations that were discovered. However, there was important information not covered in documents, which had to be gathered to get a full and accurate picture of this

particular organizational program. This situation provided a good example of the contrast between official formal aspects of a program, and the unstated informal aspects.

RECOMMENDATIONS FOR FUTURE RESEARCH

Several suggestions for future research follow. The first group focuses on the specific TNA program; the second group deals with research in training evaluation in general.

Future Research in TNA

Further useful research into TNA can be done in a number of ways. One important area is external validation of the TNA tests; the TNA results should be correlated with another measure of employee performance. Although the face validity and the internal validity of the tests have already been established, a test of the external validity of TNA modules would greatly strengthen their credibility.

The choice of external measure would have to be made by the corporate training staff. Either an existing performance index (such as total dollar sales by employee, or reservations efficiency) or a supervisor rating system could provide a benchmark with which TNA scores could be compared. If the correlation between the external measure and TNA scores were highly positive, the external validity would be established. If the correlation was low or negative, the role of the tests, or the role of job knowledge in performance would have to be re-examined.

Another recommendation in the use of TNA for future research is to develop a "composite score," which would sum up an employee's score

on all the TNA modules that he or she might take at one time. This score would combine the employee's scores on up to fifteen RES modules. As it stands now, the different modules cannot be added and averaged because they have different MAP scores. Presumably, a module with a high MAP level is either easier or more important than a module with a lower MAP score. Combining and averaging raw scores across modules could distort the desired information (score relative to MAP).

One solution would be to subtract the MAP score for each module from each individual's raw score on that module. Thus if an employee scored 87 on a module with a MAP of 85, his "converted" score would be +2, or 2 points higher than MAP; if his raw score was 65, his converted score would be -20, or 20 points below MAP. The employee's converted scores for all the modules that he or she took could be added for a composite score; this score gives a picture of how the employee performed generally relative to corporate standards.

Using this composite measure, the general performance between locations can be compared. Without the composite measure, a general comparison between locations was influenced if a greater number of employees took the "easier" modules at one location. In addition, this composite measure gives a constant indication of an individual's or a location's performance against the corporate standard. As long as the composite score is negative, it is clear that a training need exists.

Another recommendation is to administer the TNA modules periodically to a sample at each location and examine the trend of module scores and composite scores over time. If the implementation of TNA is increasing job knowledge and motivating field locations to improve training, the results should show up in future administrations of TNA. Such a time-series

design would still not give reliable training evaluation data because there are too many experimental variables that are not controlled (different personnel or management, the testing effect, the passage of time). However, it would give more information on which speculation can be based. And more importantly, even if it cannot give a reading on the effectiveness of past training, it can give a reading on the need for future training.

Another desirable experimental design is the pre-post test with a control group. If a group of employees were being trained simultaneously, this design would be feasible, especially since a proven test measure already exists. This design is only recommended if tangible evidence for the effects of employee training is needed by the training staff.

Finally, as part of a further investigation into the effectiveness of TNA as a program, it would be very helpful to include interviews and observations at the eight field locations. The evaluation of TNA done in this study had to be limited to data collection at the training headquarters simply because of financial and time constraints. However, if it were feasible in the future, a great deal of important information could be gathered through interviews with line managers at field locations. Such interviews could provide information on the managers' perceptions of TNA, their impressions of TNA's success in motivating them, and their attitudes toward TNA. Also, a part of these trips could be the observation of some actual training being done at the field location. Both the observations and the interviews could give added information and a perspective on TNA that cannot be obtained simply through data collection at corporate training headquarters.

Further Research in Training Evaluation

Industrial training evaluation is a relatively new area which has not had a great deal of coverage in the personnel, organizational communication, or industrial psychology literature. The publications that have appeared on training evaluation generally do not draw from an extensive theoretical background and thus are both limited in their applicability and disparate in their approach. Many of the articles are of a "how to" nature, or attempt to develop a theory for one particular situation. There is much practical material in the training evaluation literature, but it does not seem to have a common theoretical base. Therefore, future research in training evaluation needs to occur in three areas.

The first area is development of theory and foundations. There is a common thread to industrial training evaluation and it should be identified. This type of research simply attempts to locate the parameters which commonly occur in training evaluations, without necessarily identifying solutions. Perhaps this type of research is as concerned with definition of the area, as it is with the development of rudimentary theory. The model developed in this study is such an attempt to synthesize the different foci of training evaluation which have appeared in the literature.

The second area for future research is the definition of criteria for effectiveness in industrial training evaluation. The types of training evaluation vary widely and obviously they cannot be judged against the same specific criteria. But there must be some principle which applies to any program which attempts to evaluate training in an industrial setting. It is also just as important to define how criteria for effectiveness differ in different training situations. Clearly, the exposition

of the various goals inherent in different training evaluation programs will be the prime component in the development of effectiveness criteria.

The third area for future research is the identification of key contingencies and responses in training evaluation situations. Once the field itself and the criteria for effectiveness are defined, the next step is the development of theory to explain and predict effective training evaluation programs. This type of research is looking for the "differences which make a difference" and is attempting to identify the correct options under each of those contingencies. For example, Kirkpatrick has identified the level of evaluation as an important contingency; however the correct response to that choice has only been hinted at, not fully elaborated. Another contingency in the development of effective training evaluation is clearly the purpose of the evaluation: justification, refinement, credibility, motivation, etc. Again though, the correct prescription for each of those choices has not been developed and tested.

Future research in this area would choose the important contingencies and through a case study or experiment, test for the optimal response to that contingency. On the basis of this type of research, the original models or theories will be confirmed, refined, rejected, or elaborated.

A FINAL NOTE

It was discovered rather early in this research that all of the information on TNA, formal and informal, could not possibly be gathered. Time, finances, and geography combined to limit the extent of the information gathered in this investigative effort, and some choices were

made on the "point of diminishing returns" of additional information. Thus while some conclusions have been drawn about the effectiveness of TNA, the qualification should be added that all of the existing information on the TNA program has not been collected. Hopefully though, the most important elements of the program have been covered.

In spite of its limitations, the study afforded a fine opportunity to explore and critique the workings of one actual corporate training evaluation program, along with presenting an exposition of the training evaluation area in general. And while the focal point of this study was training evaluation, the emphasis throughout the research was on the general concept of evaluation. The framework that was set up to evaluate training evaluation programs could be applied to any number of types of organizational programs; the basic process for analysis remains the same. Clearly, cogent and effective evaluation is a prime requisite for assuring the wise use of resources and the optimal performance of both people and programs in organizations.

BIBLIOGRAPHY

Books

- Bass, Bernard M. and James A. Vaughan. Training in Industry: The Management of Learning. Belmont, California: Wadsworth, 1966.
- Belasco, James A. and Harrison M. Trice. The Assessment of Change in Training and Therapy. New York: McGraw-Hill, 1969.
- Campbell, Donald T. and Julian C. Stanley. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally, 1963.
- Campbell, John P., Marvin D. Dunnette, Edward E. Lawler III, and Karl E. Weick, Jr. Managerial Behavior, Performance, and Effectiveness. New York: McGraw-Hill, 1970.
- DePhillips, Frank A., William M. Berliner and James J. Cribbin. Management of Training Programs. Homewood, Illinois: Richard D. Irwin, 1960.
- Evaluating Training Programs. Ed. Donald L. Kirkpatrick. Madison, Wisconsin: American Society for Training and Development, 1975.
- Farace, Richard V., Peter R. Monge, and Hamish M. Russell. Communicating and Organizing. Reading, Massachusetts: Addison-Wesley, 1977.
- French, Wendell. The Personnel Management Process: Human Resources Administration. New York: Houghton Mifflin, 1964.
- Gane, Christopher. Managing the Training Function. London: George Allen and Unwin Ltd., 1972.
- Garbutt, Douglas. Training Costs. London: Gee and Co., 1969.
- Goldstein, Irwin L. Training: Program Development and Evaluation. Monterey: Brooks/Cole, 1974.
- Hamblin, A.C. Evaluation and Control of Training. London: McGraw-Hill, 1974.
- Handbook of Evaluation Research, Vol. I. Ed. Elmer L. Struening and Marcia Guttentag. Beverly Hills: Sage Publications, 1975.
- Lynton, Rolf P. and Udai Pareek. Training for Development. Homewood, Illinois: Richard D. Irwin, 1967.
- Patton, Michael Q. Utilization-Focused Evaluation. Beverly Hills: Sage Publications, 1978.
- Rogers, Everett M. and Rekha Agarwala-Rogers. Communication in Organizations. New York: Free Press, 1976.

Thorndike, Robert L. Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Warren, Malcolm M. Training for Results. Reading, Massachusetts: Addison-Wesley, 1969.

Articles

Alden, Jay. "Evaluation in Focus." Training and Development Journal, 32, No. 10 (1978), 46-50.

Blumenfeld, Warren S. and Max G. Holland. "A Model for the Empirical Evaluation of Training Effectiveness." Personnel Journal, 50(1971), 637-640.

Brethower, Karen S. and Geary A. Rummler. "Evaluating Training." Training and Development Journal, 33, No. 5 (1979), 14-22.

Bunker, Kerry A. and Stephen L. Cohen. "Evaluating Organizational Training Efforts: Is Ignorance Really Bliss?" Training and Development Journal, 32, No. 8 (1978), 4-11.

Chabotar, Kent J. "The Logic of Training Evaluation." Personnel, 54, No. 4 (1977), 23-27.

Coffman, Linn. "An Easy Way to Effectively Evaluate Program Results." Training and Development Journal, 33, No. 8 (1979), 28-32.

Fraser, Richard F., John W. Gore, and Chester C. Cotton. "A System for Determining Training Needs." Personnel Journal, 57(1978), 682-697.

Gilbert, Thomas F. "Praxeonomy: A Systematic Approach to Identifying Training Needs." Management of Personnel Quarterly, 6, No. 3 (1967), 21-33.

Hinrichs, John R. "Personnel Training." Handbook of Industrial Psychology. Ed. Marvin D. Dunnette. Chicago: Rand McNally, 1976, 845-860.

Horrigan, J. Thomas. "The Effects of Training on Turnover: A Cost Justification Model." Training and Development Journal, 35, No. 7 (1979), 3-7.

Kirkpatrick, Donald L. "Evaluating Training Programs: Evidence vs. Proof." Training and Development Journal, 31, No. 11 (1977), 9-12.

Moore, Michael L. and Philip Dutton. "Training Needs Analysis: Review and Critique." Academy of Management Review, 3 (1978), 532-545.

Newstrom, John W. "Catch-22: The Problems of Incomplete Evaluation of Training." Training and Development Journal, 32, No. 11 (1978), 22-24.

Olivas, Louis. "Auditing Your Training and Development Function." Training and Development Journal, 34, No. 3 (1980), 60-64.

Peloquin, Jerome J. "Training: The Key to Productivity." Training and Development Journal, 34, No. 2 (1980), 49-52.

Steinmetz, Cloyd S. "The History of Training." Training and Development Handbook: A Guide to Human Resource Development. Ed. Robert L. Craig. New York: McGraw-Hill, 1976, 1-17.

Thompson, John T. "How to Develop a More Systematic Evaluation Strategy." Training and Development Journal, 32, No. 7 (1978), 88-93.