PN-ACH-783

71238

FAMILY PLANNING TRAINING EVALUATION MANUAL

BY

EDWARD A. SCHOLL

DEVELOPMENT ASSOCIATES, INC.
2924 COLUMBIA PIKE
ARLINGTON, VIRGINIA 22204

1988

TABLE OF CONTENTS

Page

LIST OF TABLES AND GRAPHS v

PREFACE vii

INTRODUCTION 1

CHAPTER I: STAGES OF TRAINING EVALUATION 5

Needs Assessment 10

Evaluation of Training Inputs 13

Evaluation of Training Processes 18

Evaluation of Training Outcomes 22

Training Evaluation Feedback 34

CHAPTER II: DESIGN OF TRAINING EVALUATIONS 39

Validity of Training Evaluation Designs 40

Characteristics of Training Evaluation Designs 51

Sampling 61

CHAPTER III: DEVELOPMENT OF TRAINING EVALUATION INSTRUMENTS 73

Validity of Training Evaluation Instruments 73

Reliability of Training Evaluation Instruments 75

Developing Evaluation Instruments 81

CHAPTER IV: ANALYSIS AND INTERPRETATION OF TRAINING EVALUATION RESULTS 107

Levels of Measurement 107

Descriptive Statistics 109

Inferential Statistics 127

CHAPTER V: SPECIAL ISSUES IN TRAINING EVALUATION . 141

Evaluation of Quality of Services Provided 141

Cost-Effectiveness Analysis 149

APPENDIX: EXAMPLES OF TRAINING EVALUATION INSTRUMENTS 165

BIBLIOGRAPHY 201

LIST OF TABLES AND GRAPHS

Figures		Page
Figure 1	Stages in the Evaluation of Training	6
Figure 2	Followup Plan	38
Figure 3.1	Frequency Polygon	113
Figure 3.2	Histogram	113
Figure 3.3	Pie Chart	114
Figure 4.1	Positive Correlation	123
Figure 4.2	Negative Correlation	123
Figure 4.3	No Correlation	124
Figure 5	Learning and Performance Objectives	146
Tables		
Table 1	Guide to Training Evaluation Examples Used in the Manual	ix
Table 2	Table for Determining Sample Size	71
Table 3	Comparison of Evaluation Instruments	106
Table 4.1	Frequency Distribution	110
Table 4.2	Grouped Frequency Distribution	111
Table 5.1	Posttest Scores	118
Table 5.2	Calculation of Variance	120
Table 6	Calculation of Pearson's r	126
Table 7.1	Scores of Pretests and Posttests	129
Table 7.2	Calculation of t-Value	131
Table 7.3	Table of t-Values	133
Table 8	Followup Training Evaluation Results	135
Table 9.1	Rate of Return Visits Made by Type of Counselor Group	137
Table 9.2	Rate of Return Visits Made by Type of Counselor Group With Expected Frequencies	138
Table 9.3	Table of Chi-Square Values	140
Table 10	Training Expense Categories	156

ACKNOWLEDGEMENTS

This manual was prepared under Agency for International Development Contract DPE-3031-C-00-4078-00, Family Planning Training for Paramedical, Auxiliary and Community (PAC) Personnel II. Development Associates, Inc. is the PAC II contractor for the Latin America/Caribbean Region. An earlier version of this manual was used as a resource tool in a training evaluation workshop held in Guatemala and reviewed by over a dozen senior trainers and evaluators in the Latin America/Caribbean Region. In addition to thanking these individuals for their input, the author would like to thank professional colleagues at AID, Development Associates, and several cooperating agencies for their careful review and valuable comments.

PREFACE

This manual is intended to be an aid for those persons who are responsible for evaluating training, particularly the training of health and family planning paramedical personnel. While most of the concepts presented in the manual can be applied to the evaluation of training in general, the examples used draw upon the author's experiences in the evaluation of training of family planning workers in Latin America.

Given the contextual focus of the manual, certain assumptions have gone into its writing that may be untrue for other regions of the world and particularly for developed countries such as the United States. Among these assumptions is the fact that virtually all Latin American family planning agencies are dependent to some degree on outside donors. Many of these donors impose certain conditions on the agency such as an evaluation of the program or training being funded. Hence, the type of training evaluation that should be performed is, in some cases, predetermined by outside donors.

Other assumptions that have gone into the writing of this manual include the social/economic/political context of Latin America. To take one example, the means of communication such as telephones and postal service are often unavailable or unreliable in many rural parts of Latin America. These realities must be taken into consideration by evaluators who propose, for example, to depend upon telephone or mailed survey instruments to evaluate the effects of training.

Likewise, the severe economic constraints that most family planning agencies work under in Latin America must be kept in mind when planning evaluations of any kind. Obviously an agency cannot afford to finance a large-scale, rigorously controlled experimental evaluation design that requires the input of statisticians and computers when it cannot afford to provide minimal levels of service to its target population. What is needed is an <u>appropriate</u> level of evaluation that provides answers to the key questions that are needed in order to improve training and service delivery. It is

towards this goal, helping to facilitate the design of appropriate training evaluations that will be used to improve services, that the manual is written.

In an effort to make this manual easier to use, the following guide supplements the table of contents by showing the reader where to go in the manual to find discussion on common training evaluation questions. In addition, Table 1 presents a guide to the evaluation examples used in the manual so that the reader can refer to specific examples of interest. The evaluation stages listed in Table 1 refer to the stages as described in Figure 1 in Chapter I.

Question	Pages to Refer to
What should I evaluate before, during and after training?	6,5-34
How can the impact of training be measured?	25-34,179-200
How can I evaluate the impact of training when so many other variables enter into the picture?	32-33,42-46
How can I tell if my evaluation results are statistically significant?	127-140
How big of a sample size do I need when I want to do a followup evaluation?	67-71
How should I evaluate training if I only have a small amount budgeted for evaluation?	9-10,41,69,84
How can attitudes and skills be evaluated?	86-93,100-102,173-175
How can I use training evaluation to find more efficient ways of training?	149-164
How can I make evaluation reports more readable and understandable?	109-127
How can the quality of service delivery be evaluated?	141-148,182-184
How should training evaluation instruments be constructed?	73-106
How can management training be evaluated?	29-30

Question	Pages to Refer to
How should I conduct a needs assessment to facilitate training evaluation?	10-12
What should I do with training evaluation results?	34-38
When do I need to use control groups in training evaluation?	42-50,58-61
How can I tell how much change is needed to show that training was a success?	127-140
Are pretests always needed?	43-44,52-53,60-61

TABLE 1

GUIDE TO EVALUATION EXAMPLES USED IN THE MANUAL

Pages in Manual	Personnel Trained	Type of Training	Area of Evaluation	Instrument Used	Question Addressed	Stage of Evaluation
1	Distributors	CBD	Course evaluation	Questionnaire	Was training appropriate for knowledge level of participants?	Process
1	Distributors	CBD	Relevancy of curriculum to job tasks	Interviews	Could distributors answer users' questions?	Job Performance
1-2	Distributors	CBD	Cost-effectiveness analysis	Existing service delivery data and cost data	What was the cost of training per user recruited for distributors vs. midwives?	Input
29-30	Financial Administators	Income Generation	Application of price sensitivity analysis	Questionnaire Financial records/ service stats.	Was analysis performed by trainee? Profit/loss picture of agency and effect on service delivery	Job Performance Short-term effects
39	Community Leaders	Family Planning	Knowledge gain	Pre/posttests	What increase in knowledge occurred as a result of training?	Development of KAS
39-40	Distributors	CBD	Effect of home visits on number of users served	Monthly activity reports of distributors	What was the increase in visits as a result of training? What was the increase in no. of users as a result of home visits?	Job Performance Short-term effects

Pages in Manual	Personnel Trained	Type of Training	Area of Evaluation	Instrument Used	Question Addressed	Stage of Evaluation
80	Promoters	Communication skills	Application of training in communication skills	Observations	Were the skills taught in training being utilized on the job?	Job Performance
104-105	Distributors	CBD	Effect of training on service delivery	Monthly reports by	How many new users were served by each newly trained distributor?	Short-term Effects
105	Promoters	F.P. Promotion	Effect of training on clinical service delivery	Existing clinical records	What increase occurred in family planning clinic patients after training the promoters?	Short-term effects
121-127	Distributors	CBD	Effect of knowledge gain on service delivery	Posttest scores/service delivery data	What correlation was there between trainees' test scores and no. of users served?	Short-term effects
129-132	Family Planning Workers	Family Planning	Effect of training on knowledge gain	Pre/posttests	Was there a statistically significant increase in knowledge gain as a result of training?	Development of KAS
134-136	Distributors	CBD	Effect of training on service delivery	Service delivery data	Was there a statistically significant difference in no. of users served by groups 1 and 2?	Short-term effects
136-139	Counselors	F.P. Counseling	Effect of training sterilization counselors on patient cancellation rates	Clinic appt. cancellation	Was there a significant difference between the cancellation rates for patients counseled by groups 1 and 2?	Short-term effects
145-148	Distributors	CBD	Performance of key tasks	Posttests/ supervisory visits	Did the trainees meet minimal performance criteria for this task during training? ...on the job?	Development of KAS Job Performance
152-164	Distributors	CBD	Cost-effectiveness of alternative training types	Classroom tests/cost data	What is the projected cost per trainee for alternative training types?	Input

INTRODUCTION

Evaluation is often a word that conjures up negative images. Particularly threatening is the concept of summative evaluation: the evaluation that passes final judgement on a project or activity and may result in the project being terminated. Yet most evaluation in the training field, whether formative (taking place during the training for the purpose of making immediate changes) or summative, is rarely done for the purpose of deciding whether training should continue or not. Usually the need for training is a given and the evaluation is done for the purpose of identifying how training can be improved.

This improvement of the training function can either take the form of fine tuning the type of training already being given, or it can require a different form of training altogether. An example of the former would be when a participant reaction questionnaire used as a course evaluation revealed that the instructor incorrectly assumed a lack of knowledge about certain subjects on the part of the participants and consumed a large amount of time discussing the already familiar topic. The evaluation results thereby led the instructor to issue a pretest at the beginning of subsequent courses in order to determine what the participants already knew about the subject.

An example of training evaluation leading to a change in the type of training given would be when results of interviews with community-based distribution (CBD) workers during a followup evaluation of their training revealed that the distributors were entirely unprepared to answer users' questions about side effects of contraceptives and when medical care should be sought. A review of the curriculum revealed that what was taught was only partially relevant to the tasks that the distributors were given and omitted mention of several key tasks entirely. In light of the evaluation, therefore, the training was redesigned to become more task oriented and the priority tasks were reinforced through in-service training.

Another example of how training evaluation results might be used to alter the type of training given would be a cost-effectiveness analysis that indicated that training given to CBD distributors resulted in their serving an average of 25 new family planning users during their first year of service. The training given to them cost an average of $50 per

participant; hence, the cost in terms of training per outcome realized was $2 per new user during the first year. If the equivalent cost of training traditional birth attendants to provide family planning services was $5 per new family planning user recruited, the institution might decide to henceforth train only distributors--if serving a large number of family planning users was the sole objective of the agency.

These examples point to the various levels of complexity by which training is evaluated. Despite their differences, however, they all have in common the objective of improving--either through modifying or entirely changing--the training function.

While it is easy to glibly say that all training evaluations should be done for the purpose of improving the training function, in the real world evaluations take place for a variety of reasons. Tradition is chief among them. Countless course evaluations are done with little thought going into them and with little analysis of the results. It is as if the evaluation is being done because "there has to be a course evaluation," or because institutional norms or precedents require one to occur. Yet after the pre- and posttests are scored and the participant course evaluation questionnaires are tabulated, the results are filed for posterity and the next course proceeds as the last one did.

Another reason why evaluations occur is because of outside funding requirements. Few international donors want to be accused of providing funds to an agency without taking a serious look at the effects of the training or project sponsorship. This usually requires some sort of outcome evaluation to be done in order to justify the money already spent, or in order to program additional funds in the future.

Each of these examples sought to improve the training by focusing on a different aspect of training. The first evaluation used a participant course evaluation form to generate constructive comments from the participants that led to the subsequent adoption of a pretest in the evaluation plan so as to better match the course content with the participants' knowledge levels. The second example used a followup evaluation to identify discrepancies between what the participants learned during training and what

they were required to do on the job. A change in the training curriculum was, therefore, brought about as a result of the evaluation. In the third example, a cost-effectiveness analysis was done to compare the cost-effectiveness of training different types of personnel. The evaluation results were used to help determine the type of personnel the program would subsequently employ.

Hopefully the reader can appreciate from these examples the varied forms that training evaluation can take and the practical benefits to an agency that can be realized from undertaking training evaluations. Some of the other areas that training evaluations might focus on, and the corresponding changes that could result from undertaking them, are suggested in the following examples:

- An evaluation of the various inputs of training (trainers, curriculum, materials, etc.) can be performed using participant reaction questionnaires, or observations by the trainers themselves, to determine the quality of each of the inputs.

- An evaluation of the level of knowledge and skills of participants at the end of training, using a posttest and observations of skills performance in various role-play situations, can be used to determine whether participants are competent enough to be certified to work for the agency, as well as to identify what continuing education needs individual participants may require.

- An evaluation of the job performance of ex-participants, using questionnaires and existing agency data on output measures, can be used to determine whether the knowledge, skills, or attitudes learned in training had any effect on the participants' job performance.

- An evaluation of the effect of training new CBD workers on the number of family planning users served by the agency can be used by the agency to help determine whether expanding the CBD program, or expanding some other program such as opening up new clinics, can help the agency to better meet its goals.

These examples demonstrate further the many uses of training evaluations and how they can be used to better the type of training an agency gives, as well as help an agency determine the effects of training so as to be able to better target training resources.

Whether training evaluation occurs because of contractual requirements, because of tradition, or because of a genuine interest on the part of somebody to improve training and make it more efficient, evaluations will largely be a waste of time unless they are well designed, properly implemented, correctly interpreted, and the results communicated to all concerned. This manual attempts to facilitate these steps by outlining major points to consider in the design of training evaluations, in the development of evaluation instruments, and in the interpretation of evaluation results. In addition, Chapter V of this manual addresses two special issues in training evaluation: how to evaluate the quality of services provided by an agency and how to perform cost-effectiveness analyses.

CHAPTER I: STAGES OF TRAINING EVALUATION

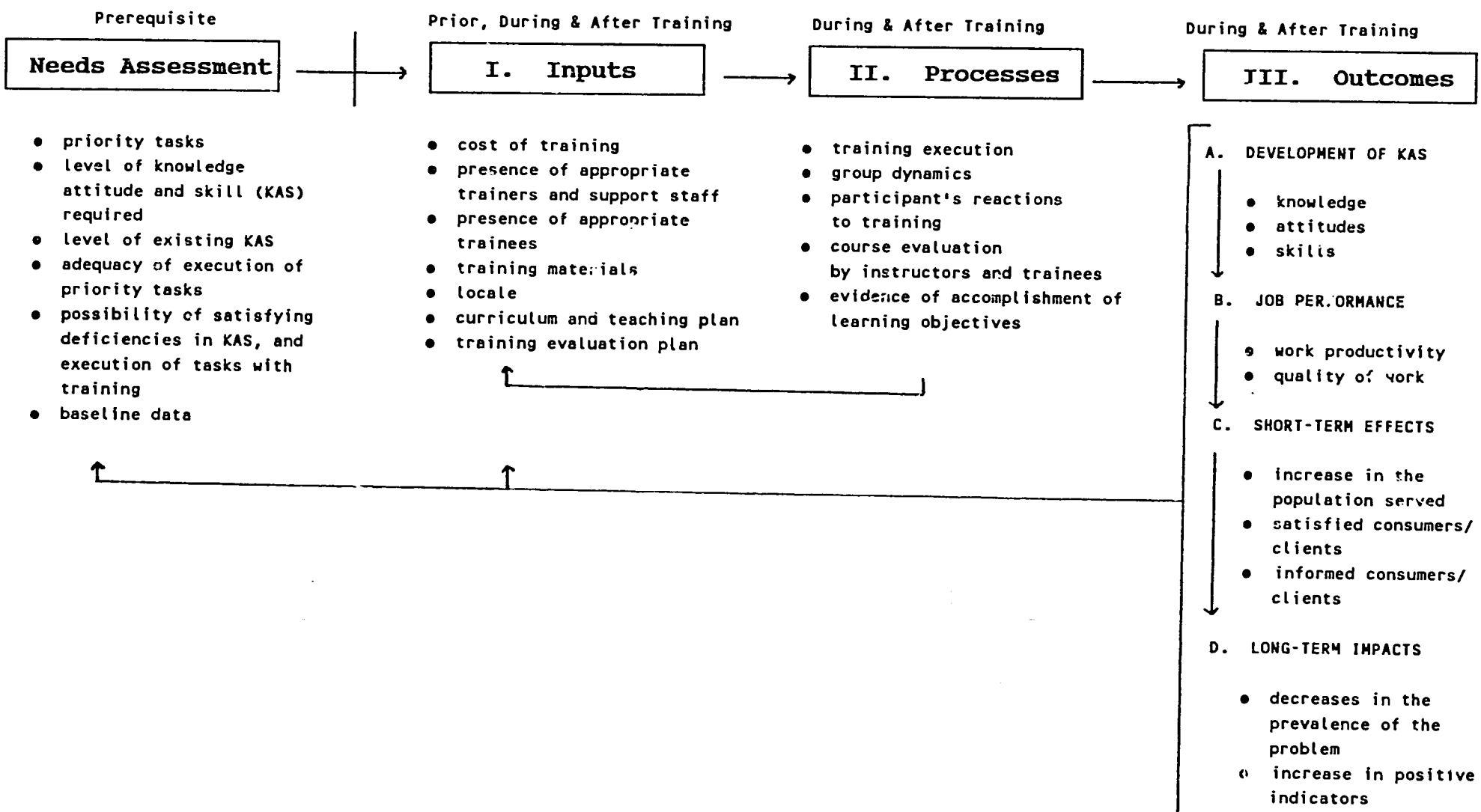
Training evaluation is not an activity that happens once at the end of a course. Training evaluation should occur both before, during, and after a training event. Even if a formal evaluation is not done at each of these stages, judgments need to be made (and are made, even if subconsciously), regarding the utility of the proposed training, the quality of its execution, the degree of learning by participants, and the resultant effect or impact of the training. Many times these judgments are made intuitively without articulating the conclusions reached. Yet whenever possible, these judgments should be made within the context of a predetermined evaluation plan with the results made available, and acted upon, by the appropriate people within the organization.

In order to help conceptualize the various stages of training evaluation and how they relate to one another, Figure 1 presents these stages along with examples of what is commonly analyzed within each stage. The three stages are preceded by a prerequisite stage which is an assessment of training needs. The stages shown in Figure 1 follow each other in a linear sequence by time. An evaluation of inputs is the first type of evaluation that should be performed prior to the initiation of the training event, as well as during and after the training event. Process evaluation occurs later, during the training itself as well as after the training. Outcome evaluation is the last stage of evaluation and is performed at the conclusion of training (in the case of assessing the development of new knowledge, attitudes and skills) and after training.

This last stage of outcome evaluation is further divided into four different sublevels of outcomes which follow each other in a linear sequence by time as well as complexity. The first sublevel, development of knowledge, attitudes, and skills (KAS), can be easily measured both at the conclusion of a training event (most commonly in the form of a posttest) as well as in followup encounters. The last sublevel on the other hand, long-term impacts, is rarely measured in the training evaluation field and would probably take many years, as well as a considerable evaluation budget, in order to measure.

FIGURE 1

STAGES IN THE EVALUATION OF TRAINING



The arrows flowing from stages II and III to stage I indicate that the information received there is fed back to trainers and managers who are responsible for determining the inputs that go into training so as to revise them for future training events. For example, if the results of the outcome evaluation measuring development of KAS reveals that the trainees learned little of the knowledge of contraceptive contraindications that they were supposed to, this information will be given to those responsible for designing the training curriculum (one of the training inputs).

There is also an arrow flowing from stage III to the needs assessment stage indicating the importance that outcome evaluation results have for subsequent needs assessments. If, for example, an outcome evaluation shows that a particular type of training resulted in no measurable improvement in a skill deficiency which the training was supposed to correct, then any subsequent needs assessment should note that that particular skill deficiency will likely not be met by the type of training in question. These examples illustrate the interconnectedness between the various stages of evaluation that make communication of the results a key issue in training evaluation.

Figure 1 is a generic representation of the various training evaluation stages. While the criteria listed under inputs and processes are fairly constant irrespective of the kind of training one is doing, the criteria evaluated as outcomes do vary depending upon the training type. Below are some of the outcomes commonly measured in the population field.

Development of KAS	Job Performance
• knowledge of contraceptive contraindications, usage, and side effects • attitudes that respect client confidentiality and elicit trust • attitudes that respect honesty and seek the good of the community • attitudes that foster informed choice • skills in clinical procedures • skills in communication • skills in administration and paperwork	• number of home visits made • number of group talks given • number of clinic references made • number of educational materials distributed • number of radio/T.V. spots broadcast • number of family planning distribution posts established • application of administrative/ management skills • quality of activities listed above

Short-term Effects	Long-term Impacts
• increase in contraceptive knowledge in target population • increase in users served • increase in client satisfaction • increase in couple-years-protection • improvement in the quality of services delivered by the organization • changes in professional, programmatic and administrative practices of the organization	• decrease in birth and fertility rates • decrease in maternal mortality rate • decrease in infant mortality rate

This framework for analyzing the various stages of training evaluation is the one that will be used in this manual but it is by no means the only one. Perhaps the most well-known and widely used framework for classifying areas of evaluation is a model developed by Donald Kirkpatrick[1]. In this model he identified four levels of evaluation that sought to answer the following questions:

Level	Questions
1. Reaction	Were the participants pleased with the program?
2. Learning	What did the participants learn in the program?
3. Behavior	Did the participants change their behavior based on what was learned?
4. Results	Did the change in behavior positively affect the organization?

This typology is attractive because of its simplicity but omits reference to an evaluation of inputs. Level 1 is roughly comparable to stage II of the framework presented in

1. Kirkpatrick, D.L., "Techniques for Evaluating Training Programs," A four-part series beginning in the November 1959 issue of the Training Director's Journal.

Figure 1. Likewise Level 2 is equivalent to stage III.A, and Levels 3 and 4 are equivalent to stages III.B and III.C respectively.

In outlining the various stages of evaluation and the questions commonly asked in each stage, it is not my intent to suggest that each evaluation stage needs to be performed for every training event that occurs. Doing so would undoubtedly bankrupt the training department within a few months. This chapter merely attempts to show some of the possible areas that a training evaluation can focus on.

Training evaluators, like other managers and administrators, need to be cost conscious and know when a particular type of evaluation is worth the time and money required to undertake it, given the likely information that it will provide. When an agency feels fairly confident that the training it is offering is both effective and efficient (having a high output/input ratio) there is probably little reason to invest substantial sums of money in an elaborate evaluation design. It is probably sufficient to just evaluate whether the objectives of each training event were met and continue to fine tune the training through input and process evaluations.

When a type of training is new, or when its effectiveness or efficiency are in doubt, an agency will probably want to invest proportionately more in evaluating the training. In such a case the other outcome measures such as job performance and short-term effects may warrant close analysis. The point that should be made is that training evaluation needs to be looked at as an investment that costs money but than has the potential of reaping rewards in terms of information that can be used to better the training function and the agency overall. Skill and experience are required to know how to maximize the benefits and minimize the costs of an evaluation so as to achieve the greatest return on the investment.

As a starting point, a training evaluator ought to consider what the present investment in training is costing the agency and then consider how much potential savings could be realized by making the training either more effective or more efficient or both. If the potential savings are greater than the projected cost of the evaluation, it is probably a good idea to undertake the evaluation. If the reverse is true then an evaluation should

probably not occur, or else only a very inexpensive evaluation should be carried out. Other situations where it would probably not be advisable to perform an evaluation are where the results will not likely be used, or where the sponsor of the evaluation is strongly motivated to prove or disprove something. The latter case would undoubtedly bias the evaluation process and lead to invalid results.

The remainder of this chapter outlines each of the various stages of evaluation and then discusses the all-important topic of training evaluation feedback.

NEEDS ASSESSMENT

Conducting a needs assessment is usually not thought of as being part of the training evaluation process but the two are vitally linked. An evaluation of training requires the prior existence of a needs assessment for three principal reasons:

- to detect deficiencies in workers' performance and prioritize training needs;
- to determine whether the dificiencies noted can be corrected or improved through training; and
- to establish baseline measures to document changes that occurred as a result of training.

Analysis of Workers' Performance

Needs assessments should begin with a comparison between the actual performance of various categories of workers and the level of performance sought from them. In this way any deficiencies can be noted and prioritized. For example, if a needs assessment points to the fact that the most serious deficiency among family planning educators is their communication skills, and not their levels of knowledge, then any future training given them should be designed to focus on imparting communication skills. Likewise, the evaluation should be designed to measure the acquisition of communication skills and their application on the job.

Training: Yes or No

Another function of conducting a needs assessment is to determine whether training is needed at all. Oftentimes needs assessments, like evaluations, are performed perfunctorily due to tradition or funding requirements. Managers and trainers may be already convinced that training is needed and know exactly what type of training they will perform. Yet for the most part, training should not be given until a documented need exists and until it is determined that training can meet that need in a cost-effective manner (see Chapter V for a discussion of cost-effectiveness).

Many times needs assessments point to existing needs that training cannot solve. If a CBD program is experiencing a lot of client dissatisfaction with the distributors, the problem could be that the distributors have a knowledge deficiency or attitude problem (which training may be able to correct), or the problem could be structural. For example, distributors may not be getting contraceptives delivered to them on a timely basis resulting in client dissatisfaction. In this case training the distributors will not correct the problem. Other factors relating to program success that training may be unable to impact upon include the organizational environment, staff morale and working conditions. In short, a host of factors can be identified as problems that need addressing, but training may or may not be the solution to those problems.

Baseline Data

Finally, a needs assessment is important to perform in order to document <u>changes</u> that occurred as a result of training. A good training needs assessment will include in it baseline data on various levels of knowledge or skills or job performance that training will attempt to improve. When measuring the post-training levels of knowledge or skills or performance of certain tasks, these results will have meaning only when compared to their pre-training levels.

Needs assessments should, therefore, always take place (whether formally or informally) prior to the initiation of training with their results reviewed by the staff responsible for developing the training evaluation plan and appropriately acted upon. Ideally, training needs assessments should seek to answer the following questions:

Question to be Answered	How this can be Measured
1. What are the key tasks that workers are expected to perform?	Review of institutional norms; interviews with program managers, supervisors
2. What is the minimum required of a worker in terms of knowledge, attitudes and skills in order to perform these tasks in an acceptable manner?	Review of institutional norms; interviews with program managers, supervisors
3. What is the present status of workers in terms of the required levels of knowledge, attitudes and skills? (quantitative if possible to serve as baseline data)	Interviews with workers, supervisors; tests; observations
4. How well are workers presently performing their key tasks? (quantitative if possible, to serve as baseline data)	Interviews with workers, supervisors, program beneficiaries; tests; review of program records; observations
5. What additional abilities (knowledge, attitudes or skills) do workers need in order to perform their key tasks according to acceptable performance criteria?	Discrepancies between #2 and #3; discrepancies between #1 and #4; interviews with program managers, supervisors
6. Can these abilities (knowledge, skills, or attitudes) be transferred cost-effectively through training?	Intuitively; analysis of projected costs; small experimental sample
7. What will be the quantified indicators of successful training?	Mutual agreement between trainers, evaluators
8. What evaluation instruments will be used to to gather this information?	Mutual agreement between trainers, evaluators

After answering these questions and deciding to go ahead with plans for a training event to occur (assuming a positive answer to question #6), the training activity can then be planned. During the planning for the training event, the first stage of training evaluation

should occur which is the evaluation of training inputs.

EVALUATION OF TRAINING INPUTS

Training Costs

As indicated in Figure 1, this stage of training evaluation involves the assessment of all the resources, or inputs, that go into the training activity. Chief among these is an analysis of the costs of training and the sources of funding these costs. The analysis of costs is not done at this stage in order to reach a decision of whether the projected training will likely be cost-effective or not. This should already have occurred during the prior needs assessment stage as part of the process for determining whether training should proceed or not. The analysis of costs during the evaluation of training inputs stage is done for the purpose of outlining in greater detail all of the expenses that go into the training (this will facilitate a retroactive cost-effectiveness analysis that can occur after the training) and the sources that paid for them (whether in the form of donated goods, fees, grants, or loans).

This analysis is important because it will have implications for the training given. To state an obvious example, if it is determined that the cost of contracting a leading authority on reproductive health to give a course will result in the training department's budget being depleted for the next six months, a knowledgeable source that charges a lesser fee will have to be contracted. Likewise, if it is determined that the costs of producing a video for training are unreasonable, a different medium will have to be utilized to convey information to the participants. Note that there is some overlap between the cost analysis done during the needs assessment stage (which is more generalized and seeks to answer the question "will training solve this problem in a cost-effective manner?") and that done during the evaluation of training inputs stage. The principal difference is that the latter seeks to elicit more detailed data on costs for the purpose of determining the exact form that training should take.

Training Personnel

Other inputs that are important to analyze prior to training include the personnel who will take part in the training, whether as participants, trainers or support staff. The proposed trainees, trainers and support staff should all be the subject of an input evaluation in order to assure that they are appropriate for the type of training and that there are neither too few nor too many of them. This evaluation should include a review of biodata forms for both participants and trainers to make sure they have the appropriate amount of education and experience that will allow them to perform their roles adequately, as well as a review of the training plan and size of the training group to assure an appropriate trainer/trainee ratio.

Training Locale

The proposed locale for training should also be subject to an input evaluation. The appropriateness of room size, physical comfort, pleasant surroundings free from distractions and its geographic location relative to the participants are all important aspects to consider. Likewise, the presence of appropriate training materials, from pens to audiovisual equipment, should be given prior thought to and made part of the overall input evaluation.

Curriculum and Teaching Plan

Probably the most important input to analyze carefully is the curriculum and teaching plan that will be used in the training. This includes both what will be taught, how it will be taught, how long the training will last, and what the learning objectives will be. A good needs assessment will allow the trainer to develop learning objectives that will hopefully be relevant to the needs of trainees and attainable through the process of training. When measurable learning objectives have been decided upon, the trainer can then proceed to develop a curriculum, including the training content and methodology, that will facilitate accomplishing the learning objectives.

Training Evaluation Plan

A final input of training is the training evaluation plan. Developing a training evaluation plan entails deciding what should be evaluated, who should be the subject of the evaluation, who should perform the evaluation, what instruments should be used, when the evaluation should occur, and what should be done with the evaluation results. Of all the inputs that are usually examined prior to training, the evaluation plan is often the one that is most commonly left out. This is unfortunate in that ad hoc evaluations, planned during or after the training has occured, are usually much less effective than if they had been planned prior to the training event.

When developing the evaluation plan, the teaching plan, with its specific learning objectives, should be followed closely so that the two plans are integrated. For example, if one of the learning objectives is to increase the ability of nurses to communicate better with family planning clients, then the evaluation plan should specifically address how this objective will be measured and how its degree of attainment will be evaluated. The plan should specify what instruments will be used to evaluate the nurses' communication skills, when they will be applied and by whom, what will constitute an "increase" in communication ability, and how the results will be used. Naturally, developing the evaluation plan is made easier if the learning objectives are stated in measurable terms, e.g. "as a result of training, the nurses will be able to score 75% or better on a communication ability scorecard when engaging in a simulated encounter with a family planning client."

The importance of developing a teaching plan and an evaluation plan together, prior to initiation of the training, can be seen in the example just given. If one of the learning objectives was the one stated above, then the evaluation plan would have to include the development of a "communication ability scorecard" that would be given to all trainees before and after training. Developing this instrument would undoubtedly require a lot of time and effort and it would be important to identify as a necessary input early in the

planning process so that it could be prepared in time for the training event.

Occasionally, a learning objective will be formulated which has no feasible, or economical, way of being evaluated. Take, for example, a learning objective such as the following: "As a result of training 50 CBD workers in one province, the contraceptive prevalence rate in that province will rise 10 percent when measured three months after the training, when compared to its level at the beginning of training." There is nothing wrong with this objective from a theoretical standpoint. It states in measurable terms what will be evaluated and when it will be evaluated. However, an evaluation plan would most likely recognize this as an unfeasible learning objective to evaluate. The lack of a control group, the small amount of change likely to occur, and the tremendous expense associated with doing statistically valid contraceptive prevalence surveys would all make this type of learning objective inappropriate, if not impossible, to evaluate in most circumstances.

This example again illustrates how developing an evaluation plan, and really thinking about how a learning objective might be evaluated, is important to perform in conjunction with the teaching plan prior to training. Not only does it alert trainers and evaluators as to what kind of instruments they need to prepare, but it can alert trainers as to the possible inappropriateness of certain types of learning objectives that cannot be easily evaluated. More will be said about the development of evaluation plans in the next chapter on the design of training evaluations.

A summary of the key questions to be looked at during the evaluation of inputs stage and how they can be measured follows.

Question to be Answered	How This can be Measured
1. What ar , the projected costs of the training?	Line by line budget preparation and analysis
2. From what sources will these costs be paid for?	Review of internal financial resources and outside funding sources
3. How can training be tailored to make it more cost-effective?	Detailed analysis of #1 and expected outcomes compared to alternatives
4. Are the projected trainees appropriate for the training?	Review of training needs assessment information; review of training curriculum; individual biod ta forms; pre-training interviews with selected applicants
5. Are the trainers appropriate for the type of training?	Interviews with trainers and their supervisors; review of trainers' curricula vitae
6. Will there be sufficient support staff to facilitate training?	Review of training curriculum and resources needed, errands to run, etc.
7. Is the training location and facility appropriate?	Visit to proposed training site; reports from those who have used it before
8. Are training materials well suited to training?	Review of materials to be used; reports from those having used the materials before; focus groups
9. Are the curriculum and proposed training methodology appropriate for the subject and the students?	Review of teaching plans (content and methodologies); discussions with other trainers who have taught the subject; review of past evaluations for similar types of training
10. Is the training evaluation plan appropriate?	Review of plan, including its projected validity, reliability, and cost; discussion with other trainers/ evaluators who have evaluated similar types of training in the past
11. Are the learning objectives appropriate and compatible with the evaluation plan, and vice versa?	Review of learning objectives and evaluation plan, including projected validity, reliability, and cost

After answering these questions and making appropriate corrections or revisions to the training plan or evaluation plan, the training can then proceed. It should be noted that these questions will be reanalyzed during the process and outcome evaluation stages

when these inputs will have their effect and be judged again by the participants and trainers during and after the training. For example, the appropriateness of the training curriculum is not only evaluated prior to training (in the context of input evaluation), but also during the training (in the context of process evaluation) and after the training (in the context of outcome evaluation). This is the significance of the return arrows in Figure 1 that show information flowing from the process and outcome evaluation stages to the input stage where these inputs will be restructured for future training events according to the feedback received in the subsequent stages of evaluation.

An analogy from the field of carpentry can perhaps make these relationships a little clearer. If nails are one of the principal inputs that a carpenter uses to build a house, the quality of the nails should first be analyzed in the factory where they are manufactured, as well as by the carpenter before he buys them (input evaluation). The nails will next be judged as worthy or not by the carpenter as they are driven into the wood during construction of the house (process evaluation). If the nails are of poor quality, the store or factory that sold or produced the nails may hear about their poor quality from the carpenter in the form of a complaint (feedback from the process evaluation stage to those responsible for producing the inputs). Finally, the nails may be evaluated a third time in terms of their ability to hold the house together over time (outcome evaluation). While nails rarely fail in this regard, the speed with which they rust or some other performance criteria may be used to rate them. Again, poor results may find their way back to the producer of the nails (feedback from the outcome stage to the input stage).

This analogy is appropriate for the training evaluation field except for the fact that training evaluation results, whether good or bad, should always be communicated to all concerned and not just when complaints or negative findings arise.

EVALUATION OF TRAINING PROCESSES

Process evaluation in the training field involves assessing three principal issues:

- how well the training conformed to its design;
- the context in which the training occurred; and
- how the training was judged by both participants, trainers, and at times, outside evaluators observing the training.

Deviation from Training Plan

In spite of well designed lesson plans, a training event usually deviates somewhat from the predetermined lesson plan. A function of process evaluation is to determine to what degree the training did go as planned. This will help to judge whether the outcomes of the training can be credited to the training design, or to the particular set of unique circumstances that unfolded during the training event. The input evaluation stage can only provide an approximation of how the training event will occur as it is looking at the training resources on the shelf as it were, before they have actually been put to use. In process evaluation, these inputs (curriculum, trainers, etc.) are reevaluated as they are put into action and exert their effect.

Context of Training

Process evaluation also involves judging the context in which training is occurring, or did occur. The context includes the entire circumstances in which the training occurred. It includes the physical environment, the emotional and intellectual level of the trainers at a particular moment, and all the human interactions taking place. This is an important consideration to make when evaluating training because countless factors external to the inputs themselves can influence the quality of learning that takes place in a training event. Factors such as the weather, how much sleep participants had the previous night, the availability of refreshments, and the comfort of the chairs being used by participants can all influence their emotional state and the degree of receptivity they have to learn or interact with others. Therefore, the context of training should always be examined to the extent possible by trainers and evaluators. Sometimes this is most easily done by another trainer or observer sitting quietly on the sidelines and observing the entire training event. He or she can usually pick up many subtle but significant factors that

take place and influence the training, but may be unknown to the trainer giving the lesson.

Perceived Quality of Training

A third function of process evaluation is to have instructors and participants evaluate, or make judgments about, the training both during and after the training event. Process evaluation should include measurements of how the training is progressing so as to be able to make mid-course corrections to the training. This can be in the form of daily evaluations by the participants who are asked to evaluate the activities of the day and changes that should be made. It can also include quizzes and exams that test the progress made towards accomplishment of the learning objectives. The training can also be evaluated at its conclusion, or even after the fact, in the form of course evaluations completed by participants and instructors (and observers if present). These evaluations should not only focus on the participants and their progress and satisfaction with the training, but on the ability of the trainers as well. Examples 2 and 3 of the training evaluation instruments in the appendix provide examples of instructor evaluation instruments.

The distinguishing feature of training evaluation at this stage is that it solicits information from the participants and instructors in terms of how they perceived the quality of the training. It is largely qualitative in nature, although the quizzes and exams used to test accomplishment of learning objectives can be quantitative in nature. Note that the results of the final posttest scores that indicate what was ultimately learned as a result of the training are not included in process evaluation analysis. This consists of one of the outcomes of training and will be discussed later.

Virtually all of the questions listed under the evaluation of inputs stage are worth asking again during the process evaluation stage in order to determine whether the inputs actually had their intended effect on the participants. Some of the additional questions to address in the process evaluation stage include the following:

Question to be Answered	How This can be Measured
1. How much progress was made towards accomplishment of the learning objectives?	Subjective impressions of instructors, observers; results of quizzes and exams given during the course
2. How well did the instructors present the information? (skills, pacing, clarity, responsiveness, etc.)	Results of daily & final course evaluations by participants, instructors; reports by observers
3. What were the obstacles, if any, to learning that were present during the training?	Same as question #2
4. How did the participants assess the quality and usefulness of the training?	Daily & final course evaluations by participants; followup questionnaires or interviews with participants
5. How did the instructors assess the quality of the training?	Course evaluation by instructors and interviews with instructors

After asking the appropriate questions and obtaining answers to those questions, the information should be fed back to those responsible for designing the training inputs as indicated in Figure 1. This is necessary in order for training to improve the next time around. The only time it would not be necessary is if the training were a one-shot occurrence and no future training of a similar type would be performed. Of course it is not always necessary for training inputs to be redesigned solely on the basis of the results of one process evaluation. It could be that the circumstances of a particular training event were unique and that its evaluation findings should not be given undue weight. For example, if a particular group of trainees found a lesson plan to be too elementary and a waste of their time, it could be that the trainees were improperly selected and consisted of overly educated participants. The lesson plan could be perfectly appropriate for the group it was intended for in which case it should not be redesigned.

Three examples of process evaluation instruments are found in the appendix of this manual.

EVALUATION OF TRAINING OUTCOMES

Evaluation of training outcomes refers to the assessment of the results of training. While input evaluation looks at what went into training and process evaluation looks at how the training progressed, outcome evaluation looks at what the training produced. The results, or outcomes, can be analyzed at four different sublevels, as indicated in Figure 1. These sublevels are the development of knowledge, attitudes and skills (KAS), job performance, short-term effects, and long-term impacts. These sublevels are presented both in the order in which they usually occur (with development of KAS occurring first) as well as by order of difficulty and expense in carrying out (measurement of long-term impacts is very difficult and very expensive to carry out).

Development of Knowledge, Attitudes and Skills

Virtually every training event seeks in some way to measure whether the participants gained some new knowledge, attitude or skill as a result of training. Most course, or learning, objectives should be phrased in terms of new knowledge, attitudes or skills that the participants must demonstrate in a way that can be observed or measured. Learning objectives are sometimes stated in terms of what the instructor will do (e.g. the instructor will teach the contraindications to the pill) rather than in terms of what the participants will learn (e.g. the participants will be able to correctly state five contraindications to the pill). The latter type of objective is much more appropriate and a truer measure of whether training is having its intended effect.

The development of KAS are usually first assessed during the process evaluation stage when quizzes and exams are given during the training to test progress towards accomplishment of the learning objectives. The assessment at this stage is only preliminary since the training is still going on and further development may occur before the end of training. When a final exam or post-test is given at the end of training this marks the beginning of the outcome evaluation stage. Final scores or evaluations of the

development of KAS is the most immediate outcome of training and the one that is most commonly and easily measured.

Classifying the development of KAS as an outcome does not imply that trainees will necessarily produce anything or perform any function differently as a result of training. They may immediately quit the job for which they were trained, in which case the training could be considered to have no effect. Nevertheless, the training did result in something being produced (for example knowledge), even if the outcome produced was subsequently unused.

The development of knowledge, attitudes and skills are all important outcomes to consider in family planning or health worker training. Knowledge is important because most functions of family planning and health workers are not purely mechanical skills, but require some degree of knowledge in order for the workers' tasks to be performed correctly. Attitude development is also important because most health and family planning workers interact with the public and are usually promoting a certain product (oral rehydration therapy or condoms for example) or providing education which require certain attitudes on the part of the promoter in order to be effective. Finally, many skills are required for health and family planning workers such as the ability to fill out forms, and the ability to speak clearly and provide information and education to the public.

Usually trainers and evaluators seek to determine not only what level of KAS trainees possess at the end of training, but what new KAS they developed as a result of the training. This determination requires the collection of baseline data which is most commonly done through the application of pretests to the trainees, as well as in needs assessment measurements. More will be said about the design of pre- and posttests and other evaluation instruments in Chapter III.

One other point to be noted concerning the development of KAS is that it can be measured at the end of training, as well as after training in the form of knowledge (or

attitude or skill) retention tests. Retention of KAS is often important to measure since KAS quickly forgotten do little good for the trainee or the program. Retention of KAS is dependent, however, on many other factors besides the training. The quality of the supervision given, the availability of reference materials, and the amount of post-training experience that workers gain will all influence whether KAS are retained or not.

The questions that should be analyzed during this level of the outcome evaluation stage are summarized below. The areas of KAS that should be tested are, of course, those areas that are related to the learning objectives, which in turn should be related to the abilities necessary for successful performance of the job for which the trainees are being trained. Some of these were mentioned under the list of outcomes relating to the development of KAS on page 7.

Question to be Answered	How This can be Measured
1. What were the levels of KAS that the the participants were expected to attain as a result of training?	Review of course objectives
2. What levels of KAS did the participants in fact obtain?	Scores of posttests; observations and interviews with participants at the conclusion of training; practicum assessments by instructors and participants
3. What new levels of KAS did the participants develop?	Comparison of #2 with results of pretests and other baseline data; interviews with participants or supervisors to elicit recall of prior levels of KAS
4. What was the level of retention of KAS for the participants after training?	Scores of KAS retention tests given at some point subsequent to training

After analyzing the development of KAS as a result of training, the next logical focus for outcome evaluation is the effect of training on participants' job performance.

Job Performance

The assessment of training's effect on participants' subsequent job performance is perhaps the most important single aspect of training evaluation. Whereas the development of KAS measures skills (and knowledge and attitudes) acquisition, job performance assessment measures the transfer of these abilities to a specific job. As was mentioned previously, the development of KAS is the most commonly and easily measured outcome of training, yet it can have no correlation at all to subsequent work performance. At times, the brightest and most gifted individuals do not have the same motivation and desire to serve others that less educated trainees have and their high test scores do not translate into effective job performance. Since training is undertaken for the purpose of preparing individuals to do specific work-related tasks (as opposed to education which is for the general intellectual enrichment of the student), job performance is far superior to scores of KAS development as an indicator of the effectiveness of training.

Although an assessment of job performance is superior to assessing the development of KAS as an indicator of the successfulness of training, it does have its limitations. Suffice it to say that a trainee who returns to his or her job and begins to implement everything taught in the training course (for example how to do home visits to promote family planning) may or may not have any effect or impact on the target population. A promoter trained to do home visits may do five visits every day and not find one new person interested in family planning if, for example, the village just had a woman die of complications from pill use and all women in the village were thereafter suspicious of family planning. In short, effective performance of work-related tasks does not necessarily guarantee that any change will occur in the target population as a result of training.

Despite this caveat, assessment of post-training job performance remains one of the best measures of the effectiveness of training and gives a fairly good idea of the potential effect of training on the target population. It is also fairly inexpensive to carry out, easy

to design, and not subject to as many confounding variables or alternative explanations as are designs that seek to measure the short- and long-ter impacts of training on the target population. Some of the most commonly measured criteria used in evaluating training's effect on family planning workers' job performance were listed on page 7.

The basic steps to follow in performing an assessment of training's effect on job performance is first to determine what the priority tasks of a given worker are that should be the foci of the evaluation. Next it should be determined what was the level of performance of these tasks by the worker before training and after, both in terms of quantity and quality of performance.

These steps and the questions that correspond to them are the following:

Question to be Answered	How This can be Measured
1. What were the priority tasks which the participant was given training to perform?	Review of course objectives; interviews with supervisors, program directors
2. What are the acceptable performance standards, in terms of quantity and quality, that the participant is expected to perform?	Review of program norms; interviews with supervisors, program directors
3. How were these priority tasks performed by the participant prior to training?	Baseline data gathering through questionnaires given to trainee, supervisors; pre-training observations of trainees; post-training recall by trainees, supervisors; program records
4. How were these priority tasks performed following training?	Questionnaires given to trainees, supervisors; observations of trainees with performance checklists; interviews with trainees, supervisors, program beneficiaries; program records
5. Are there any alternative explanations for differences in job performance other than the effect of training?	Intuitive reasoning; analysis of program environment and changes that occurred; use of control/ comparison groups; review of other threats to validity

There is no definitive answer for the question of when followup evaluations that look at these questions should occur. A sufficient amount of time should lapse that allows the workers to apply what they have learned, yet the followup should be done soon enough so that threats to validity (such as inevitable change over time) are avoided. A common period of time to let lapse between training and followup measurement is six months. There can be more or less time allowed, however, depending on the type of tasks that are performed and the length of time trainees need to apply their newly developed KAS to their job. In general, newly developed manual skills (such as the ability to take blood pressure) take little time to put into practice on the job. Abstract skills such as program management take longer to put into practice and measure. If resources are sufficient, the best choice is to do a series of followup measurements of job performance at regular intervals.

Short-Term Effects

The evaluation of the short-term effects of training refers to the analysis of changes, observable in the short-term, that occurred in the target population. These effects are typically measured in the population field in terms of changes in the target population's knowledge of contraceptives, contraceptive prevalence, the number of family planning users served, level of satisfaction and continuation rates for the various methods used, and changes in couple-years-of-protection (CYP). Other measures of short-term effects that are sometimes used include changes in the target population's attitudes about family planning, including child spacing, changes in the percentage of users correctly using their methods, and changes in the percentage of users who attend medical exams when referred by a promoter or other health personnel.

It should be noted here that an analysis of short-term effects does not require elaborate contraceptive prevalence surveys and the like to be performed. Changes in the number of users served, or simply the number of users served as a result of training are also short-term effects. For example reporting the number of new family planning users

(new users to the program, not necessarily new to family planning) served by newly-trained CBD workers is a common, and entirely appropriate, measure of training's short-term effects. Of course, without the benefit of baseline measures and an analysis of whether these users obtained contraceptives from another source previously, the conclusions that can be made about training's effect will be somewhat limited. Chapter II will discuss the issue of training evaluation designs and how followup evaluations can be effectively and inexpensively designed that allow robust conclusions to be made.

The principal distinguishing feature between short-term effects on the one hand and job performance on the other is that in job performance analysis the subject of observation is the trainee. His or her activities are scrutinized, recorded and evaluated to determine whether the KAS that he or she learned in training have been applied on the job. The analysis of short-term effects on the other hand has as its object of analysis the target population. The evaluator is now examining the target population to see what effect the trainee's activities have registered. This effect may be as simple as saying that as a result of the home visits made by promoter X, five women have decided to begin planning their families who were not doing so previously.

The principal difference between short-term effects and long-term impacts is not the subject of analysis (they both analyze changes in the target population as a result of training) but rather the time period involved and the indicators used. Long-term impact analysis looks at the overall goals of a health or family planning program such as lowered fertility, mortality and morbidity and seeks to tie together in some way changes in those ultimate indicators and the contributions of the health or family planning program. Usually this takes a very long period of time to measure and requires very sensitive instruments and large sample sizes (with control groups). Short-term effects on the other hand are really proxy indicators of progress towards these ultimate goals of lowered fertility, mortality and morbidity. For example, an increase in the number of family planning users in a given target population is a proxy measure of lowered fertility in that population. These substitute measures, like use of contraception, are used instead of ultimate measures, like the fertility rate, because the former have a high

correlation with the latter and they are much easier and quicker to measure.

Management Training

A final note should be added concerning the training of personnel who will not have as their function direct service to the public. The examples used above assume that workers are being trained to serve the public in some way and hence indicators of such service are used as measures of job performance and short- and long-term impacts. Management training is a good example of where the evaluation of short-term effects would probably best utilize the organization itself, rather than a target population, as its object of analysis. In this case the short-term effects measured could be the changes that occurred in programmatic or administrative practices as a result of training. Suppose for example, that a financial administrator of a family planning organization attended a workshop on income generation, with one of the key topics being how to perform a price sensitivity analysis. The purpose of such an analysis is to determine how much prices can be increased for services without significantly reducing demand. The evaluation of the effect of training on the participants' job performance could, therefore, be based on whether or not the financial administrator performed a price sensitivity analysis for his or her organization after attending the workshop (utilizing a self-reporting questionnaire). The assessment of short-term effects could be an analysis of the immediate profit/loss picture of the organization following the price increases, together with service delivery statistics. The long-term impact of the training in financial management could then be measured in terms of the sustained financial picture of the organization over time and its progress in generating income year after year.

In this example a specific overt action (performance of a price sensitivity analysis) that was taught in the training and expected to later occur was used as the evaluation indicator. Sometimes indirect indicators are used that indirectly evaluate mastery of a certain subject matter. The amount of employee turnover might be an indirect measure of a manager's ability to implement the principles learned in a course on

communications and improvement of employee morale. However, the validity of such a measure would probably be called into question unless it could be demonstrated that there is a clear link between employee morale and successful application of the principles learned in the course. The validity of such a measure would also be suspect if employee turnover were a rare occurrence under normal circumstances and, therefore, not likely to show obvious improvement.

Another measure by which to evaluate management training is by the use of reports that the trainees complete specifying goals that they hope the course will help them achieve. The trainees can later be asked to report on the success they had in achieving those goals and in the process, provide an indirect measure of the effectiveness of the training.

These are just a few examples of how management training can be evaluated in spite of its more abstract nature and despite the fact that managers do not usually offer direct services to the public, thus making it more difficult to isolate observable outputs to measure. Of course, the difficulty associated with evaluating management training is really related to the difficulty of developing measurable training objectives. If the objectives have been adequately defined in measurable terms, then the process of evaluating management training is really no different than for any other type of training.

The principal questions to ask during the analysis of short-term effects are the following:

Question to be Answered	How This can be Measured
1. What changes have occurred in the target population since training in terms of selected criteria?	Collection of baseline data and comparison with post-training service delivery data; questionnaires and interviews with program personnel and program beneficiaries
2. To what degree was training responsible for these changes?	Intuitive reasoning; utilization of control groups; analysis of possible threats to validity
3. What changes occurred in the trainees' organization in terms of selected criteria?	Comparison of pre- and post-training organizational records, accomplishment of organizational goals; questionnaires and interviews with selected members of the organization
4. To what degree was training responsible for these changes?	Same as question #2
5. To what extent were the changes observed, expected?	Review of the objectives of training and the results of questions #1 and #3
6. What unforeseen changes occurred as a result of training?	Same as question #5
7. How will the changes that occurred, in either organization or target population, help to meet the organization's overall goals?	Review of changes and organizational goals

Most evaluations of training stop at this stage. An evaluation of the long-term impacts of training is usually only done in theory with conjectures made as to the ultimate impact of the training. Nevertheless, the final sublevel of training outcomes, analysis of long-term impacts, is worthwhile to consider if only to understand why it is so rarely done and in what circumstances it might be appropriate.

Long-Term Impacts

An evaluation of long-term impacts of training is rarely done in the population field because of the expense and difficulty in designing a valid study. The long-term impact sought in most population programs is a reduction in fertility levels, or related measures such as a delay in age of first marriage or an increase in birth spacing intervals. In addition, health indicators are often used to measure the long-term impacts of

population programs. These indicators include the infant mortality rate, maternal mortality rate and incidence of low birth weight babies. Finally, a reduction in the incidence of abortion is also sometimes expresses as one of the long-term impacts of population programs.

An obvious feature of all these indicators is that they happen rather infrequently. Births, marriages, infant deaths, maternal deaths, and abortions are all events that occur, at most, a few times during a woman's life. In contrast, the measures of short-term effects, such as the practice of contraception, often occur continuously and can be quickly and readily measured. Likewise, the measurement of the impact of health training on specific diseases is feasible only if those diseases have a high incidence. The impact of training mothers in oral rehydration therapy, for example, could be assessed relatively quickly in an area with high mortality due to diarrhea.

The first problem, therefore, in measuring the long-term impacts of family planning training is that extremely large populations or samples must be analyzed (and have been exposed to the intervention--training or trained personnel) before any changes can be noted in the indicators. This of course requires a great deal of time and money in order to accomplish.

The other principal hurdle faced in trying to measure the long-term impacts of training is validity. Measuring the validity of an evaluation design basically involves posing the question: Can you say with certainty that the results are true? Another way of phrasing the question is: Can you say with certainty that the evaluation results (lowered birth rate, for example) were actually caused by the intervention (family planning promotion by distributors, for example)? Unfortunately, these questions can rarely be answered in the affirmative. In the first place, results such as a lowered birth rate, besides being hard to measure, are very difficult to correlate with a given input such as training. A host of other factors could have caused or helped contribute to the lowered birth rate. Socioeconomic development, female education, contraceptive social marketing, governmental promotion and provision of family planning services, as well as many other

factors usually work together to bring about lowered levels of fertility.

The other design problem in trying to relate family planning training to long-term impacts is that training is usually not the sole input from a family planning agency that impacts a target population. Usually training is given to prepare workers to provide a service in the context of an overall family planning program. If the program accomplishes a reduction of fertility levels in the target population, should the training department receive the credit? Undoubtedly the entire program, including the functions of supervision, contraceptive logistics, information and education, training, and overall management should be given credit for the results. Thus, it is usually more appropriate to discuss the evaluation of long-term impacts in terms of what an overall program accomplished rather than a specific training event or events by themselves.

If a family planning or health program had sufficient resources to seek to measure its impact in terms of changes in fertility and health indicators, then it would have to attempt to minimize the threats to validity discussed in chapter II to the extent possible. A quasi-experimental design would most likely be employed with experimental and comparison groups large enough to ensure that there would be a sufficient number of cases to allow for confidence in the results. Some of the key questions that an evaluation of long-term impacts should address are listed below. In each case the word program is substituted for training in light of the problem discussed above concerning the difficulty of isolating the contribution of the training component.

Question to be Answered	How This can be Measured
1. What are the problems that the program is trying to solve in the long-term?	Review of organizational goals; review of fertility, mortality, morbidity levels in the target population
2. How is the program expected to impact on these problems?	Intuitive reasoning; review of the literature; analysis of short-term effects
3. What other factors extraneous to the program are likely to impact on these problems?	Intuitive reasoning; review of the literature
4. After measuring changes in the prevalence of the problem, to what extent was the program responsible for those changes?	Comparison of results from experimental and comparison groups; analysis of possible confounding variables

Although an assessment of the long-term impacts of training is rarely done in the population field, it is often useful just the same to analyze these questions theoretically in an effort to gain a better understanding of what potential contribution the training effort could make towards reducing the prevalence of the social problems that most family planning and health organizations seek to redress. Often this exercise will result in a rethinking of priorities and strategies in an effort to focus more clearly on impacting the most important of the problems identified by the organization.

TRAINING EVALUATION FEEDBACK

Whenever training is evaluated the results are invariably communicated to someone. The person who receives the results may be the agency's director, the head of the training or evaluation departments, or perhaps, an outside donor who financed the training. Yet oftentimes, the people who could utilize them most do not receive the evaluation results, or receive only generalized conclusions or overall pre- and posttest scores that do not provide sufficient detail. Among the people who should definately receive detailed reports of the evaluation results are those responsible for managing or producing the inputs and those responsible for supervising the trainees and providing them with continuing education and training.

...*to redesign inputs*

The people responsible for managing or producing the inputs include those who designed the training curriculum, those who managed the logistics of the training, those who designed any of the training tools such as manuals or flipcharts, those responsible for selecting both the trainees and trainers, and those who designed the evaluation plan, including the evaluation instruments. As was previously pointed out, these people need to receive evaluation feedback in order to better redesign the inputs that went into training. Even if the same type of training will never again be repeated, it is still valuable for those who designed it to receive feedback on how the trainees benefited from it. If nothing else, it will provide them will valuable lessons to apply to future experiences when they design training of a similar nature.

As Figure 1 pointed out, the feedback to those responsible for managing or producing the training inputs should occur after both the process evaluation and outcome evaluation stages. When only the results of course evaluations and pre- and posttest scores are provided, little information is revealed about how the information was actually utilized by the trainees. For example, a person who wrote a training manual that included a section on reproductive anatomy and physiology needs to know not just if the participants liked the section of the course that dealt with this topic and scored well on the questions of the test that covered it, but whether the participants are using that knowledge on the job. A followup interview with participants could reveal, for example, that after training the participants had little occasion to use their knowledge of reproductive anatomy and physiology and felt instead that the information they received on contraceptive complications was inadequate to prepare them for all the questions they encountered in their work.

...*to evaluate test results*

One aspect of feedback that should be kept in mind is that it needs to be specific in order to be helpful. Feedback of pre- and posttest results provides a case in point. The

reporting of overall pre- and posttest results is of little value in itself. What is much more valuable for curriculum designers, evaluators or supervisors to know is what specific questions the participants had difficulty with. One useful way of providing this information is to present a list of test items and summarize the number right and wrong for each item. Each question is then classified as "RR" (right-right), "WW" (wrong-wrong), "RW", or "WR" depending on how the majority answered the question on the pretest and on the posttest. The tabulations are then used to guide trainers in modifying future training. Items that are classified as RR and WW indicate no gain, and the training might be modified by reducing instruction on RR items and increasing it on WW items. Any RW classifications should be closely examined as well to see what caused the trainees to score worse after the training.

...*to better supervise trainees*

Just as those people who managed or produced the training inputs need to be given detailed results of any evaluations that occur, so do those responsible for supervising the trainees once they return to or initiate the work for which they were trained. In large CBD programs for example, the trainers will sometimes have little communication with the supervisors of the CBD workers and the only feedback the supervisors receive, if any, is a summary of the results of the pre- and posttest scores and information through the grapevine. This is obviously not a desirable situation in that the supervisors need to know in detail how the participants did in the course, what their strengths and weaknesses were in terms of knowledge, attitudes and skills, and what topics they as supervisors should be prepared to reinforce during supervisory visits with the CBD workers or during formal continuing education encounters. The same principle holds true for any personnel situation where a trainee will be supervised after training by someone else who did not provide the training.

The solution to the problem posed above is for there to be as much detailed communication as possible, in-person if feasible, between trainers who observed and measured the participants' KAS during the training, and the supervisors who will be

expected to supervise the participants, and in many cases correct any weaknesses in KAS as the ex-trainees perform their jobs. Ideally every trainer should prepare a followup plan for each participant following their training which would include the specific topics of the course, written observations on how the participant was evaluated on those topics, and what recommendations the trainer has for the supervisor(s) with respect to reinforcing certain topics that the participant did not grasp or perform well on during the training. An example of this type of followup plan is presented in Figure 2. If such a followup plan were prepared, the supervisor would then have an excellent tool to use when visiting the participants in the months following the training. He or she could give priority to supervising those aspects of training that the participant appeared deficient in during the training and thus make more efficient use of the supervisory visit.

Naturally communication of evaluation results should not be limited to those designing the inputs or those supervising the trainees. Everyone who can benefit from the lessons learned in the evaluation should receive detailed feedback.

FIGURE 2[2]

FOLLOWUP PLAN

Name of Participant: Daysi Dorado Dates of Training: 3-5 July 1988 Today's Date: 8/23/88

Task: Provide counseling on methods and program services

I. Clinical Services	Remarks	Recommendations
A. Voluntary Sterilization		As first step, needs more practice with counseling on methods. We recommend supervised on-the-job practice with emphasis on the pill.
B. IUD		
C. Injections		
D. Diaphragm	Knows what it is--cannot explain its use	At any convenient time, further information can be provided on diaphragm and periodic abstinence methods. Has minimum concept on these and is well informed on clinical referral procedure.
E. Periodic Abstinence	Has a basic concept--cannot explain the methods.	
F. Papanicolau Test		
G. Clinical Consultation		
II. Community Services		
A. The Pill - Its use	Errors in the posttest - what to do in case user forgets, omits details in her explanations	
B. The Pill - Side Effects		
C. The Pill - Danger Signs	Named four in posttest	
D. Condom - Its Use		
E. Vaginal Tablets - Their Use		
F. Foam - Its Use	Omits details in her orientations	

Comments: Ms. Dorado participated actively in the course, demonstrating great interest in learning. Has some problems in reading.

From: _____ Trainer To: _____ Immediate Supervisor Via: _____ Director, CBD Program

2. Taken from original Spanish version in Terborgh, A., et al, 1987. Capacitacion y Control de Calidad en Programas Comunitarias de Planificacion Familiar, Development Associates: Arlington, VA (unpublished).

CHAPTER II: DESIGN OF TRAINING EVALUATIONS

In the context of training evaluation, a design is a guide that specifies what, when, from whom, how, and by whom the training will be evaluated. The evaluation design can also be referred to as the evaluation plan and should always precede the training to be performed as indicated in Figure 1 where development of the training evaluation plan is listed under the inputs stage of evaluation. As was mentioned in the previous section, the evaluation plan should be developed in conjunction with the teaching plan so that the evaluation can be sure to address each of the learning objectives. It is helpful to always have a written plan that specifies exactly what will occur so as to assure that the various components of the training evaluation will take place in an orderly fashion and the necessary resources and preparations made ready.

An example of a training evaluation plan or design that is quite simple would be the following: At the beginning of training a group of community leaders in family planning, the instructor will give the participants a pretest that will measure their entry level knowledge of the subjects to be taught during the training. At the conclusion of training the instructor will give the same pretest to the students to measure changes in their knowledge levels. Note that this example answers the following questions:

- What will be evaluated?: --entry level knowledge of certain subjects;
- When will the evaluation occur?:--at the beginning and end of training;
- From whom will the results will be obtained?: --the students;
- How will the evaluation be conducted?:--through the use of identical pre- and posttests; and
- By whom the measurements will be gathered?:--the instructor.

This type of evaluation is an example of a non-experimental design (to be defined shortly) that assesses the effect of training on development of KAS.

Another example of an evaluation design would be the following: One month after

training a group of CBD workers in how to do home visits and motivating them to do so, the CBD program manager will determine the increase in the number of new users served by the program as a result of the training in and implementation of home visits by reviewing the distributors' monthly activity reports before and after the training, noting their number of home visits and new users. Again the example answers:

- What will be evaluated?:--increase in the number of new users served as a result of increased home visits;
- When will the evaluation occur?: --one month after training;
- From whom will the results be obtained?: --the distributors;
- How will the evaluation be conducted?: --review of the distributors' monthly activity reports; and
- By whom will the measurements be gathered?: --the CBD program manager.

This example, like the one above, demonstrates a non-experimental design.

Before discussing the characteristics of non-experimental and other types of evaluation designs, the concept of design validity should be introduced which is one of the principal determining factors as to which type of evaluation design will be utilized.

VALIDITY OF TRAINING EVALUATION DESIGNS

The validity of a training evaluation design basically encompasses the analysis of two questions: "Can I be sure that the evaluation results were actually caused by the intervention (e.g. training)?" This is the concept of internal validity. The second specific question is "Are the evaluation results generalizable to other people and settings with similar characteristics?". This is the concept of external validity.

Both internal and external validity should be of concern to trainers and evaluators whenever an evaluation design is developed so as to avoid undertaking an evaluation

that could prove to be a waste of time and resources given its lack of validity and, therefore, meaning.

On the other hand, one should not become so consumed with designing a thoroughly valid evaluation, eliminating every conceivable threat to validity, that the evaluation costs skyrocket. Controlling for threats to validity do cost money and in some situations the extra cost may not be warranted. Most training evaluations contain within them one or more potential threats to validity, yet still provide results that inspire confidence and are useful to training evaluators. The key to designing good evaluations is to know when a potential threat is likely to become an actual threat and distort the results of the evaluation to a degree that would render them unuseful. When this is likely to occur, extra resources should be allocated to the evaluation design to assure a valid study.

Another way of looking at the validity question is to try to make the quality (assurance of validity) of the evaluation design proportional to the importance of the training event. Evaluation costs should typically not run higher than 10 percent of the overall costs of training, including the costs of analysis, development, and delivery. Therefore, if the costs associated with removing every conceivable threat to validity result in the evaluation costs totalling 20 percent or more of the overall cost of training, then the training objectives should probably be rethought, or else a less rigorous evaluation design allowed. Likewise, routine training events, or training events whose evaluation results will not become precedents for many succeeding events, do not probably need to become concerned with eliminating every conceivable threat to validity.

The following discussion on how to counteract each of the possible threats to validity is offered, therefore, as a guide to alert the reader to where the myriad sources of error are that can compromise an evaluation design and how these sources of error can be counteracted. Each threat may not be applicable to every training situation, and even where they are, the training evaluator must use judgement in deciding whether the extra costs associated with controlling for the threats are worth the increased confidence that

one can have in the results.

Internal Validity

Going back to the first example of an evaluation design, where the effect of training will be measured on the basis of pre- and posttest scores, note that there can be no proof that the training itself will be responsible for any changes in scores that are observed. Likewise, the design in the second example will provide no assurances that the training in home visits will be solely responsible for any increase in the number of new users that might be observed. There may be many other causes, or confounding variables, that may produce the changes in both examples. Both of these designs fail to control for confounding variables and are, therefore, subject to various threats to internal validity.

Some of the common threats to internal validity, and the means by which to counteract them, are the following:

1. History. Events other than training that occur between measurements may exaggerate or diminish the results of the training. For example, in the second evaluation design example used, the CBD program manager might find that after training the CBD workers concerning home visits there was no subsequent increase in the number of home visits made by the workers. The program manager might conclude that the training was a failure. However, it might have been that an extraneous event--the arrival of the wet season for example--prevented the workers from making many home visits that month and the training was not a failure after all.

--Means of Counteracting. The threat of extraneous events confounding the results can be controlled for by the inclusion of a control population in the study design that is not given the experimental treatment but is subject to the same extraneous events as the experimental group. For example, another group of

CBD workers who did not receive training, but did experience the wet season, could serve as a control group to compare the number of home visits they make with that of the trained group. Another safeguard is to gather earlier data on the same outcome variable on the same population. For example, in the example used above, home visit records from the previous year(s) could be examined to see if the arrival of the wet season did cause a decrease in the number of home visits made.

2. Maturation. Over time people change and become older, as well as possibly wiser, hungrier, more tired, or more bored. In training events it is not uncommon for these last three changes to occur and influence the results on a posttest. If participants do become tired and bored with a training event it is likely that they would score lower than had they not been tired or bored. The lowered scores then, would be partially due to a maturation effect and not necessarily to a lack of learning.

--Means of Counteracting. Problems with maturation can be minimized if before and after measurements are kept relatively close together in time. In addition, a control group who experienced the same maturation effect, but not the intervention, could be used to compare the results of their before and after measurements with those of the experimental group.

3. Testing. Whenever a pretest is given to a group of participants the experience of taking the test may make the participants test-wise and able to score better on subsequent tests. This occurs because the participants become sensitized and more attentive to the topics covered in the pretest and may remember the questions and the errors they made when they take the posttest. The testing threat to validity is even greater when identical pre- and posttests are given.

--Means of Counteracting. This threat can be reduced by making pre- and posttests different. It is entirely eliminated if no tests or measurements at all are given before the final test or measurement. Alternatively, a group can be randomly divided into two with one group (group A) receiving a pretest and the other control group (group B) receiving no pretest. Afterwards the groups are combined and receive the intervention and posttest. Differences in posttest scores between the two groups would then be largely due to the one group being sensitized by taking the pretest (provided that the two groups contained large enough numbers and were randomly divided so as to assure statistical "equality"). Thus, the true measure of effectiveness of the training would be determined by subtracting group A's pretest average from group B's posttest average.

4. Instrumentation. This term refers to the possibility of introducing differences between a pretest and posttest that are caused by different testing instruments. Whenever a test or questionnaire or other measurement instrument is changed between two different measurements this threat to validity becomes a possibility. For example it could be that the posttest used is much more difficult than the pretest and could account for a poorer performance on the posttest.

--Means of Counteracting. Validity threats due to instrumentation can be eliminated by using identical testing instruments in before and after measurements. In situations where this is not desirable, for example in seeking to reduce the validity threat associated with repeat testing (producing test-wise participants), one can structure pre- and posttests to be similar but not identical. Equality of difficulty between the two tests might be achieved by previously selecting a certain number of questions, with at least two for each topic of training, and then randomly selecting half (including one for each topic) to be included in the pretest and half to be included in the posttest.

5. Statistical Regression. Anytime an individual scores extremely high or low on a test in comparison to the mean score of other participants it is probable that on the next test the individual will score closer to the mean. This phenomenon is known as regression. Statistical regression becomes a threat to validity only when individuals are chosen because they exhibit extreme characteristics such as very high test scores.

--Means of Counteracting. Since this threat applies only to situations where individuals are chosen because of some extreme characteristic that was evidenced in a one-time measurement, it can be lessened by measuring individuals more than once, or by not selecting individuals on the basis of extreme one-time measurements.

6. Selection Bias. This common threat to validity occurs whenever participants assigned to experimental and control groups differ in a significant way that could affect the outcome being measured. The differences between the two groups may be unknown to the evaluator but can exert their effect and distort the results just the same. Differences such as age, sex, and socioeconomic status, as well as differences in specific indicators related to an outcome measure being analyzed can all bias the results.

--Means of Counteracting. The best means of preventing this threat to validity is to randomly assign participants to experimental and control groups and thereby randomly distribute any major differences present in the sample. If random assignment is not possible, then matching pairs of experimental and comparison subjects on selected "important" characteristics can be done although some degree of selection bias will undoubtedly remain.

7. Mortality. Also known as attrition or dropout, this threat to validity exists whenever participants in a program or training event drop out before the final

testing has occured. Since people who drop out of a course are likely to be different from those who remain (with the dropouts probably being less motivated and less successful) there will be a threat to the validity of any study that measures differences between two groups that experience very unequal attrition rates. Any difference between final test scores between the two groups will be subject to the suspicion that it was not the intervention that caused the differences but rather a difference in dropout rates.

--Means of Counteracting. This threat is best controlled for by seeking to prevent dropouts from occurring to the extent possible, and when they occur, it is usually preferable to eliminate them, including their pretest scores, from the study results. Random assignment of subjects to experimental and control groups also allows one to avoid forming nonequivalent groups who may have characteristics that would predispose one or the other to a higher attrition rate.

External Validity

External validity refers to the ability of evaluation results to be generalizable to other people and settings with similar characteristics. Evaluation results that are published or communicated to an outside audience are usually done so for the purpose of sharing the lessons that were learned as a result of the evaluation so as to benefit other people and organizations who might be involved in the same type of intervention. Implicit in this communication is the belief that the results of one evaluation will have significance for others in different settings. However, there are several threats to external validity that can cause the results of one evaluation to be of little relevance to others because of certain situations or conditions present in the evaluation. Among these threats are the following:

1. Hawthorne Effect. This refers to the effects that experimental procedures may have on evaluation findings that would limit their generalizability. The name

came from an early experiment in industrial hygiene at the Hawthorne electrical plant which found that workers increased their productivity when the lighting was increased, and they increased their productivity when the lighting was lowered. The conclusion was that increased attention paid to the workers resulted in an increase in productivity. The possibility exists, therefore, of making evaluation results unrepresentative by paying too much attention to your subjects.

--Means of Counteracting. This threat can be minimized by utilizing unobtrusive measurements to the extent possible and by not notifying the subjects that they are being evaluated (unless professional ethics dictate otherwise). One example of how this might be applied could be where a course given to family planning clinic counselors is evaluated after the fact by using observers who pose as family planning clients. The pseudo-clients then enter the clinic and evaluate the interaction they have with the counselors who are unaware of the evaluation taking place.

2. Social Desirability Effect. This effect refers to the situation where subjects seek to please or impress the evaluator and, therefore, give answers or demonstrate behavior that they think is expected of them. Tests of attitudes commonly elicit this effect where the participant states an attitude that he or she thinks is the "correct" one sought by the instructor.

--Means of Counteracting. This threat can also be reduced by utilizing unobtrusive measurement and by not notifying the subjects that they are being evaluated. When subjects know they are being evaluated, efforts can be made by the evaluators to impress upon the subjects that there is no "right" answer or behavior and they should answer or behave as they normally would. Interviewers or evaluators should naturally not give any clues as to what they are looking for or hoping for in answers or behaviors.

3. Placebo Effect. This effect is commonly known in the medical sciences where patients who are given a treatment respond favorably solely because of their belief that any treatment given by a doctor will improve their condition. In the area of training evaluation, this could occur when participants expect training to help them and consequently report that it did (even if other objective measures indicated that it did not) in a followup questionnaire.

--Means of Counteracting. The placebo effect can be minimized by utilizing objective measures of an outcome rather than subjective, self-reporting measures. Likewise, a placebo effect could be averted by utilizing control groups and by "blinding" the subjects--that is, not informing the subjects as to whether they are in the "special" educational treatment or in a control group.

4. Experimenter Bias. When an experimenter or educator has a vested interest in a particular individual or group, or believes a certain experimental approach to be superior, it is quite possible that he or she will be biased in recording or interpreting the results of test scores from that individual or group or experimental approach. This is known as experimenter bias and can be due to either overt bias or subconscious bias where an experimenter hopes for or expects certain results and is, therefore, more likely to find them.

--Means of Counteracting. It is possible to limit the threat of experimenter bias by seeking to choose educators who do not have any particular vested interest or bias towards a particular group or experimental stragegy. Likewise, experiments or evaluations can be done so that they are "double-blind"--where the subjects as well as the people delivering the training or other intervention do not know which group is the experimental group and which is the control group. A "triple-blind" study--where the subjects, trainers, and the evaluators or data gatherers all are unaware of which groups are experimental and which are control--is the safest means of entirely eliminating experimenter bias.

However, both double-blind and triple-blind evaluations have very few applications in training evaluation given the nature of training and the difficulty in disguising the presence of the training intervention. When no control groups are utilized in a training evaluation, it is still possible to partially blind the study by asking another party, who does not know the intervention under review or the results that are expected, to code and interpret the results.

5. Novelty Effect. This refers to the effect on evaluation results that can come from the mere fact that subjects being evaluated are having their normal routine interrupted in some way (unless the evaluation is entirely unobtrusive) and, therefore, may act differently than they would under normal circumstances. For example, if a supervisor of CBD workers evaluates the performance of the workers by accompanying them for a day, it can be expected that this very obtrusive measurement will result in the CBD workers behaving differently than they would if they were alone or being evaluated on a routine basis by their supervisor. The novelty effect would also likely occur if a guest speaker delivered part of a training course and other courses did not have the same "novelty" of a guest speaker being present.

--Means of Counteracting. The novelty effect is counteracted by seeking to not disrupt the normal environment that subjects' experience to the extent possible. Avoiding special, one-time, guest speakers or evaluators who deliver an intervention or evaluate the subjects' performance will further the external validity of the evaluation design.

As can be seen from the examples above the threats to the validity of evaluation designs are numerous. In fact it is hard to escape the conclusion that all evaluation results are in some way tainted by one or more threats to internal or external validity. This is particularly true in the field of training evaluation where true experiments that control

for most of the threats to internal validity are rarely done. Nevertheless, meaningful results can still be achieved even when all threats to validity are not controlled for. As was said earlier, the training evaluator must balance the need for a scientifically pure evaluation, free from all validity threats, with the need to design cost-effective evaluations.

The reader may have noticed that some of the threats to internal and external validity seem to be in conflict with one another. That is, efforts to minimize one threat serve at the same time to increase another threat to validity. This is true for many of the validity threats mentioned above. For example, testing and instrumentation can at times be at cross purposes. In order to reduce the testing threat to validity, a trainer may want to give different pre- and posttests to his or her students so as not to alert them to the questions that will be on the posttest. In so doing, however, the trainer increases the validity threat of instrumentation because of the fact that two different tests will be employed.

An evaluator who controls for every threat to internal validity will probably be increasing the threats to external validity in the process, and vice versa. This is because evaluation designs that seek to control for threats to internal validity typically require a control group, before and after testing, and randomized allocation (if it is to control for selection bias) of subjects into experimental and control groups. Yet introducing repeated testing and randomized allocation into treatment and control groups tends to increase threats to external validity, such as the Hawthorne effect and novelty effect, because the participants tend to become very conscious of the fact that they are being evaluated and may not act as they would under normal circumstances. On the other hand, utilizing entirely unobtrusive measures that make the subjects unaware that they are being evaluated, and hence favor external validity, are hard to employ in experimental and quasi-experimental evaluation designs that usually require altering the natural setting and making their presence known to the subjects.

While efforts to reduce both internal and external threats to validity should be made, it is usually always preferable to favor internal validity over external validity if a choice between the two must be made. This is because it is much better to have evaluation results that you know truly reflect the effects of the intervention under study, even if they are not very representative of other situations, than it is to have results that are generalizable but paint an untrue picture.

CHARACTERISTICS OF TRAINING EVALUATION DESIGNS

As has been seen in the examples above, the threats to internal and external validity are numerous and counteracting these threats is the major determinant, along with the resources that one has available, in deciding what type of evaluation design to utilize. There are three principal levels of evaluation design that are used in the training evaluation field that will now be discussed. In discussing these designs it will be helpful to use a common notation developed by Campbell and Stanley[3] that utilize symbols to indicate various components of a design. These symbols are the following:

"R" indicates random assignment of people to either an experimental or control group

"----" drawn horizontally, separating the two groups, indicates that the groups were not randomly assigned and, therefore, nonequivalent

"O" indicates a measurement of some kind, an observation

"X" indicates a program intervention; for example, training

Nonexperimental Designs

The examples of evaluation designs mentioned at the beginning of this section were both nonexperimental designs. The distinguishing features of nonexperimental designs are

3. Campbell, Donald T., and Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research. Chicago:Rand McNally, 1966.

that they do not include control groups (formed by random assignment) nor comparison groups (formed by non-random assignment) and, therefore, assert little control over the major factors that confound (or cause misinterpretation of) an observed effect. Nevertheless, nonexperimental designs are the principal design used for evaluating most routine training events and appropriately so. It does not make sense for large expenditures to be made in order to evaluate a routine course by selecting participants, randomly assigning them into an experimental group (the course) or control group (perhaps another course that presents a different subject) and taking before and after measurements of both groups. This type of experimental design should be reserved for special circumstances that will be discussed later. The point to make here is that nonexperimental designs are appropriate to use in many situations as long as one is careful not to draw too many conclusions about the effects of the training in the absence of stricter controls.

The following paragraphs provide some examples of common nonexperimental evaluation designs, along with guidelines as to when they are appropriate to use and some of the problems associated with them.

Posttest Only Design

Experimental group X_1 O_1

The posttest only design is a very common evaluation design but unfortunately it is the least valid design. It consists of a single group evaluated only once after an intervention has occurred. There are no data collected on the group prior to the intervention and no control or comparison groups to compare the treatment group's performance to.

Problems with the design: Because there are no pretests included in this evaluation design, it is impossible to definitively state that training had a positive effect and to what degree.

Appropriate uses in training evaluation: The appropriate uses of this design are when it is not important to document the effect that training had or when there is very little time allowed for a training intervention and giving a pretest may consume too many precious minutes. It is also appropriate when you may be able to assume a total absence of knowledge for a given subject and could assume a pretest score close to zero.

One-Group, Pretest-Posttest Design

Experimental group O_1 X_1 O_2

This design incorporates a pretest which allows a comparison to be made with the posttest and consequently a reasonable estimate of the degree of learning that took place can be arrived at. This design is probably the most widely used in the field of training evaluation because it is simple, much more valid than the posttest only design for drawing conclusions as to the effect of training, and does not require a great deal of resources to undertake.

Problems with the design. While much more valid than the posttest only design, this design is subject to several threats to internal validity including history, maturation, testing, and instrumentation. Evaluators using this design will have to qualify the statements they make about learning increases in light of these validity threats.

Appropriate uses in training evaluation: This design has many appropriate uses in training evaluation. In spite of the fact that it does not control for the effects of history, maturation, testing, and instrumentation, these effects do not always confound results and various means can be employed to keep their potential confounding effect to a minimum. These means were discussed in the previous section following each threat to validity. The most appropriate use of this design is probably for the typical training course where the development of knowledge, attitudes, or skills is the outcome measure

of interest. For other outcome measures such as the effect of training on job performance and its short-term effects on the target population, this is a less desirable evaluation design. For evaluating outcomes other than development of KAS, it is usually desirable to choose a quasi-experimental or experimental design.

Static-Group Comparison Design

Experimental group	X_1	O_1
Comparison group		O_2

This design includes a comparison group formed by non-random means. Observations are made of both the experimental and comparison groups after the experimental group has received an intervention. This design is an improvement over the posttest only design in that it allows for a comparison to be made with a group that did not receive the intervention. In the training field it is sometimes performed when the same posttest that was given to a recently trained group of individuals is given to similar individuals who did not receive training. It may be tempting to conclude that differences between the scores of the two groups reflect the effect of training, but there are validity threats with this design that prevent this conclusion from being assumed.

Problems with the design: The validity threats to this design include selection bias (by virtue of the fact that the groups were not randomly assigned) and mortality (subjects in the experimental or comparison group may have left the groups for some reason).

Appropriate uses in training evaluation: This design is appropriate where it is difficult to obtain more than one measurement on a group, or where the threat to validity of repeated testing is high. Likewise it is appropriate for routine courses where a posttest only design might otherwise be used. Comparing the posttests from the trainees with another group of like individuals who were not trained can sometimes provide meaningful information even if it cannot be used to prove the effects of training.

Quasi-experimental Designs

Quasi-experimental designs are most often used in the training evaluation field when there is concern for establishing strong evidence for the effects of training. As mentioned above, this often occurs in the context of evaluating the effects of training on job performance as well as the short-term effects on the target population. Quasi-experimental designs are favored over non-experimental designs for these types of evaluations because the former are more powerful designs that exert greater control over confounding variables.

The distinguishing characteristics of quasi-experimental designs are their establishment of experimental and comparison groups formed by methods other than random assignment. In addition, before and after measurements are made of both groups. Quasi-experimental designs exercise varying degrees of control over many, but not all, of the factors that affect the internal validity of results. They are often accepted as the best design that can be utilized in the training evaluation field given the frequent impossibility of randomly assigned participants to either experimental or control groups.

Single Group, Time Series Design

Experimental group	O_1	O_2	O_3	X_1	O_4	O_5	O_6

The single group, time series design is commonly considered a quasi-experimental design even though it does not include a comparison group. The design includes a series of measurements (more than one) both before and after the intervention that help to control for the effects of history and maturation. Naturally the more observations that are made, the more valid will be will be the results. This design is particularly useful when performing followup evaluations of the effects of training, such as participants' job

performance or the short-term effects on the target population. One of the best uses of this design is to identify serial trends in data. Repeatedly measuring outcome criteria such as the number of home visits a distributor makes can identify trends that might not appear in only two measurements, one taken before and the other after training. It could be, for example, that the number of visits the distributors were making before training was increasing month by month and the subsequent increase in visits following training was no greater than the increase experienced in earlier months.

Problems with the Design: Unfortunately this design is subject to the validity threats of testing, instrumentation (if the same instrument is not used), and mortality, as well as history and maturation to a certain extent. The latter two can be greatly reduced, and perhaps eliminated entirely, by adding a comparison group to the design.

Appropriate uses in training evaluation: As alluded to above, the uses of this design in training evaluation are many. It is a powerful tool that can be easily employed provided that the repeated measurements required in the design can be practically carried out in a consistent manner, and by utilizing the same instrument so as to eliminate the validity threat of instrumentation. It is particularly appropriate when the measure that one is interested in is already being collected. It is not practical if participants, or the subjects of the evaluation, are far removed from the evaluators and would require large expenditures in order to be reached.

Pretest-Posttest Comparison Group

Experimental group	O_1	X_1	O_2
Comparison group	O_3		O_4

This is a design in widespread use in the fields of education and training because it fits the realities that trainers often face. Usually trainers have to work with already formed groups or classes and cannot assign participants to these groups, or other control groups,

randomly (as is done in the experimental designs). Nevertheless, trainers can observe a comparison group that is similar to the experimental group for key variables. The design involves selecting a comparison group and then measuring the variable under consideration for both the experimental and comparison groups before any intervention such as training takes place. Following the experimental intervention, the two groups are measured again and differences between the two groups can be said, with proper qualifications, to be due to the training. The strength of this design is in its ability to largely control for the effects of history and maturation, and partially control for the effects of testing and instrumentation.

This design can be strengthened by incorporating it with the previous design to make a Two Group, Time Series Design where repeated measurements are taken on both the experimental and comparison groups before and after the intervention. This makes for one of the most valid quasi-experimental designs.

Problems with the design: The flaw with this design that prevents the results from being able to be used to show conclusively the effects of training is its lack of random allocation of participants to the experimental or comparison groups. This leaves open the possibility that there was selection bias in forming the experimental and comparison groups and, therefore, they are non-equivalent. This will always be the major threat to the validity of quasi-experimental evaluation designs and one that is often unavoidable. The most that the evaluator can do to prevent the effect of selection bias is to use great care in matching comparison groups to experimental groups so that they are equivalent on key variables such as age, sex, occupation, or whatever variables are believed to influence the outcome. Besides the effect of selection bias, the effect of instrumentation can occur with this design if different instruments are utilized on the pre- and posttests. Likewise, the effect of testing can confound the results of both experimental and comparison groups.

Appropriate uses in training evaluation: The appropriate uses of this evaluation design are many. It is a useful design whenever the results of a training evaluation are important enough to warrant the increased validity that comes with a quasi-experimental design versus a non-experimental design. This may be the case when a large amount of resources are being devoted to a particular type of training, or when the training is new and being evaluated for the first time. Likewise it is very useful when target populations are the subject of the evaluation (such as when measuring the short-term effects of training on a target population) rather than the participants themselves. Since target populations are much more exposed to outside effects, it is important to support the validity of these kind of evaluations by including a comparison group in the analysis.

Experimental Designs

Experimental designs are usually an expensive luxury in the field of training evaluation but they can sometimes be appropriate, particularly if large resources are already going into a particular type of training and it must be determined what the true effect and magnitude of the training is. An experimental design includes random assignment of subjects into control and experimental groups, and observations of both groups, usually before and after application of an intervention. Experimental designs yield the most interpretable, definitive, and defensible evidence of effectiveness. An experimental design should be utilized when there is reason to be concerned with the possibility of any validity threats affecting the outcomes, and when a feasible means for randomly assigning participants to experimental and control groups exist, and when the resources are present that permit it.

Pretest-Posttest Control Group Design

Experimental group	R	O_1	X_1	O_2
Control group	R	O_3		O_4

In this classical true experiment, a sample of participants are randomly divided into

experimental and control groups and then an intervention, such as training, is applied to the experimental group only. The pretest scores can be used to check that the groups started out as more or less equivalent. If at the end of the training, the mean posttest score for the experimental group is significantly higher than the mean posttest score for the control group, this difference can be attributed to the effect of the training. This is a very powerful design which eliminates all threats to internal validity except for the effect of testing (and instrumentation if different instruments are used). The threats of external validity can still pose a problem, however. This design is useful when a large group of subjects are available and they can be randomly assigned (without their knowledge if possible, so as to increase external validity) to experimental or control groups.

Problems with the design: In addition to the potential validity threats of testing and the various threats to external validity, a major obstacle to performing this design is the frequent impracticality of randomly assigning subjects into experimental and control groups. There are few people who will consent to be in a control group and tested twice without receiving any discernible benefit from it. While some evaluators get around this obstacle by "blinding" the subjects and guaranteeing them only a 50 percent chance of falling into the experimental group (and thereby receive some perceived benefit) it is difficult to blind subjects as to whether they are receiving training or not. Unlike medicine where doctors can either inject a certain drug and observe its reaction or inject distilled water, trainers cannot very easily disguise the presence or absence of training. Perhaps all that can be done to blind subjects is to provide a type of training to both groups but differ them along the lines of the key independent variables.

Besides the difficulty of finding subjects willing to be part of a control group, trainers are often confronted with already formed groups, such as a class of high school students, who require training and it is impractical to split them up into experimental and control groups. Matching them with a similar class in the same or another high school (and thus forming a comparison group) is usually the most feasible course to pursue.

One caveat that should be mentioned here is that random assignment does not automatically assure equivalence between two groups. The size of the samples must be large enough so that any differences between the individuals to be randomly distributed will not fall predominantly in either of the two groups. Naturally the more dissimilar the individuals are, the greater the sample will have to be in order to create two "equivalent" groups. Sample size will be further discussed later in this section.

Appropriate uses in training evaluation: This design, like other experimental designs, is appropriate whenever the concern with threats to validity (especially possible selection bias) warrant the extra time, money, and effort required to implement it. Provided that the sample sizes are large enough to allow the random assignment to achieve "equality" between the two groups, the evaluator will be able to confidently assert that training was responsible for the differences observed between the two groups.

Posttest Only Control Group Design

Experimental group	R	X_1	O_1
Control group	R		O_2

This is also a true experimental design. It is exactly like the previous design except it contains no pretest of either the experimental or control group. This is an easier design to implement than the previous one because of the absence of a pretest. Although pretests have many important uses, they are not required in true experimental designs (as they are in quasi-experimental designs) in order to validate the fact that the groups are equivalent at the beginning of the experiment. If the sample size for each group is large enough (at least 20), one can have confidence that random assignment will result in two equivalent groups.

Problems with the design: The principal problem associated with this design in comparison to the previous one is that the absence of pretests will prevent one from

determining the amount of change brought about by the intervention. This design will show that there is a difference between the two groups at the posttest, but the extent of the change attributed to the program cannot be determined precisely without the use of a pretest.

Appropriate uses in training evaluation: The principal advantage this design has over the previous one is that it eliminates the validity threat of testing. It should be used, therefore, whenever an experimental design is called for and there is reason to believe that a pretest might interfere with the training results, or when it is difficult to give a pretest.

The examples mentioned above are just a few of the myriad designs available to the training evaluator, but they include the most commonly used designs. There are no clear-cut rules for when to use one design versus another. These decisions should be made on the basis of careful coordination between trainers, evaluators, supervisors and anyone else in the organization who might have a role to play in training evaluation. The basic rule of thumb in deciding which design to choose is to use the most powerful design (that controls for the threats to validity relevant to the situation at hand) that can be practically and affordably carried out. For most routine training evaluation situations this will be a nonexperimental design. For some larger, more complex evaluations, including operations research studies and cost-effectiveness analyses, a quasi-experimental design may need to be used. Finally, for special situations where resources permit and where a large group of willing subjects is at hand, an experimental design may be appropriate.

SAMPLING

At the beginning of this section it was noted that one of the key components of an evaluation design consisted of specifying from whom the evaluation results would be collected. One obvious answer to this question would be "from all the people trained."

But what if hundreds of people were trained in large one-day seminars? Perhaps not all of the trainees need to be evaluated to get a sense of what the reaction to or effect of the seminars was. Or suppose the effects of training CBD workers in contraceptive method counseling is evaluated on the basis of interviews with the CBD workers' clients. Do all the clients have to be interviewed? In both these cases it may be more appropriate to choose a sample of all the participants who attended the seminars or of all the clients served in the program. The alternative of measuring the entire population (who in these cases would be all the participants who attended the seminars and all the clients served in the program), would probably be too costly and impractical. How a sample can and should be selected and how large the sample size should be form the remainder of this section.

Sample Selection

A sample should be a miniature version of the population to which the evaluation findings are going to be applied. In order to be useful, samples must be representative of the population they were chosen from. If the samples closely approximate the population they were chosen from on characteristics that are relevant to the interests of the evaluation, then it will be representative of those populations. The best way for assuring representativeness is by using a probability sample and by using a large enough sample size.

Probability Sample

A probability sample is a sample in which all the members of a population have a known probability of being selected. Probability samples are more likely than non-probability samples to be truly representative of the larger population and are, therefore, usually preferable to non-probability samples. Methods involved in probability sampling also reduce the level of conscious and unconscious sampling bias. The following are common methods for selecting probability samples.

1. Simple Random Sampling: In simple random sampling participants are selected on a random basis, without any underlying system or pattern. Each person in the population being studied has an equal chance of being included in the sample. The most common process for selecting the sample is by assigning each member of the population a number and then choosing numbers from a random number table until the desired sample size has been reached.

2. Systematic Sampling: Systematic sampling is easier and less time consuming than simple random sampling. It consists of selecting elements for the sample in regular intervals. For example, the population can be listed in an arbitrary order (e.g. by alphabetical order) and then every 5th or 25th or nth number may be picked from the list depending upon the size of the population and the size of the sample being selected. Of course, if the population is already listed in a particular order, it is not necessary to rearrange them. Choosing elements at regular intervals from any list of the population will result in a random sample provided that the list was not arranged in a pattern that coincides with the sampling interval.

3. Stratified Sampling: In simple random sampling or systematic sampling it is possible to randomly select a sample that has differing percentages of certain characteristics than the population. For example, it is possible to select a sample of women, 20 percent of whom use an IUD, even though the sample was randomly chosen from a population of whom 10 percent use an IUD. If it is important that the sample closely mirror the population on certain key variables such as type of contraceptive method used, then stratified sampling should be used. This method involves first dividing the population into categories or strata of interest to the evaluator. For example, one such strata could be the type of contraceptive method used. Next random samples are chosen from each of the strata in order to arrive at the desired make up. If the population contained

family planning users of whom 60 percent used the pill, 30 percent used barrier methods, and 10 percent used an IUD, the population would be divided into these three categories and then IUD users would be randomly selected until a sufficient number were chosen to make up 10 percent of the sample size. Overall, stratified sampling provides a basis for a sample to be more representative of the true population. It can also assure that a sufficient number of population subgroups, whose behavior is of special importance to the evaluation, is present in the sample.

4. Cluster Sampling: This method is very useful to simplify the sampling process, especially when the population is very large and geographically dispersed. It involves first dividing the population into smaller groups or clusters. Often this can be done by identifying natural divisions or clusters within the population. For instance, if a trainer wishes to do a followup interview of a sample of 30 of the 100 CBD workers given training during the past year and at the same time conserve resources, he or she may not want to randomly select the 30 since these individuals may live in completely different areas of the country. Instead a cluster sample could be performed where the trainer first identifies nine provinces from which all 100 of the CBD workers live in roughly equal proportions. The trainer could then randomly choose three provinces and then randomly select 10 CBD workers for followup from each of these provinces. Naturally, cluster sampling in this example would save a great deal of time, money, and effort for the trainer compared to any of the other probability samples.

Non-probability Sample

A non-probability sample refers to the selection of a sample that is not based on known probabilities. Since it does not include random selection of participants it is open to the possibility of selection bias, whether conscious or unconscious. Inevitably, non-

probability samples underselect or overselect people with certain characteristics that are not representative of the population under review. Since non-probability samples cannot be counted on to yield samples representative of the population, the information gained from the study cannot be assumed to apply to any group other than the sample itself.

Despite this limitation of generalizability, non-probability samples have many uses. There are many times when samples are needed but obtaining a probability sample would be impossible or prohibitively costly. If, for example, an evaluator wanted to survey a periurban community to find out the attitudes of men towards vasectomy, it would be next to impossible to assign a number to each man in the community and then randomly select them. Even a cluster sample would be very difficult to perform, particularly if the community was not neatly divided into equal sized portions. Even if there were square blocks that divided the community into roughly equal-sized neighborhoods, there may be no way of knowing how many men reside on each block and in the entire community. In such cases, non-probability samples are usually the only means by which to obtain information on the characteristics of a population.

There are three principal types of non-probability samples. These are purposive samples, quota samples, and samples of convenience.

1. Purposive samples: Purposive samples include situations where a sample is sought of either extreme cases, rare elements, or key informants. Extreme cases include people who fall into some extreme category, for example, women who have over 10 children. If a study were interested in the attitudes of women who have over 10 children towards family planning, these women would need to be specially sought out, as a random sample of women in general might never yield enough women in this category. Purposive sampling could be done, therefore, to purposefully seek women out (in a non-random fashion) who had over 10 children and then obtain the information the study is collecting.

Purposive samples based on rare elements include situations where a sample is sought that contains individuals with a rare characteristic. Collecting information on AIDS patients, for example, would be a situation where purposive samples would probably be called for. As in the previous example, the AIDS patients would be sought out in a non-random fashion, perhaps by visiting a local AIDS support group.

Purposive samples based on key informants refers to situations where information is sought from a few strategically chosen individuals, such as community leaders, who might provide enlightening and articulate information that other members of the community perhaps could not provide.

2. Quota samples: While non-probability samples inevitably introduce some selection bias, utilizing a quota sample will help to reduce, although not eliminate altogether, such bias. A quota sample is similar to a stratified sample except that the sample is not chosen randomly from the various strata isolated. In quota sampling the characteristics thought to be relevant to the study are first determined and then the percentage of the population with those characteristics is determined. The quota sample will then be selected so as to have those same percentages of the various characteristics.

3. Sample of convenience: In this sample type, individuals are chosen for inclusion in the sample solely on the basis of their availability or proximity to the evaluator. For example, if an evaluator steps out into the street to interview women who pass by concerning their attitudes toward family size, this is a sample of convenience. There is absolutely no assurance that the sample is representative at all of women in general. This sample type is very convenient and inexpensive to perform but is also the least valid. Appropriate qualifications should always be made when reporting the results.

Sample Size Determination

In addition to how a sample is selected, the size of the sample is of utmost importance and can determine whether or not the sample is representative of the population. If a sample is too small its results will be meaningless when applied to the population it was taken from and if the sample is too large it could waste valuable time and resources. Many non-technical books state a rule of thumb that samples should consist of at least 30-35 people in an experimental group, and an equal number in a control group if one is used[4,5]. However, these guidelines can under- or overstate the sample size required for a given population with its unique determinants.

Their are five basic determinants that go into selecting the appropriate sample size: the size of the population, the variation of the data being observed, the precision or degree of accuracy sought, the reliability or confidence level sought, and finally, the resources available.

Size of the Population

A large population naturally requires a larger sample to represent it than a small population. However, as populations increase in size they require samples that are smaller as percentages of the population. This can be seen in Table 2 at the end of this section, where the sample size needed for a population of 100 is 80 but the sample size needed for a population of 10 is 10 (i.e. the entire population).

4. Brindis, D., et al, Evaluation Guidebook for Family Planning Information and Education Projects, San Francisco: Center for Population and Reproductive Health Policy, U.C.S.F., page 101.
5. Mayo, G.D.; DuBois, P.H., The Complete Book of Training, San Diego: University Associates, page 175.

Data Variability

The sample size also depends on the degree of variability of the data being observed. This variability is known as the standard deviation and will be described in section V. If a population is very heterogeneous and, therefore, has much variability in its observed characteristics, it will take a larger sample to be representative than if the population were very homogeneous. Imagine a group 100 seemingly identical nurses who are trained in family planning and a group of 100 people randomly chosen from the street who are given the training as well. If the trainer wanted to do followup interviews with the participants after the course, he or she would have to pick a larger sample of the general population participants, relative to the nurses, in order to gain a representative sample and make conclusions as to how all of the participants would have responded had they been followed up.

Precision

The degree of precision sought by evaluators also influences the size of the sample. For example, an evaluator who seeks to determine the effect of training CBD workers in family planning service delivery might do a mini-survey of contraceptive prevalence in the target community before and after training. If the evaluator wants to be able to estimate the prevalence rate in the target community to the nearest percentage point (for example, to know that the prevalence rate is 45% and not 44% or 46%) a larger sample will be required than if the evaluator only wants to be able to estimate the prevalence rate to the nearest 10th percent (for example, to know that the rate is 50% and not 40% or 60%).

Confidence Level

Whenever samples are used to predict the value that would be found if the entire population were surveyed, there is the possibility of error. Despite the utilization of a

large sample, randomly selected, it is still possible that the sample will produce erroneous estimates for the characteristics of the population being analyzed. The confidence level is the degree of confidence that one has in the estimate made on the basis of a sample. In the example above, the evaluator may want to be 95 percent confident that the prevalence rate in the target community is 43 percent after training (utilizing the .01 level of precision) on the basis of the sample surveyed. If the evaluator wants to be 99 percent confident that this estimate of the prevalence rate is the true rate then the sample size would have to be increased accordingly.

Resources Available

Last but not least, the resources that one has available will help determine the sample size that should be used. If resources permit, training evaluators will want to select a sample size equal to or greater than the minimum size necessary for statistical significance to be guaranteed. However, if that statistically minimal sample size is greater than what can be feasibly or affordably carried out, training evaluators may need to compromise by using a smaller sample size and accepting results that may still be useful, even if they cannot be used to establish statistically valid conclusions.

Given the four determinants of population size, standard deviation, precision, and confidence level, a statistician can mathematically determine what the size of the sample should be by utilizing various formulas. It is beyond the scope of this manual to go into those formulas here, however. Persons responsible for training evaluation should seek advice from statistical reference books or statisticians before embarking on a major evaluation where determining a statistically valid sample size is important. When seeking the advice of statisticians, however, trainers should be prepared to state the population size, the estimate of the standard deviation, the level of precision sought, and the confidence level sought. After the required sample size is determined mathematically it will be up to the trainer or manager to decide whether the resources are present to obtain the required sample size or if a smaller sample size will have to suffice.

Table 2 presents estimates of the required sample sizes for given population sizes where N is the size of the population and S is the size of the sample. Naturally using this table will not be as accurate as determining the sample size mathematically (and may overstate the sample size needed) given one's own unique determinants such as the standard deviation and level of precision required, but it is a short-cut method that may be useful.

TABLE 2

TABLE FOR DETERMINING SAMPLE SIZE[6]

N	S	N	S	N	S
10	10	220	140	1200	291
15	14	230	144	1300	297
20	19	240	148	1400	302
25	24	250	152	1500	306
30	28	260	155	1600	310
35	32	270	159	1700	313
40	36	280	162	1800	317
45	40	290	165	1900	320
50	44	300	169	2000	322
55	48	320	175	2200	327
60	52	340	181	2400	331
65	56	360	186	2600	335
70	59	380	191	2800	338
75	63	400	196	3000	341
80	66	420	201	3500	346
85	70	440	205	4000	351
90	73	460	210	4500	354
95	76	480	214	5000	357
100	80	500	217	6000	361
110	86	550	226	7000	364
120	92	600	234	8000	367
130	97	650	242	9000	368
140	103	700	248	10000	370
150	108	750	254	15000	375
160	113	800	260	20000	377
170	118	850	265	30000	379
180	123	900	269	40000	380
190	127	950	274	50000	381
200	132	1000	278	75000	382
210	136	1100	285	1000000	384

6. R.V. Krejcie and D. Morgan, "Determining Sample Size for Research Activities", Educational and Psychological Measurement, 30:607-610, 1970.

CHAPTER III: DEVELOPMENT OF TRAINING EVALUATION INSTRUMENTS

After completing the training evaluation design and specifying what, when, from whom, how, and by whom the training will be evaluated, the next task of the training evaluator is to develop the instruments that will be used to obtain the information sought in the evaluation. Evaluation instruments include the tests, questionnaires, interview guides, and all the tools that evaluators use to obtain information to make the judgments called for in an evaluation. Taking care in designing the evaluation instruments is every bit as important as taking care in the overall evaluation design. A poorly constructed instrument can lead to invalid results as surely as a poorly constructed design. This section will outline the issues of validity and reliability as they relate to the subject of developing evaluation instruments, and present guidelines for the development of evaluation instruments.

VALIDITY OF TRAINING EVALUATION INSTRUMENTS

Validity as it relates to instruments is the degree to which the instrument measures what the evaluator wants it to measure. A thermometer, for example, is a valid instrument for measuring temperature. Asking a person how hot they feel the weather is, on the other hand, is a less valid measure of temperature. In the field of training evaluation, a valid instrument for measuring participants' learning might be a multiple-choice posttest. Administering a questionnaire on how the participants enjoyed the course would not, however, be a valid measure of their learning.

There are four basic approaches to determining if an instrument is valid. These approaches are: content validity, concurrent validity, construct validity, and predictive validity.

Content Validity

Content validity refers to the extent to which an instrument accurately represents the knowledge or skills it is intended to measure. A test given to a group of trainees will have content validity if its questions measure knowledge of a representative sample of all the information presented in the training. To ensure content validity, no important items, behaviors, or information presented in the training should be omitted from the test. In addition, there should be a balance of the material such that topics covered more thoroughly in the training should have more detailed questions concerning them in the test. An instructor who spent 5 minutes in a two hour course covering the male reproductive system and then devotes one-third of the test to questions about this topic will not have a test with content validity. Likewise, if the test includes questions that were never discussed in the course it will lack content validity.

Concurrent Validity

Concurrent validity refers to the extent to which an instrument agrees with the results of another instrument that you know to be a truly valid and accepted measure. For example, a home pregnancy kit would have concurrent validity if its results matched those of another test, such as a sonagram, whose results never failed. In training evaluation, a newly developed test would have concurrent validity if its results concurred with those of another more extensive (and perhaps more costly) test developed by a panel of experts whose validity was unquestioned.

Construct Validity

Construct validity refers to the extent to which an instrument represents the construct it purports to measure. A construct is an abstract variable such as a skill, attitude, or ability that the instrument is supposed to measure. When an evaluator questions whether a particular test or questionnaire really measures intelligence, achievement,

positive attitudes or some other construct, he or she is asking about construct validity. One way to test the construct validity of an instrument is to test it in the field to see if the results match the theoretical expectations. In order to measure trainees' job satisfaction, for example, a questionnaire might be administered that is supposed to measure the trainees' degree of job satisfaction. In order to test the construct validity of the questionnaire, the evaluator might seek out a number of people who are known to be highly satisfied with their jobs. If they score high on the questionnaire (indicative of high job satisfaction) then the questionnaire could be considered to have construct validity. Likewise, if the questionnaire were given to individuals who were known to have low job satisfaction, they should score low on the questionnaire if it possesses construct validity.

Predictive Validity

Predictive validity is the ability of a test or other instrument to predict the subsequent performance or behavior of the individual taking the test on the basis of the test results. If a posttest score for family planning supervisors has a high predictive validity with respect to being a good supervisor, then those who score highly should subsequently show themselves to be good supervisors. Unfortunately, knowledge test scores, while easy to administer, frequently fail to have a high predictive validity with respect to subsequent job performance. Other instruments with a higher predictive validity in this regard (such as performance tests or on-the-job observations) will normally need to be used.

RELIABILITY OF TRAINING EVALUATION INSTRUMENTS

The reliability of an instrument is its ability to consistently achieve the same or similar results from a given individual over time, assuming no change in the individual. If an attitude survey given to a person one week reveals to the evaluator that the person has attitudes that would make him or her a good counselor, and then one week later the

same person takes the same survey but this time the results lead the evaluator to conclude that this person would make a very poor counselor, then the survey has very little reliability. This unreliability, or inability to achieve consistent results, could be due to internal measurement error, to external or coder error, or to random error. These three sources of unreliability will each be discussed briefly.

Internal Measurement Error

Internal measurement error refers to the extent to which an instrument is intrinsically "error-prone" by its very design. Certain instruments will always tend to produce more unreliable results than others on the basis of their design. This will of course affect their validity (whether they are really measuring what they are supposed to be measuring) as well as their reliability. The degree of internal measurement error can be assessed by a number of methods. Three common procedures, which help ensure than the instrument itself is reliable, are: test-retest, multiple-form, and split-half.

Test-Retest

Measuring reliability by using the same test or other instrument at two points in time is called test-retest reliability. If there is a high degree of correlation (the strength of the relationship) between the results of the test given at two different times, then the test is relatively reliable. Of course the longer the time interval between the two applications of the test, the more likely it is that some other effect of maturation or history will cause the differences in the scores.

Multiple-Form

This approach to assessing reliability involves constructing two tests, or two forms of the same test, that theoretically measure the same thing, and then having the respondents answer both tests at the same session, mixing the order in which the questions are given

from the two tests. If there is a high positive correlation between the results from the two tests, then they are considered reliable. Another approach that trainers can utilize for determining the reliability of individual questions on a test, interview, or questionnaire is to repeat the question, perhaps worded slightly different, at a different location within the test. If the question is reliable it will elicit the same response.

Split-Half

This method involves splitting the instrument into two equal parts and comparing the results. To do this, one randomly assigns items in the instrument to two groups (for example dividing all the even and odd numbered questions), then sums the scores of the two groups, and measures the degree of correlation between the two sets of scores. A high correlation between the two sets of scores indicates a reliable instrument.

External Error

The three methods described above are means by which to measure whether an instrument is reliable or not. However, the intrinsic nature of an instrument is not the only source of unreliability when measuring peoples' responses. Unreliability can also arise due to errors in the instrument's score that are caused by the observation, rating, or coding processes. These sources of unreliability are referred to as external error, also known as coder error, or interrater reliability.

Interrater reliability becomes a significant issue whenever there is a great deal of data to process in a particular evaluation and there are multiple scorers or coders assigning scores to the completed instruments. Imprecise instructions given to the coders, different levels of attention and preciseness on the part of the coders, and simple mistakes in tabulating a score all contribute to external error and reduce the reliability of the instrument. Of course, it does not take a large team of coders to make these kind of mistakes. One solitary trainer scoring the pre- and posttests is also perfectly

capable of making mistakes in the process that will increase the unreliability of the results.

Random Error

A final source of unreliability of an instrument is random error. Random error is applicable to both reliability and validity, and consists of an error in measurement brought on by the environment, the condition of the participant, and how the instrument is administered.

The environment can play an often subtle, yet powerful role in affecting the results of a questionnaire or other instrument. Environmental determinants are many, including the weather, the location and context of the measurement, and the presence of others at the time of the measurement. For example, it should be easy to imagine the different results one might get when interviewing a woman about her contraceptive practices in her home, in the presence of her husband and mother-in-law, compared to interviewing her in private. Although the set of interview questions might be considered valid and internally reliable, the condition of the environment can cause answers that are neither valid nor reliable.

Likewise, the condition of the participant can cause random error to be produced that reduce the instrument's validity and reliability. Conditions such as hunger, fatigue, restlessness, and anger can all cause results to be invalid and unreliable.

Finally, the administration of the instrument can determine whether its results are both valid and reliable. For example, the instructions for how to fill out a questionnaire may be different from one group to another, causing the groups to perform differently on the questionnaire. Standard instructions are necessary when giving any instrument. Directions should be made clear and they should be repeated in the same way for different groups in order to assure reliability.

Any or all of these sources of error, whether internal, external, or random, can result in an instrument producing unreliable results. Care should be taken to eliminate all of the above sources of error by first testing the instruments to be used with one or more of the approaches for measuring internal error. Next, care should be taken when scoring and tabulating results to eliminate sources of external error. Finally, every effort should be made to reduce the possibility of random error by following certain protocol (e.g. always giving the same instructions to participants) and by attempting to control for validity threats such as maturation and history by the means discussed previously in the context of design validity.

Relationship Between Validity and Reliability

In this discussion on reliability, the topic of validity recurred, particularly in the context of sources of random error. This is because there is strong link between validity and reliability that is difficult to separate. As was noted above, random error, or changes in measurements produced from temporary conditions in the person or situation, affects both validity and reliability. A test given to a temporarily impatient participant will produce results that are both misleading (invalid) and inconsistent (unreliable). In this case the instrument lacked both validity and reliability (at least in the context in which it was given). Yet it is not always the case that validity and reliability are in agreement. It is possible for an evaluation instrument to be reliable but not valid. A ruler that had the lines on it spaced unevenly will always measure the distance between two objects in the same way, thus demonstrating reliability, even though the resulting measurement is invalid. A training evaluation instrument can also show reliability yet not measure the characteristic accurately that it was designed to measure.

On the other hand it is impossible to have an unreliable instrument that is at the same time valid. If the instrument is unreliable (either due to its design, the way it was coded, or the context it was given in), then it cannot be an accurate measure of a given

characteristic, at least at that particular moment in time. If the instrument is unreliable due to internal error then it can never be a valid instrument. If the instrument is unreliable due to external or random error then the instrument will be invalid for as long as the conditions persist that caused the external or random error.

This distinction between the validity and reliability of instruments is important to make because it is relatively easy to check for reliability. A trainer or evaluator simply has to check for consistency of results and perform tests such as the test-retest method for assessing reliability. If there is a lack of reliability then there is necessarily a lack of validity for one or more of the measurements made. Yet the reverse does not hold true. The presence of a reliable instrument does not guarantee its validity. Therefore, the more difficult checks for validity should still be done, even when an instrument appears to be reliable.

To briefly illustrate this relationship, suppose an evaluator hires two people to observe a particular family planning promoter on the job in order to evaluate the effect of communication training on the promoter's job performance. The two observers might report back very different results with one concluding that the promoter is utilizing the communication skills taught in training and the other observer reporting that these skills were not being utilized at all. The disagreement between the two observers demonstrates a lack of reliability with the instrument (in this case apparently due to external error or interrater unreliability). Therefore the evaluation is both unreliable and invalid, since there was inconsistency of results and doubt as to the correctness of the conclusions. However, if the two observers produced the same conclusions, the evaluator could only be somewhat confident of the reliability of the instrument but not its validity. The observations could still be very inappropriate and misleading (and therefore invalid) measures of whether communication skills were being applied on the job, in spite of their consistency of findings.

DEVELOPING EVALUATION INSTRUMENTS

Before entering into a discussion of how to design specific evaluation instruments, there are several rules that should be followed when developing any evaluation instrument. These are:

- design instruments to yield a wide distribution of responses;
- design instruments that are easy to administer;
- design instruments that are simple and succinct;
- design instruments that are precise and unbiased;
- design instruments that will be economical to employ;
- field test instruments before using them; and
- do no unduly burden agency staff with data collection.

Distribution of Responses

In general an instrument should be designed so as to yield a wide distribution of responses. Questions used in interviews, questionnaires, and tests should not be asked in such a way that everyone responds the same. Rather the instruments should discriminate between the respondents on the items of concern to the evaluator. A question such as "Do you believe it is important to be a responsible parent?" will elicit only one kind of response. A more discriminating question is "Do you believe that practicing contraception is a sign of responsible parenthood?". Field testing is a useful way to test the ability of an evaluation instrument to provide a variety of responses that will be of use to the evaluator.

When constructing a test this concept of distribution of responses is particularly important and is a major concern of psychometricians who specialize in testing design and analysis of results. Psychometricians usually seek to construct a test item that will yield a mean score of about 50 percent correct answers so that there will be a distribution resembling a bell-shaped curve. Educators, on the other hand, usually construct test items that are easier, often wishing to build confidence and reward the student by having the majority of them be able to answer the question correctly. Their

test scores are often clumped in the 70-90 percent correct range and few students if any score below 50 percent. This can sometimes create a conflict between educators and evaluators who prefer normal bell-shaped distributions to be able to perform many statistical analyses.

In general it is not so important to have a normal bell-shaped distribution for qualification tests or for posttests since one is not so interested in having a wide distribution in scores as in identifying those individuals who scored poorly and, therefore, need more training. It is more important to have a wide distribution in scores for pretests or tests that will be subject to extensive statistical analysis. A wide distribution of scores on a pretest will help the trainer to identify those content areas that should be emphasized or deemphasized in the training. Obtaining this wide distribution of scores will require in many cases making the test more difficult to avoid the tendency for everyone to score high. Likewise tests that will be used to select a chosen few for a particular award or task must be difficult enough to be able to discriminate the "good" from the "excellent".

Ease of Administration

Instruments should be made as easy to administer as possible. The mark of a good evaluation instrument is not how many questions it asks or how complex it appears but rather how efficiently it obtains the information sought after (as well as whether it is valid and reliable). This is especially important when the person who designed the instrument will not be the one who administers it. If it is very time consuming, difficult or awkward to administer, it may not be administered in the exact way that was intended by the designer or it may not be administered at all. This is why instruments, whether they are tests, questionnaires or interviews, should be made as easy to carry out as possible and have clear, explicit directions to the person administering it. The instrument should also be easy for the subject or participant to complete. Subjects of evaluations should not be made to feel extremely uncomfortable, embarrassed or have a

very unpleasant experience by participating in an evaluation.

Simplicity and Succinctness

Evaluators should strive to make their instruments simple and brief. If one clearly worded question will obtain the information sought there is no reason to include two or three questions on the subject. Instruments that are longer than necessary will only tend to make the subjects weary and frustrated. Likewise, questions should be stated simply, not couched in professional jargon that the subject may be unfamiliar with.

Precision and Lack of Bias

It is important to state questions precisely, particularly when the trainer or evaluator will not be present when the instrument is administered in order to clear up any confusion that might arise among the subjects. One common mistake that demonstrates a lack of precision is when a question asks more than one piece of information at a time. For example, the question, "Do you believe that oral contraceptives are safe, and that their use should be promoted more vigorously?" is really two questions. A person who believes that oral contraceptives are safe but that they should not be promoted more vigorously would be in a quandary as to how to answer this question. Questions should ask for only one piece of information and ask that information in a precise manner.

Being precise also entails carefully defining all possible response alternatives to remove any doubt a subject might have concerning how to answer a particular question. For example, the question, "On a scale of 1 to 10, how would you rate the success of the training session?" is unclear, because the 1 to 10 scale is not defined. Designing meaningful scales will be discussed later in this section.

Another common error made with respect to precision is when a complete list of alternatives to a question is not given. For example, the question, "How much do you

know about the advantages and disadvantages of the Papanicolaou Test: a) a little; b) a moderate amount; c) a great deal?" does not include an option for a person who knows nothing about the Papanicolaou Test.

Instruments should also be neutrally worded so that the responses are not biased. For example, the question, "How will the skills which you learned in the training help you in your job?" suggests that all the trainees found the training to be helpful. It would be better to phrase this question as, "Did you learn any skills in the training session which will help you in your job?" _____"If so, how will your new skills be helpful?" _____

Economical

Instruments should also be kept economical. Costs should be kept in mind both when deciding what type of instrument to select (e.g. interview versus questionnaire), as well as when deciding how to design a particular instrument (e.g. one page close-ended test versus five-page open-ended test). The total cost of developing and administering the instrument, as well as any possible field testing that may be necessary, should be analyzed to help select among various options. A common tradeoff that training evaluators face is between cost and response rate. A very inexpensive instrument to use to do a followup evaluation of training is a mailed questionnaire. Mailed questionnaires typically have low response rates, however, and may be completely out of the question for participants living in rural areas. The resultant low response rate presents a validity threat, not only due to the possibility of not meeting the required minimal sample size, but also due the threat of "mortality", i.e. participants who do not respond to mailed questionnaires may have very different characteristics from those who do. Another option is to personally visit each participant and interview them. While this will guarantee a high response rate it will also be very costly. Usually only a sample of all participants is selected when it is decided that followup interviews will be done.

Field Test

Instruments should be field tested before they are used for the first time in an actual training evaluation. Despite careful design, it is impossible for a trainer or evaluator to foresee all the possible sources of confusion, anger, or anxiety a question might present to a respondent. An evaluation instrument that will require direct contact and communication between an evaluator and a subject (as in a questionnaire, test, or interview) should be field tested first with a small group of people, similar to those to whom the instrument will later be applied. Validity and a reasonable degree of reliability should be assured for each question in the instrument and for the instrument overall. After this small group takes the test, questionnaire, or is interviewed, the evaluator should attempt to find out whether any questions were unclear, were too complex, or were particularly uncomfortable to respond to. The instrument should then be revised accordingly. Field testing is also important to verify an adequate distribution of responses as was mentioned earlier.

Burden of Data Collection

Except for the use of existing records, all training evaluation instruments represent a burden on someone. Whether a training evaluator utilizes questionnaires, tests, interviews, or observations, he or she will often need assistance from additional agency personnel in order to prepare the instruments, apply them, code them, and interpret them. This of course consumes resources and often puts pressure on staff time that could otherwise be used for other purposes.

This holds true not only for the agency that is undertaking the evaluation, but for the individuals and agencies being the subject of an evaluation as well. When individuals or agencies are asked to collect new information for an evaluation, this represents an additional burden on them.

In order to conserve resources (and amiable personal relations) within the agency performing the evaluation, as well as on behalf of the agency or individual being evaluated, a training evaluator should seek to make data collection as simple as possible, while still obtaining the information necessary. This will involve, first of all, utilizing existing data whenever it is relevant rather than creating new data. When existing data are not available, then new data should be gathered in a manner that does not demand too much of staff time and resources. This guideline is particularly important for followup evaluations that are commonly quite elaborate and require a great deal of data collection.

Given these general guidelines that apply to the development of any type of evaluation instrument, several of the most common instruments used in training evaluation will now be analyzed. When discussing these instruments the following definitions will be used:

Questionnaire: A form of data collection where the subject answers a set of questions in writing designed to elucidate information about the subjects' behaviors, beliefs, or attitudes.

Test: A form of data collection where the subject answers a set of questions or exhibits a particular skill in order to obtain information about the subjects' level of knowledge and skills.

Interview: A form of data collection where the subject verbally answers a set of questions designed to elucidate information about the subjects' behaviors, beliefs, or attitudes.

Observation: A form of data collection where the evaluator watches and records certain behaviors of the subject.

Existing Records: Data collection where the evaluator uses existing records of a subjects' knowledge, behavior, beliefs, or attitudes and does not have any direct contact with the subject.

Questionnaires

In training evaluation, questionnaires are most often used by trainers to elicit information from the participants concerning the progress of the course (e.g. daily evaluation questionnaires), the participants' overall rating of the course and the instructors (final course evaluation), and, at times, the participants' beliefs and attitudes towards the subject at hand. Questionnaires are also commonly used, perhaps more than any other method, in followup evaluations where participants are either mailed or handed questionnaires and asked to answer a series of questions relating to the effect that training has had on their job performance.

The advantages of questionnaires over many other instruments are many. They can be answered anonymously, be administered to many people simultaneously at relatively low cost, and help maintain validity and reliability by requesting the same information from all respondents. They are also usually easy to score and process the data, compared with instruments such as interviews or observations. One disadvantage of questionnaires, however, is that the respondent has to be literate and understand the instructions accompanying each type of question.

A questionnaire can either be mailed, distributed to subjects to fill out while the evaluator waits, or distributed to subjects who return them later. Within training courses, the latter two forms of distribution are used, while for followup evaluations, any three of the forms of distribution can be used. Mailing questionnaires is the least expensive means of sending questionnaires to former participants, but the poor response rate often makes this alternative undesirable.

Questions within a questionnaire can be in the form of open-ended questions, ranking scales, Likert scales, rating scales, checklists, two-way questions, and multiple-choice questions. Examples of each follow, together with a brief discussion of their uses.

1. Example of an open-ended question:

What do you feel were the most important topics covered in this course? _____
_____
_____

This format is useful whenever it would be difficult or impossible to foresee all the different responses that a subject might have to a question, or when qualitative data are sought that can only be obtained by allowing subjects to freely write. For example, asking a question and then following it up with the question "Why do you feel this way?" (or something to that effect) requires an open-ended question. As was previously noted, neutral wording is important. The question "Explain how the topic presented this morning will benefit you in your work," presumes that the topic was relevant to everybody's work and was beneficial. The reader may even feel that the sample question presented presumes too much. Perhaps it should read "Were there any important topics presented in this course; if so what were the most important ones?".

2. Example of a question using a ranking scale:

The following list contains the six principal topics covered in this course. Please place a one (1) by the topic that you feel was most helpful to you for your work as a promoter, and so on. The six (6) will be the topic that was the least helpful to you.

The demographic situation in country X _____
Contraceptive methods: their correct use _____
Contraceptive methods: their contraindications _____
Contraceptive methods: their side-effects _____
Reproductive anatomy and physiology _____
The church's view on family planning _____

This type of scale is useful when the evaluator wants to obtain information on how a subject ranks various items. Simply asking "What topic was most helpful to you" does not provide the comparative information that this ranking scale question does. Note the importance of defining the scale. The subjects must be explicitly told that one (1) represents the most helpful topic and six (6) represents the least helpful topic.

3. Example of a question using a Likert Scale:

Please circle the response that best fits your response to the statement.

A. This workshop met my expectations.

strongly agree	agree	disagree	strongly disagree

A Likert scale is often used when an evaluator is interested in how strongly a person favors or disfavors a certain attitude or value. As was mentioned earlier, a complete list of alternatives should be provided and an equal number of positive and negative responses should be given. If only the first three responses were given, there would be no appropriate response for a person who strongly disagreed. Sometimes, Likert scales include a neutral response category such as "no opinion", "neither agree nor disagree", etc. Such a response category may be appropriate for respondents who are likely to be indecisive about a certain subject.

4. Example of a question using a rating scale:

Please circle the number that corresponds to your feelings about each statement.

A. This workshop met my expectations

7	6	5	4	3	2	1
strongly agree		agree		disagree		strongly disagree

A rating scale is useful whenever more responses are desired than a Likert scale provides. A person who somewhat agrees or somewhat disagrees, for example, can circle number 4 in the scale above whereas they would have no good option with the Likert scale above. Rating scales can be made as long as necessary; ten or more possible responses are sometimes used. However, it is usually unwise to present more than seven responses on a rating scale, otherwise confusion will likely arise as to the subtle distinctions between an 8 and a 9, for example.

4. Example of a question using a checklist:

In the following list check all the items that correspond to your beliefs.

__ I believe that the woman should bear the primary responsibility for practicing contraception.

__ I believe that contraceptives should never be given to unmarried girls under 18 without the consent of their parents.

__ I believe that only women who have the consent of their husbands should be allowed to become sterilized.

__ I believe that condoms should be officially made available in secondary schools to all who want them.

Checklists have many uses in questionnaires as they allow more than one response to a question. Directions should be clear to state, however, that more than one response can be checked.

5. Example of a two-way question:

I am presently protecting myself from pregnancy by use of a method of contraception, either by myself or by my partner. YES ____ NO ____

Two-way questions are only appropriate when there is a clear yes or no answer to a question without other alternatives or exceptions. They are useful for surveys when large numbers of people are asked a question and the evaluator wants to be able to quickly code the responses. The amount of information they are capable of providing is very limited, however, when compared to other question formats. As was previously mentioned, it is very important that only one piece of information be asked at a time. If the statement above read "I am presently protecting myself....and I go to see the doctor at least once a year," the responses received would be invalid because the evaluator would not know whether the answer corresponded to whether the person was using contraception, seeing a doctor annually, or both.

6. Example of a multiple-choice question:

Please circle the letter corresponding to the average number of home visits (to promote family planning) you make each month. a. 4 or less b. Between 5 and 9 c. Between 10 and 14 d. 15 or more

Multiple-choice questions are widely used in questionnaires because they permit control over the responses (unlike short-answer questions) yet provide a range of responses (unlike two-way questions) that a respondent can usually choose from comfortably. They are also usually easy for the respondent to complete and can be scored objectively and quickly. Many times, a category such as "other" has to be included in a multiple-choice question so as to accommodate all possible responses. In the example above, however, the choices presented cover all possible responses. In constructing multiple-choice questions it is important that the wrong answers be plausible choices so that it is less likely that the correct answer will be chosen simply by a process of elimination.

Each of these types of questions has its proper place in a questionnaire depending on the type of information being solicited and the type of subject responding to the questionnaire. Typically evaluators use open-ended questions, ranking scales, Likert scales, and checklists when seeking information about the subjects' beliefs or attitudes. Two-way questions, short-answer questions, and multiple-choice questions are typically used to solicit information from subjects concerning their behaviors. Of course any of the seven types of questions can be adapted so as to gain information about attitudes or beliefs.

One important consideration to keep in mind when selecting a particular format for questions is how easy they will be able to be coded. Obviously coding an open-ended question takes much more time and is much less amenable to quantitative analysis than a multiple-choice or two-way question. Whenever possible, questions should be formatted in a way that can be easily coded and numerically tabulated so as to present frequencies and percentages for each response. This generally requires a close-ended question (where all the responses are given and the respondent is asked to select from among them). However, open-ended questions, while less appropriate for statistical analysis, are more appropriate for obtaining information that cannot easily

be provided in the form of close-ended questions. For example, a trainer who wants to know what the participant liked best about a workshop can never provide all the possible answers to this question in the form of a multiple choice question. This question is best presented in an open-ended question format where the participant if free to answer from an infinite range of possibilities.

Of course, a trainer is free to include several types of questions in a questionnaire, each appropriate to the particular type of information sought. When there is this mixing of question types, it is generally better to include the close-ended questions first. These are usually easier for the respondent to answer and will make it more likely that a respondent will complete the questionnaire. It is also preferable to include near the end of the questionnaire any questions that may be very personal or sensitive for the respondent.

Tests

Tests are given in a variety of forms in the field of training evaluation. Tests of knowledge can either be in written form or be given orally. Tests of skills, on the other hand, are known as performance tests and require the subject to exhibit a particular skill that is then evaluated. Knowledge tests, whether in the form of pretests, midtests, posttests, or tests of knowledge retention, are ubiquitous among training evaluators. They are usually inexpensive to perform, can be easily carried out on large groups, and can usually be readily scored and tabulated.

Performance tests are used less often by training evaluators because they are usually harder to develop and because simulating the actual job skills of many workers is difficult to do, particularly when they perform abstract tasks rather than manual skills. Nevertheless, performance tests should probably be used to a greater extent than they are because they are often a better predictor of how a worker will perform on the job than are tests of knowledge.

Tests of Knowledge

Two major types of knowledge tests are norm-referenced tests and criterion-referenced tests. A norm-referenced test compares participants with each other or to other groups rather than to specific learning objectives. When norm-referenced tests are scored, the raw score for each participant is converted to a norm-referenced percentile, which shows how the participant ranked with respect to the "norm" or average score. This average score can either be taken from a larger population of participants who took the test or from the participants at hand in a particular course. This type of test is commonly used in universities for entrance exams where test score norms have been established and individuals who take the entrance exam have to score above a certain percentile in order to be admitted. It may be used in training evaluation if the trainer wants to pass, or select for a certain task, only the top ranking individuals in a course. For example, a program might want to select only 20 promoters for a particular program but allows anyone interested to attend a one-day training course. If 35 people show up and take the training, the trainer could use a norm-referenced test and choose only the top 20 scores on the posttest for admission into the program.

The other type of knowledge test is a criterion-referenced test. This is more common and more appropriate for most training evaluation contexts. It involves administering a test with a pre-determined cut-off score. With a criterion-referenced test, the concern is not with how participants rank in comparison to each other or another pool of participants, but rather whether or not they meet the minimum standards for the training. Criterion-referenced tests usually report the raw score of the participant and then a pass/fail judgment. Sometimes scores will be divided into three categories: pass, "monitor closely", and fail. The second category might be used to alert the participants' supervisors or other personnel that these participants passed the course but need close supervision and continuing education. This type of communication should be made as specific as possible with instructions given as to <u>what</u> should be monitored closely. Figure 2 presented an example of how this communication might take place between

trainers and supervisors.

The types of questions that are included in tests of knowledge are many of the same that are included in questionnaires. Open-ended questions, two-way questions (true/false questions), and multiple-choice questions are all common formats for presenting questions in tests of knowledge. The same precautions mentioned for these types of questions under the questionnaire heading also hold true when they are used in tests.

In addition to open-ended, true/false, and multiple-choice questions, there are two other formats often used for asking questions in tests; these include matching items and completion items. An example of each follows.

1. Example of a matching item question:

Column I lists internal organs of the male and female reproductive systems and column II lists functions of internal reproductive organs. Please match one function from column II with the appropriate internal organ it corresponds to in column I. Each lettered entry in column II may be selected once, more than once, or not at all.

Column I	Column II
___ ovaries	A. secretes a mucoid fluid which helps nourish and transport sperm
___ vas deferens	B. causes growth of the penis and testes
___ seminal vesicles	C. develops and releases mature ova
___ uterus	D. passageways for sperm
	E. protects and nourishes the fertilized egg
	F. pulls the ovum into the fallopian tube after ovulation

With matching items, it is important to state the directions clearly (e.g. "each lettered entry may be selected once..."), so as to avoid possible confusion, which often occurs with this type of question. It is also a good idea to provide more possible responses in column II than there are items to match in column I; otherwise respondents can more easily guess at answers by a process of elimination. For matching items in particular (but for any type of question as well) it may be appropriate to include an example at the beginning, with the answer provided, so respondents can better understand what they are supposed to do. This is particularly important for respondents with limited amounts of formal education.

2. Example of a completion item question:

Please fill in the blank with the correct response. A. The _____ hormone is responsible for keeping the corpus luteum alive when the egg is fertilized and is the hormone tested for in pregnancy tests.

Completion items, or short-answer items, are useful only when there is one and only one correct word or phrase that can fit in the blank. In the example above, "HCG" is the only appropriate response. Completion items are often useful when asking for information that respondents had to memorize. The advantage of completion items are their ease of construction; an evaluator does not have to invent many believable but wrong answers as in multiple choice questions. Their principal disadvantage, however, is their limited applicability to situations where there is only one right answer.

Performance Tests

Performance tests are used to evaluate participants' performance of particular skills. The problem with these types of tests as mentioned previously, is the difficulty of creating realistic simulations of actual job skills. For example, how can one test in the classroom whether a participant possesses sufficient skills in verbal communications in order to give talks to large audiences on family planning? Role-plays can be done in the classroom, but just because a participant appears capable of giving a talk to a small group of people he or she is well acquainted with does not mean the participant will be able to perform as well in front of a large group of potential strangers. One other problem with performance tests is that they generally have to be administered individually, often with the use of special equipment or volunteers.

Despite, these problems, however, performance tests are very useful because they do attempt to test the performance of certain behaviors (even if these are weak simulations of the real thing) rather than just knowledge or attitudes, which can have very little correlation with job performance. Because they simulate real-world situations, performance tests generally have higher predictive validity than knowledge tests for estimating which participants will be successful on the job.

Performance tests can be used to evaluate a number of different skills required by workers. In the family planning field these might include the following:

Communications Skills

- ability to communicate to couples during a home visit
- ability to communicate to a large audience during a community talk
- ability to communicate program objectives when recruiting promoters

Clinical Skills

- ability to perform a pelvic exam
- ability to insert an IUD
- ability to take blood pressure

Administration Skills

- ability to correctly and neatly record family planning user information
- ability to correctly fill out contraceptive supply request forms
- ability to correctly record revenues received from sales

In order to test these and other skills, training evaluators can use methods such as role-plays, practical assignments, and special simulations using instruments or "dolls".

1. Role-Plays. With this common method, one or more participants are asked to perform a certain role that simulates a real-life situation. For example, one participant could play the role of an oral contraceptive user who comes to a promoter for advice on side-effects that she is experiencing. Another participant (the one being evaluated) plays the role of the promoter and responds to her questions. The "promoter" is evaluated in the process by the instructor, and often by fellow participants, on the basis of how well he or she did in giving appropriate advice to the "user". When using role-plays, it helps if the instructor has prepared an evaluation sheet with items to be checked-off that will help the instructor remember what a particular participant did or did not do. In this way specific feedback that will benefit the participant can be given by the instructor. This is important both to help the instructor remember specific behaviors evidenced during the role play, as well as to help the instructor be more impartial when assigning a score to the participant. Such a checklist might include items such as the following:

___ The participant correctly distinguished between common side-effects and potentially life-threatening complications.
___ The participant recommended a visit to the clinic if the problem fails to go away in the next week.
___ The participant listened carefully to all the user's complaints and asked appropriate followup questions to fully understand the nature of the complaints.

2. Practical Assignments. Participants can be given practical assignments to test particular skills. For example, as part of a course for family planning promoters, the participants could be asked to spend one afternoon of the course doing home visits in the community where the course was taking place. The promoters could be accompanied by instructors who would evaluate the performance of the promoters during several practice home visits. In this example, like the one before it, instructors should have a list of items representing key behaviors that the promoters should demonstrate during their home visits. The instructors should probably not fill out the checklist, however, until after the visit has been completed so as to not interfere with the communication taking place between the promoter and the community resident.

3. Instrument Simulation. Sometimes instruments, including mannequins or dolls, can be used to test specific skills. For example, a number of lifelike dolls exist for demonstrating and practicing how labor and delivery are performed. Likewise, one company has developed model breasts and testicles made of silicone and rubber for teaching how to detect breast and testicular cancer. Unfortunately, many teaching mannequins and dolls are very expensive. In addition, they can never create totally lifelike situations. Nevertheless, if there is a doll or other instrument that can be used to simulate a family planning patient, for example, it may be appropriate to acquire, particularly if it would be impractical, impossible, or ethically ill-advised to use actual family planning patients to train participants.

If instruments are to be used in evaluating specific skills, then the evaluator would observe the behavior of the participant with the instrument, just as he or she would if a live patient were being used. Again checkoff lists are probably the best means for evaluating performance of skills. Sometimes the instrument itself may be designed to automatically perform the evaluation. For example, many dolls used for teaching cardiopulmonary resuscitation contain a computer printout

within them that records the pressure and rate of the hand compressions used in the process of CPR.

Interviews

Structured Interviews

Interviews are used less frequently to evaluate training courses but are common in many followup evaluations. As defined earlier, an interview is a form of data collection similar to a questionnaire except that it is given verbally. Interviews can either be structured or unstructured. Structured interviews are much like questionnaires; specific questions are asked with little room for deviation. Structured interviews are usually used instead of questionnaires when the respondents have a low level of literacy, when a higher response rate is needed than a questionnaire might provide, or when it is desirable for an interviewer to be present to clarify an incomplete or incoherent answer to a particular question.

Unstructured Interviews

Unstructured interviews allow for more flexibility on the part of the interviewer. They generally consist of a fewer number of general questions that the interview asks. The responses to these questions are then further probed with followup questions until the interviewer feels he or she has a sufficient level of understanding regarding how the respondent thinks or feels about a certain subject. Unstructured interviews contain all the advantages and disadvantages that such flexibility brings. On the positive side, they permit the interviewer to use his or her judgment in asking questions and in forming followup questions so as to gain a more complete understanding of the subject being dealt with. Typical followup questions that an interviewer can use include "can you give me an example of that," or "is this what you mean...." They are also more appropriate for dealing with sensitive or awkward subjects since the interviewer, if skillful, can use

lead-in questions that try to put the respondent at ease. The interviewer can also be alert for body language from the respondent that signals that a particular line of questioning should stop.

On the negative side, unstructured interviews require more experienced, skilled interviewers than do structured interviews. They are also less reliable than structured interviews because they are never given in the same way. A different interviewer might gain different answers from the same respondent because different followup questions were asked. The validity of unstructured interviews, however, will often be higher than that of structured interviews, particularly when information might be suppressed by a respondent that would only be revealed upon followup probing. However, validity might be less for unstructured interviews if the interviews given to different respondents vary between each other too greatly.

Face-to-Face Interviews

Interviews, whether structured or unstructured, can be done either face-to-face or on the telephone. The advantage of face-to-face interviews are that they are more personal and respondents will probably be more willing to take the time to answer questions. They also are more appropriate for unstructured interviews which depend, in part, on the interviewer observing non-verbal reactions from the respondent so as to be able to formulate appropriate followup questions and know how to approach sensitive topics. The interviewer may also gain a greater knowledge about how the respondent really feels about a subject, apart from what is said, by observing clues in the tone of voice, gestures used, etc. The major disadvantage with face-to-face interviews are their expense both in time and travel costs.

Telephone Interviews

Telephone interviews have the advantage of being inexpensive and convenient for both the interviewer and respondent. Some respondents may even feel more comfortable responding to sensitive questions when they don't have to look someone in the eye. The major disadvantage with telephone surveys, however, is that many people do not have telephones and can only be interviewed in person. If an evaluator wanted to do followup interviews with a 50 percent random sample of the participants in a particular course, he or she could not interview by telephone unless all the participants had telephones. If only half the group had telephones and the evaluator chose these 50 percent to interview by phone, it would be very likely that the results would be unrepresentative of the entire group because only the participants with phones (and therefore, probably of a higher socioeconomic status) were chosen for followup.

One final note to make about conducting interviews in general is that interviewers should be instructed to write down the responses of respondents as close to verbatim as possible. There is a great tendency to overgeneralize and/or paraphrase what the respondent has said. It is usually much better for interviewers to record what subjects actually say and then later, during the coding session, to lump responses into more general categories. This can be done by using a tape recording device, but in many circumstances it is preferable to make notes during the interview and then to write a detailed summary immediately afterwards.

Observations

Observations are used principally to provide contextual, anecdotal, or process evaluation information during training courses. Virtually every course evaluation includes a summary by the instructor of his or her overall impressions of how the course went. Many of these impressions are formed, not just by analyzing test information or reading the results of questionnaires given to the participants, but by observing the course in its

overall context. By closely observing a course in progress, information can be obtained on how participants responded to a particular type of small group exercise, the extent of cohesion and positive group dynamics among the participants, the comfort of the surroundings, and hundreds of other large and small bits of information. Often another instructor who is not currently speaking, or an outside observer, is in the best position to note many of these subtle aspects of the course.

Observations can also be quite useful as an evaluation tool for recording on-the-job behaviors. Since self-reported behaviors are sometimes suspect, observations of actual behaviors can often yield more realistic pictures of what ex-participants are doing on the job as a result of training. The problem with observing behavior, however, is the problem of external validity associated with making subjects aware that they are being evaluated. The moment that most workers know the "boss" or a special evaluator is on hand watching their behavior, they will usually begin to act atypically.

Observations are rarely used as the sole means for evaluating training or the effects of training. This is due to the time and expense that is involved in systematically observing participants, particularly when a followup evaluation is required. It is much simpler and quicker to mail or drop off a questionnaire when doing followup evaluations, or even to interview ex-participants. Since observations generally require prolonged periods of observation, resulting in a high cost per unit of observation, it is considered a relatively expensive form of data collection. Nevertheless, observations are very useful for obtaining qualitative evaluation information that is unobtainable with other forms of evaluation, except perhaps for open-ended questions and unstructured interviews. Observations are also especially useful for validating data obtained on small samples using other methods. If these other methods appear valid using observations as a check, then they can then be used with greater confidence with larger samples.

If observations are to be used to systematically collect information on participant behaviors, then there should be a checklist or some other recording instrument to

facilitate the gathering and coding of the information. As with performance tests that are observed, evaluators should first determine the information they wish to collect that will be of relevance for evaluating the training. The number of items to be observed should probably be under ten. The evaluator should then use the checklist or recording instrument to record each time a certain behavior occurs. For example, the evaluator may observe a nurse providing information to a couple about contraceptive choices. The evaluator may be interested in whether the couple is really being provided with adequate information about all the methods available to them. The evaluator might note on a checklist, one by one, whether the nurse informed the couple about the correct use, benefits and risks of each method the clinic has available, as well as other methods than can be obtained from other pharmacies or clinics. Of course, observations can also be used in less structured situations to supplement other evaluation data and provide background information that will make quantitative data more easily understood.

Existing Records

A final form of data collection for evaluating training is the use of existing records. Utilizing existing records will most likely be relevant only for evaluating the outcomes of training, including the effect of training on job performance, and the short and long-term effects on the target population. The advantages of using existing records are several: they are quickly and easily obtained, usually without placing additional demands on staff time, they are often viewed as objective and, therefore, credible, and they are inexpensive to obtain. The principal disadvantage is that no control can be exerted over them; the only information that can be collected is the information that exists in the records. In short, it is information that somebody else deemed important but may lack vital information for the purposes of a particular training evaluation.

One example of using existing records to evaluate the effects of training on job performance would be to use a CBD program's existing records for the number of new users who are served by the program each month. This information is usually collected

from each distributor every month. If a CBD program provides training to 40 new distributors in a given province, an evaluator can easily access the monthly reports from those 40 distributors to find out the number of new family planning users each has served over a given number of months. The evaluator can then draw a conclusion as to the effect of training on the program's delivery of services.

Another example might be where an evaluator was interested in documenting the effect of training dozens of new family planning promoters in a particular community. The outcome measure might be the increase in the number of family planning patients coming to the local clinic in relation to a comparison clinic in an adjacent community. If so, it would probably be very easy to access old and new clinic records for the number of visits of family planning patients to both the experimental and comparison clinic.

These are just two examples of the many uses that existing data can provide in training evaluation. Of course, oftentimes an evaluator will find the existing records incomplete or inappropriate for his or her purposes. In the second example above, it could be that records are only kept on the number of patients who come to the clinic with no tabulations done as to the primary reason for the visit. It could be that the evaluator would have to pore over each chart to look for the reason for the visit before it could be determined that a given visit was a "family planning visit." Worse yet, many of the charts could be illegible or not provide any clues as to the primary reason for the visit. In those cases, a review of existing records will not suffice as an evaluation instrument.

By way of summary, Table 3 presents what has been said regarding the various evaluation instruments and their advantages and disadvantages.

TABLE 3

COMPARISON OF EVALUATION INSTRUMENTS

Instrument	Advantages	Disadvantages
Questionnaire	Low cost Easy to administer May be made anonymous Can be given to large groups	Low response rate when mailed Can be confusing for respondents No assurance questions were understood Must be literate
Knowledge Test	Low cost Easy to administer Can be given to large groups Easily processed	Often poor correlation with behavior Often threatening to respondents
Performance Test	Tests skills which often correlate with job performance Does not require literacy Close to real-world situations	Costly and time consuming Difficult to create simulations Can require special equipment
Interview	Flexible Generates qualitative information Can increase validity Answers can be clarified Sensitive topics may be probed	Less reliable Costly and time consuming Requires trained interviewers Difficult to process
Observation	Can be more valid for info. on behaviors Generates qualitative information Often less threatening	Costly and time consuming Less reliable Requires trained observers Can induce behavior changes
Existing Records	Convenient and economical Avoids burdening staff with additional demands May be more objective	Inflexible Data often incomplete or overly generalized Records often poorly organized

CHAPTER IV: ANALYSIS AND INTERPRETATION OF TRAINING EVALUATION RESULTS

Interpreting training evaluation results is perhaps the most challenging phase of training evaluation. After specifying a particular training evaluation design, developing the evaluation instruments that will be used to collect the data, and then actually collecting the data, the training evaluator is then faced with the difficult task of analyzing the data collected, presenting them in an evaluation report, and interpreting their significance and meaning. This latter task of data analysis, presentation, and interpretation is by no means easy and requires careful thought and attention when being carried out. For complex analyses, it requires the aid of statisticians familiar with the myriad formulas used in data analysis and interpretation. This section will attempt an overview of how to present, analyze, and interpret training evaluation results using some of the more common and simpler methods available. Before entering into this discussion, however, a brief summary of the different types of data that exist is presented.

LEVELS OF MEASUREMENT

Training evaluation results, like any other data, can be measured and reported in any one of four ways. Each has its own limitations as to the type of statistical analysis that can subsequently be performed on it.

1. Nominal. In nominal measurement, data are presented in unordered groupings that differ in name only, without any value being placed on the various categories. Nominal data categories include things such as race, sex, occupation, etc. The categories used in nominal measurements can only be said to be different from one another, without any conclusions made as to how different they are or which is better or worse than the other. In training evaluation, an example of results in nominal form might be a list of participant characteristics which said, for example, that 40 percent of the class was male, 60 percent female, 30 percent

from urban areas, and 70 percent from rural areas.

Nominal data are the weakest of all measurements in terms of what types of statistical manipulations can be performed on them. With nominal data a mode can be calculated (the most frequently occurring category) as well as a percentage distribution (the percentage of total cases pertaining to each category). However, a mean and other measures of central tendency cannot be calculated using nominal data.

2. Ordinal. Ordinal data permit descriptions in terms of rank, but with no unit of distance between the classifications. Presenting the results of a course evaluation according to the number or participants who thought elements of the course best, second best, third best, etc., is an example of ordinal data. These classifications do have values which can be used to rank them and determine whether one element of a course was perceived to be better than another, but there is no way to numerically state how much better one element was from another using ordinal data. There is no way to tell, for example, what the distance between best and second best is numerically, only that one is better than the other.

With ordinal measurements, not only can the mode and percentage frequencies be calculated, but various other statistical calculations can be performed as well including the median and percentile. However, a mean and standard deviation (to be discussed later) cannot be calculated from ordinal data.

3. Interval. With interval data, measures are both rank-ordered and have equal distances between measures. A test score, for example, is an interval measurement. Not only does a score convey rank (e.g. a person scoring 80 scored higher than someone scoring 70), but it conveys a known distance as well (e.g. a score of 80 is 10 points higher than a score of 70). One characteristic of interval data that set them apart from other numerical data is the absence of a true zero

point. For example, a score of zero on a test does not mean the total absence of knowledge. Temperature measurement is another interval measure since the difference between 0 and 1 degrees centigrade is a known and constant distance. However, 0 degrees does not indicate the total absence of any heat.

All the statistical manipulations that can be performed for nominal and ordinal data can be used with interval data. In addition, a mean and standard deviation can be calculated when working with interval data.

4. Ratio. The final type of measurement is a ratio measurement. Ratio measurements are identical to interval measurements, with the only difference being that the former are based on a true zero point. Age, parity, and number of years of schooling are all ratio measures since the number 0 in these examples does indicate a total absence of the characteristic. The reason that measurements based on a true zero point have significance is that a ratio can be calculated in such cases. For example, a person who is 30 years old is exactly twice as old as someone who is 15 years old. Using an interval measurement such as a test score cannot be used to calculate a ratio, however. A person who scored 80 on a test is not necessarily twice as knowledgeable as a person who scored 40. The practical significance of the difference between interval and ratio measurements is probably not great for the subject of training evaluation, however, and in many cases the literature omits reference to ratio measurements entirely.

DESCRIPTIVE STATISTICS

Descriptive statistics refer to how numerical terms, entities, proportions, and characteristics are described. The other major body of statistics that will be discussed later is inferential statistics which, as the name implies, is used to make inferences or conclusions about data and their significance.

Describing data correctly and meaningfully is the central task of descriptive statistical analysis. Often training evaluators simply report the scores of pretests and posttests, for example, and leave it up to the person reading the report to make meaningful conclusions from the results. This may be appropriate in some circumstances when the number of scores reported is small, but for most circumstances, training evaluators should use one or more of the methods described in this section to help make the evaluation results more understandable and convey a clearer picture.

Frequency Distributions and Percentages

A frequency distribution is a simple, useful way to present raw data. It consists of reporting data in tabular form by categories and by the frequency with which the categories appear in the raw data. The following table is an example of a frequency distribution showing the results of a posttest.

TABLE 4.1

FREQUENCY DISTRIBUTION

Score	Frequency
60	1
66	2
67	1
69	1
70	1
71	2
73	2
74	4
75	3
76	4
79	5
80	3
81	4
82	1
84	3
85	2
88	1
90	1
	n = 41

From this frequency distribution it can be seen how many individuals received a particular score. It is easier to gain an overall picture of the results using a frequency distribution than it is by looking at raw data because a frequency distribution arranges the scores in order and allows the reviewer to quickly gain a sense of where most people scored.

A variant of a frequency distribution is the grouped frequency distribution, in which scores are clumped together or grouped. A grouped frequency distribution for the data above might look like this:

TABLE 4.2

GROUPED FREQUENCY DISTRIBUTION

Score Range	Frequency
60-64	1
65-69	4
70-74	9
75-79	12
80-84	11
85-89	3
90-94	1
	n = 41

Grouped frequency distributions have the advantage of being able to clump frequency data that are very lengthy into manageable units that are more readily comprehended.

Percentages are another useful way to present data. Percentages are calculated by dividing the frequency for a particular score or score range by the total frequency. For example, one might wish to report the percentage of participants scoring less than 70 (perhaps the pre-determined failing point) in the posttest example used above. If the scores are already tabulated as in the grouped frequency distribution above, then it is

easy to simply take the frequency of people scoring less than 70 (5) and divide it by the total frequency (41), which yields a percentage of 12. Usually frequencies and percentages are presented together under two columns labeled "frequency" and "percentage" situated side-by-side next to the score or score range.

Displaying Data

While the presentation of frequency distributions and percentages are a big improvement over the simple reporting of raw scores, at times it is desirable to display data using a variety of types of graphs. Graphing data, particularly when there are many data points, is an excellent way of making a strong visual impression of what the data reveal and minimize the possibility of misinterpreting the data. Graphs are also especially appropriate when reporting data in staff meetings, conferences, or other forums where quick, easy-to-understand messages are needed. The two most common types of graphs for illustrating frequency distributions are frequency polygons (line graphs) and histograms (bar graphs). Each is illustrated below utilizing the data presented in Table 4.2. A pie chart is another form of graph that is particularly suited for illustrating how a particular total or "pie", is subdivided. Figure 3.3 illustrates a pie chart showing how the budget of a hypothetical training course was allocated.

FIGURE 3.1: FREQUENCY POLYGON

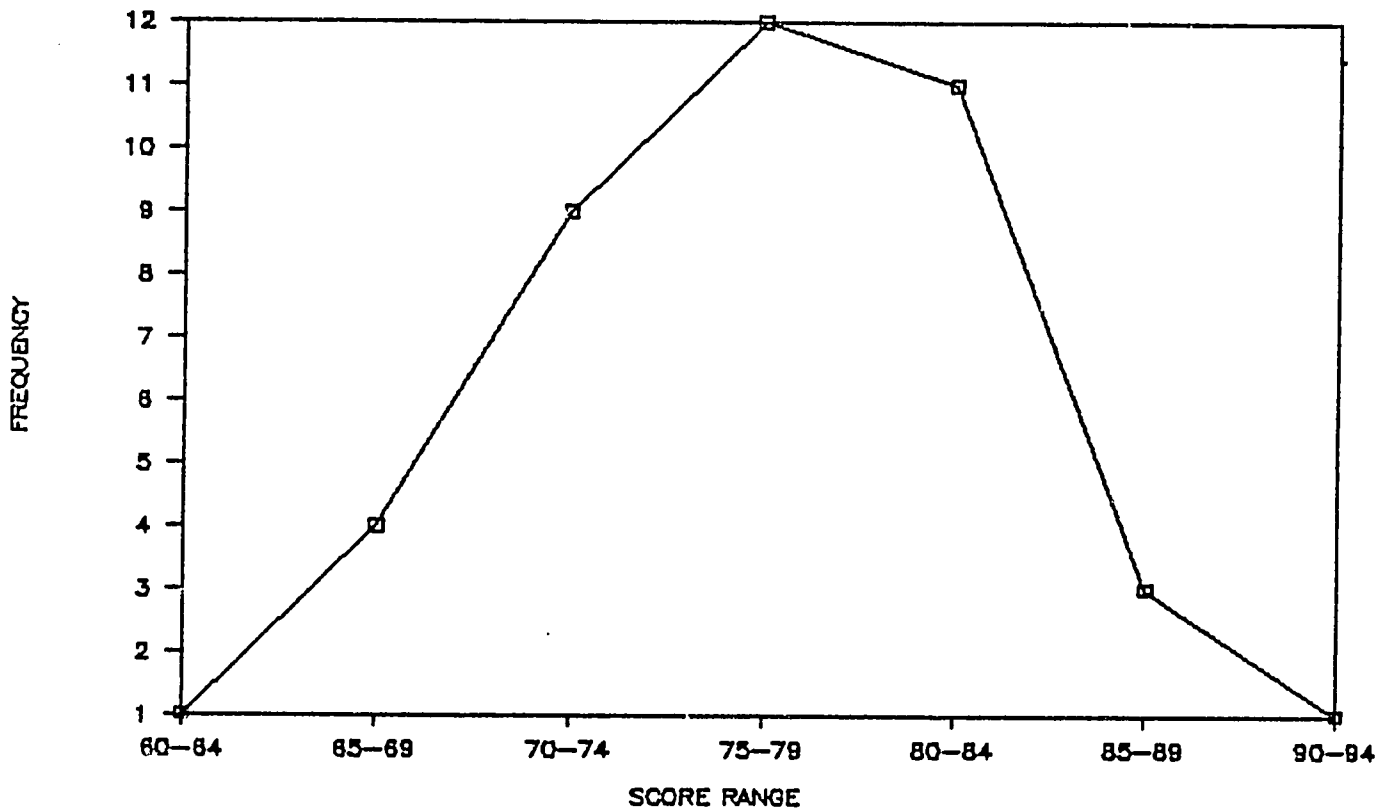


FIGURE 3.2: HISTOGRAM

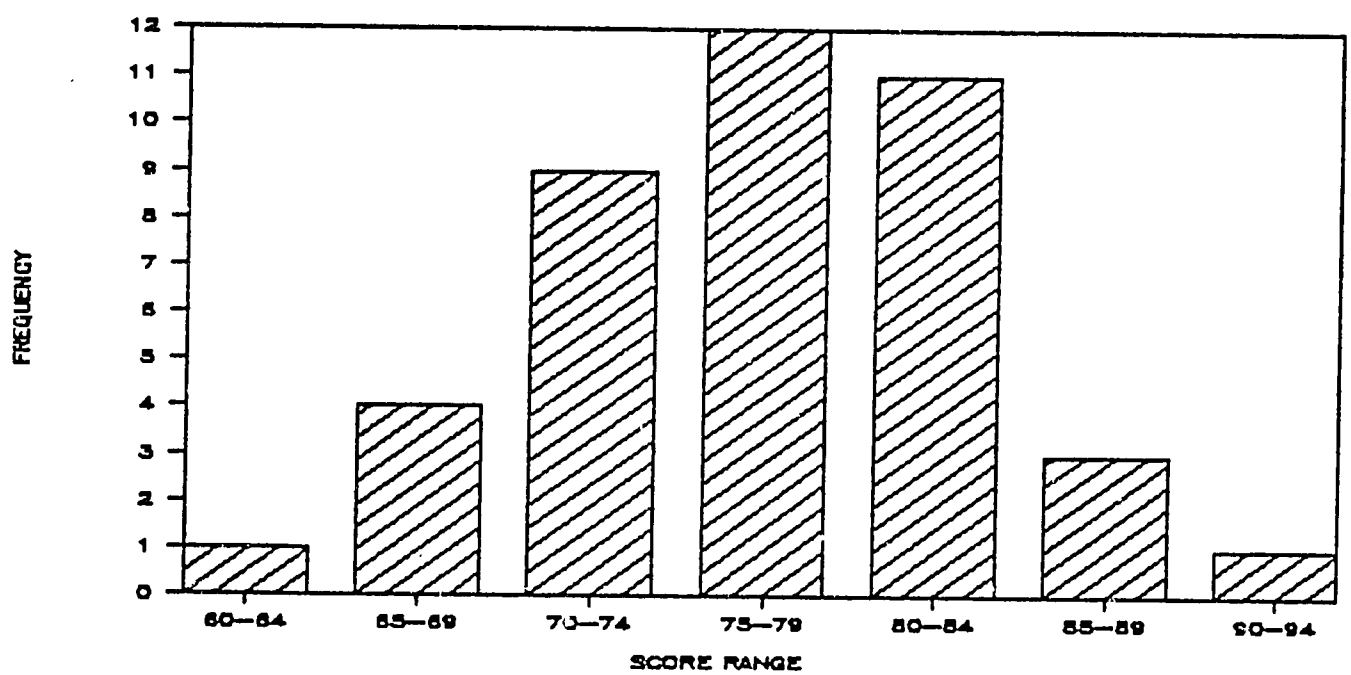
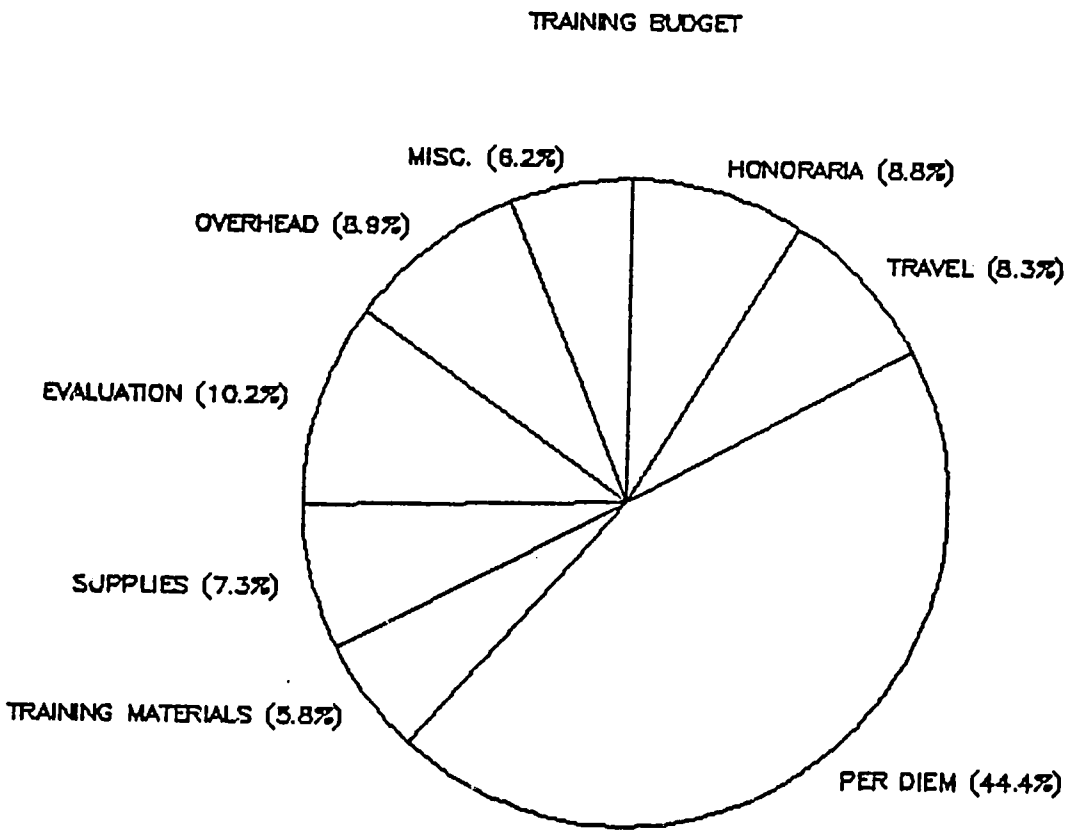


FIGURE 3.3: PIE CHART



Figures 3.1 and 3.2 convey the same message using different formats. In one quick glance the reader can tell from the graphs that the most frequent score on the posttest was between 75-79, with few people scoring below 65 or above 90. The frequency distribution tables also convey that information but not as rapidly as the graphs. The pie chart example also conveys a rapid message of how the budget of a course was allocated. Reading the budget figures would provide the same information, but not nearly as rapidly nor would it make as much of an impact on the reader as seeing the data visually.

Each graph has its own particular characteristics that people find desirable or undesirable and there is no one correct graph to use in a particular situation. Whatever seems easiest to present and causes the least amount of confusion to viewers should be used.

Graphs such as these can be easily made with computer software programs such as Lotus 1-2-3. In addition, they can be made by hand with rulers and compasses. Thin colored tapes sold in art stores are very handy to use when designing histograms or frequency polygons. Pie charts are the most difficult to do by hand and require the use of a protractor to measure the appropriate angles and a compass to draw the appropriate size circle.

Measures of Central Tendency

Presenting data in the form of frequency distributions and percentages, and graphing the resulting tables, can help to convey important information about the data, but only in a general sense. Looking at the grouped frequency distribution in Table 4.2 readily transmits the information that most of the participants scored between 75 and 84. What is not known from a cursory look at the raw data or at the tables listed previously is what the average test score was, or what the score was that divided the top 50 percent of the class from the bottom 50 percent of the class. These measures are the mean and median, respectively, and together with the mode they consist of measures of "central tendency", or the typical performance of the group.

1. Mean. The arithmetic mean is the average for a group of numbers. It is obtained by adding all the values and dividing by the total number of scores. As was mentioned earlier, the mean can only be calculated for interval or ratio data. Nominal and ordinal data cannot have a mean calculated for them, as should be obvious by trying to think of what the "mean nationality" would be for a group of 20 Ecuadorians and 15 Mexicans.

When data are presented in the form of frequency distributions, the mean can be calculated by multiplying each score by the frequency corresponding to it. For example, from the frequency distribution in Table 4.1, the mean score would be calculated in the following manner:

(X) Score	(F) Frequency	F (X)
60	1	60
66	2	132
67	1	67
69	1	69
70	1	70
71	2	142
73	2	146
74	4	296
75	3	225
76	4	304
79	5	395
80	3	240
81	4	324
82	1	82
84	3	252
85	2	170
88	1	88
90	1	90
	n = 41	Total 3,152

Mean ($\bar{X}$) = 3,152 ÷ 41 = 76.9

2. Median. The median is the middle score of the distribution of ordinal, interval, or ratio measures when these measures are arranged in order. In other words, it is the value that separates the measures into two groups, with 50 percent of the measures above the median and 50 percent of the measures below the median. If there are an even number of data, the median is the average of the two middle values.

Using the same data from Table 4.1, the median would be obtained by first calculating what the middle score is. In this case it is the 21st score, since there are a total of 41 scores and the 21st score is the exact middle score with 20 other scores each lying above

and below it. The next step is to count from either the top or bottom of the frequencies until the 21st score is found. In this case it corresponds to the value of 76. The median score would, therefore, be reported as 76.

Note than this was also the approximate value of the mean. If a distribution contains no great extremes of scores, then the mean and median will be close to each other. However, extremely high or extremely low scores can alter the mean greatly while the median is relatively unaffected. For example, if the three people who scored 75 in the posttest example above had instead scored 0, the mean would be reduced from 76.9 to 71.4, while the median would continue to be 76. In cases where there are extreme values that would affect the mean to such an extent that it would give a false impression of how most people did, it is better to report the median, or better still, to report them both.

3. Mode. A final measure of central tendency is the mode, or most frequently occurring score in any distribution. Modes can be reported using any kind of data, including nominal data. The mode is already calculated when data are arranged in frequency distribution tables since the score with the greatest frequency is the mode. In the example of posttest scores from Table 4.1, the mode would be 79 since its frequency is five--more than the frequency of any other individual score.

Measures of Dispersion

Measures of central tendency can provide very useful information about data, but by themselves they do not provide a very complete picture. Evaluators usually want to know not only where the center of a distribution lies, but how widely dispersed or spread out the scores are in the distribution. To quantify this dispersion, three measures are used which are the range, variance, and standard deviation. These three measures all indicate how far the individual values in a grouping of data tend to be from each other.

To appreciate the importance of calculating measures of dispersion, consider two groups of trainees who were given a posttest and whose scores were the following:

TABLE 5.1

POSTTEST SCORES

Group A's Scores	Group B's Scores
51	62
58	64
63	65
67	65
72	66
79	68

The mean and median for the scores of both group A and group B are 65. If the trainer reported only this information in an evaluation report, the person reading the report would likely be led to believe that both groups were fairly similar in terms of their performance in the training. But as can be clearly seen by looking at the raw scores, group A's participants experienced very different outcomes with two individuals scoring very low and two individuals scoring very high. The individuals in group B, on the other hand, had much more similar results.

This information may be important, especially when criterion-referenced tests are given that require participants to achieve a score higher than a given cut-off point. If the failing score for the test given to the participants above was 60, then two individuals from group A would have failed the test while none in group B would have failed it. Reporting measures of dispersion is also important for conveying information (to those who may not have the raw scores to review) about the likelihood of there being individuals who scored very low on a particular posttest and who will, therefore, require more individualized attention and continued in-service training in order to perform adequately.

Calculating measures of dispersion (particularly the standard deviation) on pretests is also a useful way for trainers to alert themselves to the existence of a heterogeneous group. If pretest scores show a high level of dispersion, the trainers will want to be alert to the presence of individuals who may require extra attention during the training and to the presence of individuals who may become easily bored due to their high entry level of knowledge.

1. Range. The range is the easiest measure of dispersion to calculate. When using interval or ratio data, it is simply the largest score or value minus the smallest score or value. In the prior example, the range for group A would be 79 - 51, or 28. The range for group B would be 68 - 62, or 6. The range can also be calculated for nominal and ordinal data, and in fact, is the only measure of dispersion that can be calculated for these types of data. When using nominal data, one reports the range of frequencies observed in each of the categories. For example, when reporting the nationalities of a group of participants in an international training event, it might be reported that "the largest contingent were the Guatemalans with 8 participants, and the smallest contingent were the Peruvians with 2 participants."

Usually the range is not reported as the sole measure of dispersion (unless it is for nominal or ordinal data) since it has the drawback of being heavily influenced by extreme values just as the mean is.

2. Variance. The variance is a more complicated measure of dispersion to calculate. It is a measure of how individual scores in a group vary from the mean of the group. If the scores vary little from the mean, then the variance is small. If they vary a lot from the mean, the variance is large. The variance, like the standard deviation soon to be discussed, can only be calculated for interval or ratio data. To calculate the variance, each individual score is subtracted from the mean of all the scores, the difference is squared, the squared differences for all scores are then added together. Finally, this

total is divided by the number of scores. To illustrate the process, the scores of the posttests from group A in Table 5.1 will be used to calculate the corresponding variance.

TABLE 5.2
CALCULATION OF VARIANCE

X Group A's Scores
51
58
63
67
72
79

Mean $\bar{X}$ = 65

	$\bar{X}$ - X (mean - score)	$(\bar{X} - X)^2$ (mean - score)^2
(65-51)	14	196
(etc.)	7	49
	2	4
	-2	4
	-7	49
	-14	196
Total		498

Variance = 498 ÷ 6 = 83

In this case the variance is very large since the scores were quite dispersed. Calculating the variance for the scores for group B yields a variance of only 3.3, reflecting the close grouping of the scores.

3. Standard Deviation. The standard deviation is simply the square root of the variance. In the example above, the standard deviation for group A would be the square root of 83, or 9.1. The standard deviation for group B would be 1.8. The standard deviation is the most common measure of the dispersion of data and also has many uses in other statistical calculations. The standard deviation can also reveal important information

about a group of scores or values that are normally distributed. A normal distribution is one that resembles a bell-shaped curve when represented graphically. In general terms, it refers to a distribution that has approximately equal numbers of high and low scores that are equal distances from the mean (e.g. symmetric around the mean), and whose mode is near the mean as well. Many distributions such as heights, weights, or blood pressures, are normally distributed. This cannot be assumed for all distributions of test scores, however. Test score distributions should be graphed first to see if they have an approximately normal distribution. Figure 3.1 is an example of an approximately normal distribution since it is somewhat symmetric around the mean.

Assuming that one does have an approximately normal distribution, the standard deviation indicates how many of the scores measured will most likely fall in a certain range. For any normal distribution, approximately 68 percent of the values will fall within one standard deviation of the mean. To use the example of group A in Table 5.2, if it can be assumed that the distribution is approximately normally distributed (in reality it is impossible to tell with only six data points), then one could estimate that 68 percent of all the scores would fall within 9.1 points of the mean, which was 65. In other words, 4 of the 6 scores (.68 X 6) would be expected to fall between 56 and 74 (65 ± 9.1). Reviewing the raw data, this is indeed the case.

Just as normal distributions will have approximately 68 percent of their values falling within one standard deviation of the mean, they will have approximately 95 percent of their values falling within two standard deviations of the mean, and approximately 99.6 percent of their values falling within three standard deviations of the mean.

Measures of Association

The final descriptive statistical measures that will be discussed are measures of association. Measures of association can be very important for training evaluators to consider. Anyone who has ever pondered questions like "I wonder if the high test scores

were related to ____"; or "Perhaps the most productive CBD workers are those who are ____", has already considered the subject of association. The blanks in these two examples could be variables such as age, race, sex, past experience, or anything that might appear to be related to the measure being analyzed. If a relationship exists between any two variables, then there is said to be correlation between them. The strength of that relationship is measured by a correlation coefficient. Correlation coefficients range from +1.0, indicating a perfect positive correlation (i.e. they vary in the exact same direction), to -1.0, indicating a perfect negative correlation (i.e. they vary in exactly the opposite direction). A correlation coefficient of 0 indicates no correlation or relationship at all.

While correlations can be determined mathematically, they can also be observed visually by plotting the values of any two variables (that use interval or ratio data) on a diagram. Such a diagram is called a scatter diagram. Figures 4.1-3 show three such scatter diagrams in which the two variables analyzed are CBD workers' posttest scores and the number of family planning users they reported serving six months after their training. Each square represents a particular CBD worker and the point corresponding to his or her posttest score and number of users.

FIGURE 4.1: POSITIVE CORRELATION

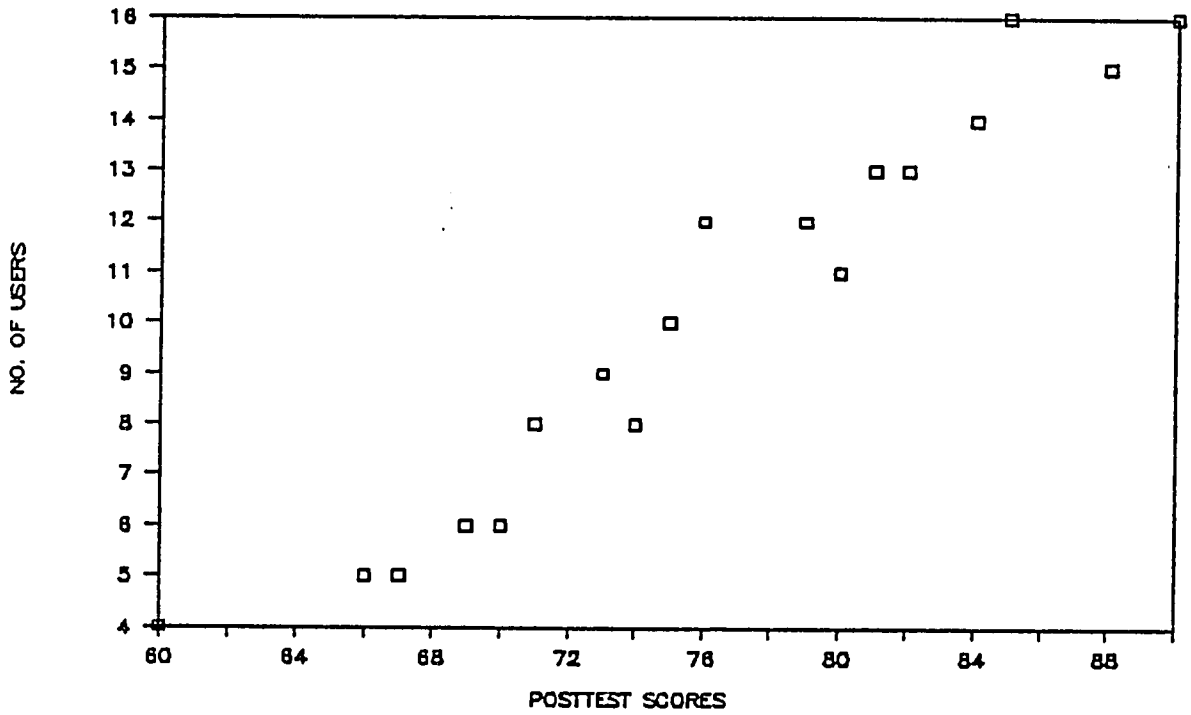


FIGURE 4.2: NEGATIVE CORRELATION

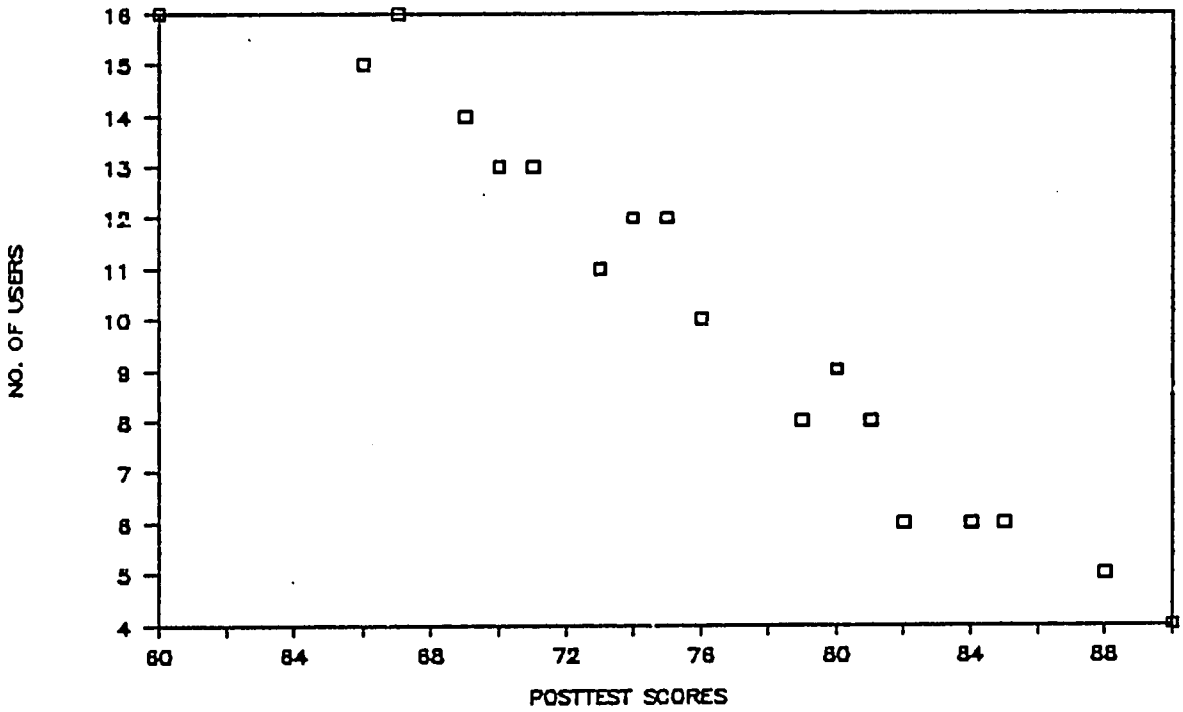


FIGURE 4.3

NO CORRELATION

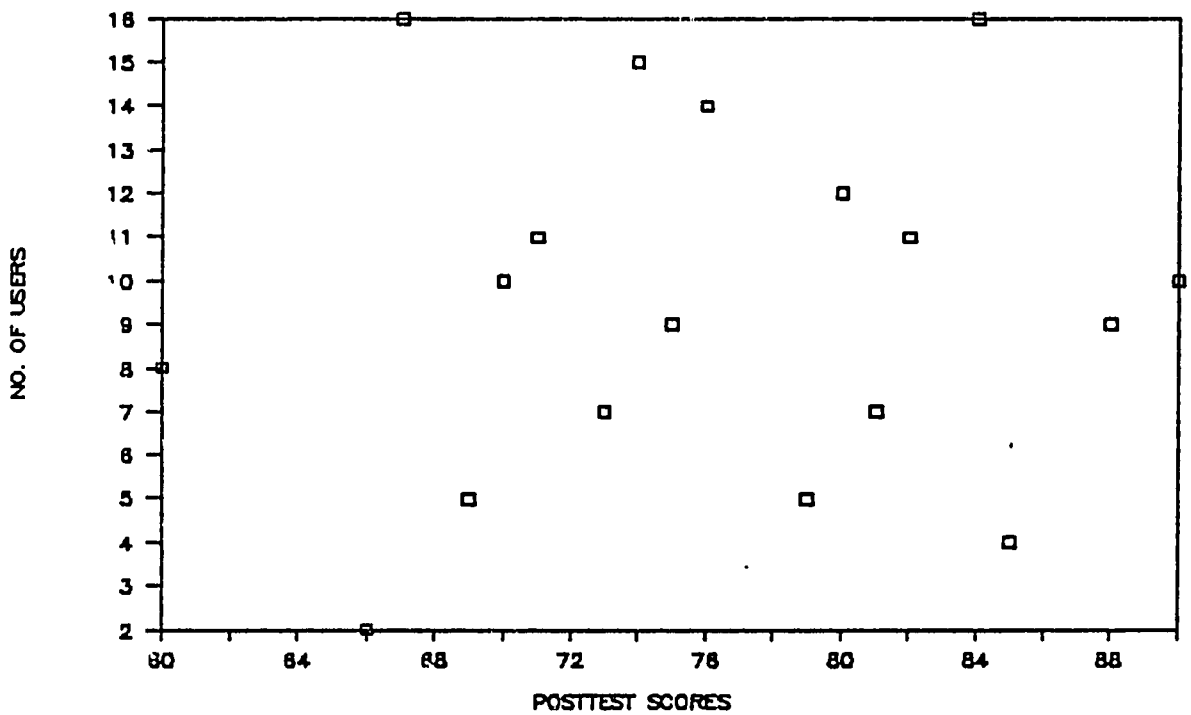


Figure 4.1 shows an example of positive correlation, where the workers who scored higher on their posttest exams also were serving more family planning users. The correlation coefficient corresponding to this set of data would be positive and close to 1.0, indicating an almost perfect positive correlation between posttest scores and the number of users served. Figure 4.2 shows the example of a negative correlation, where the workers who scored highest were the ones with the least number of users. This set of data would have a correlation coefficient close to -1.0, indicating an almost perfect negative correlation. Figure 4.3 illustrates the example of a set of data with no correlation, or a correlation coefficient close to 0. Here the test scores have no relationship to the number of users the CBD workers subsequently recruited and served.

A means by which to more precisely calculate the correlation is through computing a

correlation coefficient known as Pearson's r. Pearson's r is one of many different correlation coefficients that can be used to measure associations between variables. The Pearson's r correlation coefficient, like all correlation coefficients, varies between -1.0 and +1.0. In order to use Pearson's r, a few criteria have to be met. One is that the data be interval or ratio data, another is that the two variables have a linear relationship, i.e. resemble a straight line when plotted on a scatter diagram with no curve in the appearance of the graph. If these criteria cannot be met, then other types of correlation coefficients may be able to be used. A statistician or statistics book should be consulted in such a case.

To calculate Pearson's r, the following formula should be used:

$$r = \frac{n\,\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\{n(\Sigma x^2) - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}}}$$

where:

r = Pearson's Correlation Coefficient
n = sample size
x = values on the X axis
y = values on the Y axis

Table 6 illustrates the calculation of r using the same test score and productivity data that were used to make the scatter diagram in Figure 4.1.

TABLE 6

CALCULATION OF PEARSON'S R

(x) Posttest Score	(y) No. of Users	x^2	y^2	xy
60	4	3600	16	240
66	5	4356	25	330
67	5	4489	25	335
69	6	4761	36	414
70	6	4900	36	420
71	8	5041	64	568
73	9	5329	81	657
74	8	5476	64	592
75	10	5625	100	750
76	12	5776	144	912
79	12	6241	144	948
80	11	6400	121	880
81	13	6561	169	1053
82	13	6724	169	1066
84	14	7056	196	1176
85	16	7225	256	1360
88	15	7744	225	1320
90	16	8100	256	1440
1,370	183	105,404	2,127	14,461

n = 18

Inserting the appropriate numbers into the formula, the coefficient is

$$r = \frac{n\,\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{\{n(\Sigma x^2) - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}}} = \frac{18\,(14{,}461) - (1{,}370)(183)}{\sqrt{\{18(105{,}404) - (1{,}370)^2\}\{18(2{,}127) - (183)^2\}}}$$

$$= \frac{9588}{\sqrt{(20{,}372)(4{,}797)}} = .97$$

This coefficient, representing an almost perfect positive correlation between the posttest scores and the number of users the workers subsequently served, confirms what was observed in the scatter diagram of Figure 4.1. Whenever a correlation coefficient is between .6 and .8, one can conclude that a high degree of positive correlation exists

between two variables. A correlation coefficient of .8 or greater indicates a very high degree of positive correlation. Conversely, a coefficient of between -.6 and -.8 indicates a high degree of negative correlation, and a coefficient between -.8 and -1.0 indicates a very high degree of negative correlation.

A final point should be made about correlation. Correlation should not be confused with causality. Just because two variables are correlated or associated with each other does not mean that one causes the other. There can always be intervening variables or outside factors that influence the outcomes of the two variables and make it appear that one causes the other. In the example used above, it would be inappropriate to conclude that high posttest scores cause CBD workers to recruit and serve more family planning users. There could be a third factor--perhaps self confidence--that is the real cause for people scoring high on tests and having high work productivity. Perhaps people with a high sense of self confidence score higher on tests in general and also have higher work productivity. This would make it appear that the high test scores were causing the high work productivity when in reality it was a high sense of self confidence that was causing both to occur.

INFERENTIAL STATISTICS

Whenever training evaluators want to compare two different groups of trainees, or two different measures on the same group of trainees, inferential statistics will have an important role to play. Virtually every training event collects pre- and posttest scores. These scores are often included in the evaluation report along with statements such as "As can be seen from the differences in the pre- and posttest scores, the trainees learned a great deal in the course." But how does one know whether a "great deal" was learned or not? If the mean score of the pretests was 60, on a 100 point exam, and the mean score of the posttests was 95, it is fairly obvious that the participants learned a great deal, or, to put it another way, there was a significant difference between the two scores. However, if the mean posttest score was 65, would that also be a significant difference?

Perhaps the increase was only due to chance, since scores will always tend to fluctuate from one measurement to another even when no change in learning has occurred.

Inferential statistics help to answer these and other questions related to the difference between two groups or the difference between two scores of the same group. Inferential statistics employ a variety of tests that can be used to determine the amount of observed change between two groups of scores and give the probability that a real, or significant, change occurred between the two groups. If the calculated probability is less than 5 percent that the observed change was due to chance, then it is generally accepted that there was a significant change.

Probably the two most useful tests for training evaluators to use for determining whether differences are significant are the Students t-Test and the Chi-Square Test. Deciding which of the two tests to use requires that one first answer the following questions:

1. Are the scores or values interval or ratio data? If the data are nominal or ordinal, then the chi-square test will have to be used. The t-test can be used if the data are interval or ratio.

2. Are the groups being compared independent? If a pre- and post-measurement are made on the same group of people, then the two sets of scores are dependent. If, however, scores are obtained from two different groups of people and they were not matched on some characteristic, such as age or sex, then the two sets of scores are independent. The t-test can be used with both dependent and independent groups, but it is necessary to have independent groups in order to perform the chi-square test.

3. Are the data from the two groups approximately normally distributed? (Refer to the discussion on standard deviation for an explanation of a normal distribution.) If the data have an approximately normal distribution then a t-test can be performed. A chi-square test, on the other hand, does not require a normal distribution in the data.

The following is an example of when a t-test is appropriate to use and how the t-value is calculated:

Case #1: t-Test, Dependent Groups

Suppose that you are responsible for the training department of your organization and have been handed a course evaluation written by the trainer who conducted the course. The course was given to 20 participants and the pre-and posttest scores for each participant are included in the report, together with the mean score for the group and the percentage change between the mean pretest score and the mean posttest scores. Table 6.A presents the data given to you.

TABLE 7.1

SCORES OF PRETESTS AND POSTTESTS

Student	Pretest	Posttest
1	72	87
2	86	100
3	72	90
4	65	79
5	32	82
6	54	90
7	81	92
8	68	87
9	66	92
10	44	98
11	74	100
12	49	96
13	37	92
14	66	84
15	66	82
16	82	100
17	82	98
18	38	88
19	74	90
20	74	93
	1,282	1,820

Pretest Mean: 1,282 ÷ 20 = 64.1 Posttest Mean: 1,820 ÷ 20 = 91
Percentage increase of posttest over pretest: [(91÷64.1)(100)]-100 = 42%

With this data you know the mean posttest score was 27 points higher than the mean pretest score, a 42 percent increase. It appears obvious that there is a significant difference between the posttest scores and the pretests scores and, therefore, you conclude that a change in the participants' knowledge did take place as a result of training. However, to be on the safe side you decide to use a statistical test to be absolutely certain that the higher posttest scores didn't just happen by chance. Since the scores are interval data and are dependent (being pre- and posttests on the same group of people) the t-test would seem to be the appropriate test to use. One requirement of the t-test is that the data be approximately normally distributed. When graphing the data above, however, you find that they only somewhat resemble a bell-shaped curve. Nevertheless, they are close enough to normality that statisticians will allow the t-test to be used (in reality, the t-test works fairly well even when data do not resemble a normal distribution; there is a great amount of leeway in what is considered to be a "normal distribution").

Having decided that you will perform the t-test you then proceed according to the following formula:

$$t = \frac{\bar{D}}{\sqrt{\dfrac{\Sigma D^2 - \dfrac{(\Sigma D)^2}{n}}{n(n-1)}}}$$

where:

D is the difference between the pretests and posttests
$\bar{D}$ is the mean of the differences (D's)
ΣD is the sum of the differences
ΣD^2 is the sum of the squared differences, and
n is the number of participants (or pairs of scores).

Table 7.2 again presents the pretest and posttest scores given to you along with the calculations that you should perform to arrive at the t-value.

TABLE 7.2

CALCULATION OF T-VALUE

PRETEST SCORES X	POSTTEST SCORES Y	D (Y-X)	D^2
32	82	50	2500
37	92	55	3025
38	88	50	2500
44	98	54	2916
49	96	47	2209
54	90	36	1296
65	79	14	196
66	84	18	324
66	82	16	256
66	92	26	676
68	87	19	361
72	90	18	324
72	87	15	225
74	90	16	256
74	93	19	361
74	100	26	676
81	92	11	121
82	98	16	256
82	100	18	324
86	100	14	196
1,282	1,820	538	18,998

$\bar{D} = 538 \div 20 = 26.9$

$\Sigma D = 538$

$\Sigma D^2 = 18{,}998$

$n = 20$

$$t = \frac{\bar{D}}{\sqrt{\dfrac{\Sigma D^2 - \dfrac{(\Sigma D)^2}{n}}{n(n-1)}}} = \frac{26.9}{\sqrt{\dfrac{18{,}998 - \dfrac{(538)^2}{20}}{20(19)}}} = \frac{26.9}{3.45} = 7.80$$

Having obtained a t-value of 7.80, you then consult a table of t-values to determine whether the t-value you obtained indicated that the difference between the two sets of scores was statistically significant, or could have been expected from chance alone. Table 7.3 is a partial table of t-values. The left column marked "d.f." refers to the number of degrees of freedom in the calculation and is equal to n-1 for dependent groups as in the current example. The two right-hand columns are the t-values corresponding to the .95 and the .99 confidence levels. These refer to the probability that the difference between the two groups was due to chance alone. With the .95 level of confidence there is a 5 percent chance that the difference was due to chance alone. With the .99 level of confidence there is only a 1 percent chance that the difference was due to chance alone. Both of these levels are commonly used by evaluators and the use of one or the other depends on how sure you want to be that the difference you observed is not due to chance.

Returning to the t-value of 7.80 that you obtained and reading across from row 19 (d.f. 20-1) in Table 7.3, you note that the t-values associated with the .95 and .99 levels of significance are 1.729 and 2.539, respectively. Since 7.80 is larger than either of these numbers you can be assured that the difference between the pre- and posttests was not due to chance alone. If the t value you had computed were less than 1.729, you would have had to conclude that chance alone could have produced the difference in test scores at the .95 level (i.e. there would be at least a 5 percent probability that the difference was due to chance). If the t-value had been between 1.729 and 2.539 then you could have concluded that the difference was statistically significant at the .95 level, but not at the .99 level (i.e. you would be willing to accept a probability of between 1 and 5 percent that the difference was due to chance).

TABLE 7.3: TABLE OF T-VALUES[7]

d.f.	.95	.99
1	6.314	31.821
2	2.920	6.965
3	2.353	4.541
4	2.132	3.747
5	2.015	3.365
6	1.943	3.143
7	1.895	2.998
8	1.860	2.896
9	1.833	2.821
10	1.812	2.764
11	1.796	2.718
12	1.782	2.681
13	1.771	2.650
14	1.761	2.624
15	1.753	2.602
16	1.746	2.583
17	1.740	2.567
18	1.734	2.552
19	1.729	2.539
20	1.725	2.528
21	1.721	2.518
22	1.717	2.508
23	1.714	2.500
24	1.711	2.492
25	1.708	2.485
26	1.706	2.479
27	1.703	2.473
28	1.701	2.467
29	1.699	2.462
30	1.697	2.457
35	1.690	2.438
40	1.684	2.423
45	1.679	2.412
50	1.676	2.403
60	1.671	2.390
70	1.667	2.381
80	1.664	2.374
90	1.662	2.368
100	1.660	2.364
120	1.657	2.351
200	1.652	2.345
500	1.648	2.334
infinity	1.645	2.326

7. This table adapted from Table A.4 of Olive J. Dunn, Basic Statistics: A Primer for the Biomedical Sciences, 2nd edition, Wiley:New York, 1977.

Case #2: t-Test, Independent Groups

In this example, suppose that you have developed a new type of training that is supposed to be superior for motivating CBD workers to promote and distribute family planning methods in their communities. As the evaluator responsible for determining whether this new type of training is effective in motivating CBD workers to be more productive, you decide to test the "new" training by providing it to half of the CBD workers scheduled to receive training (group 1). The other half (group 2) of the CBD workers are given the standard training. If at all possible, you want to randomly distribute these CBD workers into either the new or standard training. If this is impossible, you may want to match groups on some characteristic and use a dependent group t-test.

Having given the new training to 20 CBD workers in group 1 and the standard training to 15 CBD workers in group 2, you then perform a followup evaluation six months later to determine the number of family planning users each is currently serving. The results you obtain from reviewing their service delivery statistics are found in Table 8. Since the results are interval data, are approximately normally distributed, and pertain to two independent groups, you decide to test the significance of the difference between the mean number of users served by the two groups by using the t-test for independent samples. The formula for the t-test for independent samples is the following:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{\Sigma(X-\bar{X})^2 + \Sigma(Y-\bar{Y})^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where:

$\bar{X}$ is the mean of scores in group 1

$\bar{Y}$ is the mean of scores in group 2

$\Sigma(X-\bar{X})^2$ is the sum of the squared differences of scores from the mean in group 1

$\Sigma(Y-\bar{Y})^2$ is the sum of the squared differences of scores from the mean in group 2

n_1 is the number of scores in group 1

n_2 is the number of scores in group 2

TABLE 8: FOLLOWUP TRAINING EVALUATION RESULTS AND CALCULATION OF T-VALUE

GROUP 1 # OF USERS X	GROUP 2 # OF USERS Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})2$	$(Y-\bar{Y})2$
23	15	5	1	25	1
15	12	-3	-2	9	4
20	18	2	4	4	16
18	10	0	-4	0	16
24	19	6	5	36	25
19	14	1	0	1	0
17	13	-1	-1	1	1
16	17	-2	3	4	9
17	12	-1	-2	1	4
19	9	1	-5	1	25
13	38	1	24	1	576
14	15	-4	1	16	1
12		-6		36	
20		2		4	
18		0		0	
20		2		4	
17		-1		1	
19		1		1	
18		0		0	
15		-3		9	
360	192			154	678

$\bar{X} = 18 \qquad \bar{Y} = 16$
$n_1 = 20 \qquad n_2 = 12$

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\Sigma(X-\bar{X})^2 + \Sigma(Y-\bar{Y})^2}{n_1+n_2-2} \; \frac{1}{n_1} + \frac{1}{n_2}}} = \frac{18 - 16}{\sqrt{\frac{154 + 678}{32 - 2} \; \frac{1}{20} + \frac{1}{12}}} = 1.04$$

The degrees of freedom for the t-test for independent groups is $n_1 + n_2 - 2$, which equals 30 in this example. Referring again to Table 7.3, you notice that the t-values associated with the .95 and .99 levels of confidence with 30 degrees of freedom are 1.697 and 2.457, respectively. Since the t-value you calculated is 1.04, and less than either of these numbers, you must conclude that there is no statistically significant difference (at the .05 level) between the performance of the two groups and, therefore, the greater number of users served by group 1 could have been due to chance alone.

Case #3: Chi-Square Test

To illustrate the use of the chi-square test, suppose you wish to evaluate the effect of providing refresher training in sterilization counseling to clinic staff who will counsel clients interested in obtaining a sterilization procedure. You have long suspected that the high rate of clients cancelling or not showing up for their sterilization procedures after their initial orientation encounter has been due, in part, to the failure of the counselors to fully explain the nature of the operation and correct some of the misconceptions surrounding the operation. Therefore, you devise an experiment to test the effect of providing new "client-oriented" training on the subsequent rate of return visits for sterilization made by clients after their initial interview. You randomly select half of the counselors to receive the refresher training (experimental group) and then observe the behavior of the clients counseled by both the experimental group and control group. Table 9.1 lists the number of return visits for sterilization made by clients pertaining to each of these two groups.

TABLE 9.1: RATE OF RETURN VISITS MADE BY TYPE OF COUNSELOR GROUP

	No. of Return Visits Made	No. of No-Shows, Cancellations	Total
Experimental Group (given training)	25	12	37
Control Group (no training)	19	14	33
	44	26	70

With this information you have a pretty good idea that the training was successful in that a higher percentage of clients counseled by the experimental group made return visits for sterilization than those counseled by the control group (68 percent versus 58 percent). However, you wish to eliminate the possible conclusion that the differences could be due to chance alone. Therefore, you decide to perform a chi-square test of significance. You perform the chi-square test because the groups are independent and because the data are nominal in nature. Although the results are represented with numbers, the data are nominal because they represent proportions rather than actual values such as test scores. In other words, each individual client was not given a "score", but rather was classified as "return visit" or "no-show".

You next proceed to calculate the chi-square value according to the following formula:

$$X^2 = \Sigma \frac{\{f(o) - f(e)\}^2}{f(e)}$$

where:

X^2 is the chi-square value
f(o) is the observed frequency of responses
f(e) is the expected frequency of responses

The expected frequency of responses is the frequency that would have been theoretically expected, if there had been no differences between the groups being compared. The next step you take is to calculate the expected frequencies for each cell assuming there is no difference between the two groups. These expected frequencies are placed in parentheses.

TABLE 9.2: RATE OF RETURN VISITS MADE BY TYPE OF COUNSELOR GROUP WITH EXPECTED FREQUENCIES

	No. of Return Visits Made	No. of No-Shows, Cancellations	Total
Experimental Group (given training)	25 (23)	12 (14)	37
Control Group (no training)	19 (21)	14 (12)	33
	44	26	70

These expected frequencies were calculated by multiplying the percentage of the total corresponding to each group by the totals listed in the two columns for return visits and no-shows. For example, the expected frequency for return visits for the experimental group was calculated as follows:

Proportion of experimental group visits of all visits = 37 ÷ 70 = .53
Proportion of return visits expected for experimental group = (.53)(44) = 23

Having calculated the expected frequencies, you now use the formula above to calculate the chi-square value:

$$X^2 = \frac{(25-23)^2}{23} + \frac{(12-14)^2}{14} + \frac{(19-21)^2}{21} + \frac{(14-12)^2}{12} = .17 + .29 + .19 + .33 = .98$$

The number of degrees of freedom for the chi-square test is found by multiplying the number of columns less one by the number of rows less one. In this case the degrees of freedom would be (2-1)(2-1) = 1.

Having calculated the chi-square value and degrees of freedom, you then consult a table of chi-square values to determine whether the difference between the two groups is statistically significant. Table 9.3 is a partial listing of chi-square values. Reading across from the row corresponding to 1 d.f., you find that the values of 3.84 and 6.63 corresponding to the .95 and .99 levels of confidence, respectively. Since the chi-square value you calculated is less than these you conclude that chance alone could have caused the differences in the results observed between the experimental and control groups. Therefore, you cannot state that the training achieved its desired effect.

One final point should be made concerning all three of the statistical tests discussed in this section. Whenever a t-value or chi-square value is less than the .95 level of confidence, most statisticians claim that the differences observed in the groups under comparison could have been due to chance and, therefore, no statistically significant differences are said to exist. However, as previously discussed, the .95 level of confidence corresponds to a probability of 1 in 20 that the results were due to chance. Just because a t-value or chi-square value is less than the .95 level of confidence does not mean that the group's differences actually <u>were</u> due to chance. It only means that the differences were not sufficiently pronounced so as to be sure (over 95 percent sure) that chance alone did not cause the differences. If a particular test fails to show a significant difference at the .95 level, it may be because the sample size used was too small. As larger samples are more likely to show evidence of significant differences when they actually exist, evaluators may wish to do a restudy with a larger group if the results of one study fail to detect significant differences.

TABLE 9.3

TABLE OF CHI-SQUARE VALUES[8]

d.f.	.95	.99
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.73
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.00
18	28.87	34.81
20	31.41	37.57
24	36.42	42.98
30	43.77	50.89
40	55.76	63.69
60	79.08	88.38
120	146.57	158.95

8. This table is adapted from Table A.5 of Dunn, Ibid.

CHAPTER V: SPECIAL ISSUES IN TRAINING EVALUATION

EVALUATION OF QUALITY OF SERVICES PROVIDED

Training evaluation is often concerned with numerical outcomes: how many trainees were trained?; how many clients did the trainees subsequently serve?; how have the organization's numerical goals been furthered as a result of training? These are legitimate concerns that should be attempted to be answered as a result of training evaluations. But another aspect of the outcomes of training should also be uppermost in evaluator's minds and that is the evaluation of the quality of services provided by trainees.

In section II it was noted that the outcomes of training with respect to job performance include not only the quantity of activities that the trainee performs on the job, but the quality of that work as well. Likewise, outcomes of training with respect to short-term effects on the target population include not just indicators relating to an increase in the population served, but an increase in the number of satisfied consumers/clients as well.

This focus on quality of services is often overlooked in training evaluation but is just as important as evaluating training's effect on the quantity of services delivered. One reason why training's effect on quality of services is often not addressed is because quality is a hard concept to define. It is easy to know whether a trainee has provided a service to a particular person or not, but how does one know whether that service was of good quality or poor quality or somewhere in between? There is no easy answer of course. Evaluating the quality of services requires that the organization first define for itself what it considers a quality service to be.

Analysis of Institutional Norms

In the area of training CBD workers, for example, an organization might develop a list of norms that indicates what a CBD worker should do and not do when providing a contraceptive method to a new user. The list of norms might include items such as the following:

Before providing a new user with a contraceptive method, the CBD worker shall perform the following:

- describe all methods that the program offers, including their uses, advantages and disadvantages;
- verify the absence of any contraindications in the potential user after a particular method has been selected;
- in the absence of contraindications, give the user a one month's supply of the method and make an appointment for a followup visit in one month;
- if case of any doubts concerning the possible existence of a contraindication, refer the potential user for a medical exam before giving any oral contraceptives;
- complete the client record form and supply distribution form.

These institutional norms provide guidelines for defining whether a CBD worker has provided a quality service or not when he or she distributes a method to someone. However, institutional norms are at times overly general to serve as guides to defining the quality of a service. For example, the first norm states that the CBD worker will describe all methods that the program offers, including their uses, advantages and disadvantages. Suppose that a worker gives a presentation to a prospective user and

mentions all the methods offered by the program, including their uses, advantages and disadvantages, but forgets to mention one minor point. Perhaps the CBD worker forgot to mention that a pill user should take the pill at the same time every day to avoid forgetting to take it. If the CBD worker remembered the other instructions about taking the pill, and its advantages and disadvantages, should it matter that the worker forgot to mention one minor point, that if ommitted, would not necessarily endanger the user? Probably not. This points to the importance of specifying not only institutional norms, but the minimally acceptable performance standards for each institutional norm.

Specification of Performance Standards

Minimally acceptable performance standards outline what a worker has to do as a minimum in order for the service to be considered of a quality acceptable to the institution. Such standards do not indicate what one hopes a worker will perform (hopefully workers will exceed the standards), they only indicate what a worker should do for his or her work to be considered acceptable. In the example used above, minimally acceptable performance standards for norm #1 could include the following with respect to oral contraceptives:

As a minimum the distributor will:

- explain how the pill is used (taken orally every day)
- explain that the pill is a highly effective form of contraception
- explain that the pill is reversible
- explain that decreased or absent menstrual flow occurs as a result of taking the pill
- explain that the pill is not 100 percent effective in preventing pregnancy
- explain that there can be reduced milk flow when taking the pill (if the woman is currently breastfeeding)
- explain that there is a slight risk of complications when taking the pill which include ...
- explain that there are normal side effects associated with taking the pill which include...

These eight items of information are hypothetical minimal performance standards that could be used be a CBD program wishing to evaluate whether one of its norms were being fulfilled or not, and, in so doing, whether its CBD workers were providing a quality service. These items apply only to one component of one norm (description of oral contraceptives to a new user). It is a lengthy process to establish minimal performance standards for each institutional norm and each task of each worker but it is worth it in the long run so as to be able to evaluate systematically and uniformly whether quality services are being provided by the institution or not.

The foregoing discussion has relevance to the topic of training evaluation in that it is often of interest, or should be of interest, for training evaluators to look at the effects of training on the quality of services given by trainees after their training. If this is to be done objectively and uniformly, minimal performance standards need to be developed in order for an evaluator to rate the quality of service that an individual provides, both before and after training. If the outcome indicator of training is a vaguely perceived notion of "quality", without specific observable measures of quality specified, then the evaluation will most likely be void of any validity and reliability.

Minimally acceptable standards are useful not only for evaluating the quality of workers job performance, but for evaluating the accomplishment of learning objectives as well. A performance test in the classroom might require that a participant perform a role-play in which the "promoter" has to explain to a potential user the location, costs, and types of clinical family planning services offered by a particular agency. Minimally acceptable standards might include that the promoter mention the following:

- mention the name, address, and phone number of the closest clinic to the user;
- mention how to reach the closest clinic to the user by public transportation;
- mention the hours of the clinic;
- mention that the IUD, diaphragm, and sterilization are all offered at the clinic;
- mention that the clinic's services are not very expensive

While this might be the minimal amount of information that the promoter has to mention to pass the role-play exercise, it would be desirable if he or she mentioned other aspects of the clinic's services as well, such as the exact prices of various services and some information about the methods offered at the clinic. However, it would not be a cause for failing the participant if he or she forgot to mention these items.

Figure 5 is a summary outline of how minimal performance standards can be used to evaluate the accomplishment of learning objectives and on-the-job performance related to a particular task of a CBD worker: verifying the absence of contraindications to the pill before giving the pill to a new user. Note how the outline illustrates the knowledge, skills, and attitudes that the participant should perform both during the training and on the job. Developing an outline similar to Figure 5 for each key task of a given worker before training occurs would be an invaluable way to help assure that trainers and supervisors focus on evaluating the same indicators that measure whether quality services will be provided by the program or not.

However, to perform this type of integrated evaluation of training in the classroom and on the job requires close communication and coordination between trainers and supervisors as was mentioned in section II concerning the feedback of training

FIGURE 5[9]

LEARNING AND PERFORMANCE OBJECTIVES

Key Task: Verify the absence of contraindications before providing the pill to a new user.

	During Training		On-the-Job Performance	
	When	How	When	How
Demonstrates Level of Knowledge	During the posttest...	Identifies all contraindications according to program standards, without errors or omissions.	Supervisor asks...	Names all contraindications according to program standards, without errors or omissions.
	Provides a list of medical terms for all contraindications.	Writes down correct local terms for each medical term.	Supervisor names the medical terms for all contraindications...	Names one or two local terms for each medical term.
Demonstrates Skills	Interviews three fellow trainees who have been assigned different contraindications...	Correctly detects the contraindication in each case.	Supervisor involves trainee in a simulated interview...	Correctly detects the contraindications selected by the supervisor for the test.
			Supervisor involves trainee in simulated inverview...	Correctly distinguishes common discomforts from those that signal contraindications.
			Interviews performed with women referred to clinic with suspected contraindications...	Women report they made the clinical visit.
Demonstrates Positive Attitudes	During training...	Takes time to learn about contraindications despite failures in other subjects.	Supervisor or evaluator visits users...	Users report distributors interviewed them on health problems prior to giving them the pill.
	During training...	Asks questions to clarify doubts on contraindications.	Supervisor or evaluator visits users...	No pill users with contraindications are found.
			Supervisor reviews records...	Distributor identifies and asks about cases that were referred on suspicion of contraindications in order to clarify his or her doubts.

9. Taken from the original Spanish version in Terborgh, A., et al, 1987. Capacitación y Control de Calidad en Programas Comunitarios de Planificación Familiar, Development associates: Arlington, VA (unpublished).

evaluation information. An organization's managers have a special role to play in helping facilitate this communication and in making sure that trainers, supervisors, and evaluators all know what the norms of the institution are, what the tasks of each worker are, and what the minimally acceptable performance criteria for each task are. In this way, trainers, supervisors, and evaluators can all orient their focus towards helping trainees achieve, and hopefully surpass, this minimal level of performance, both in the classroom and on the job.

One other aspect of Figure 5 that should be noted is how the quality of services was evaluated on the job. To evaluate the worker's knowledge, questions were posed by the supervisor in order to test whether the worker knew the contraindications or not. To evaluate the worker's skills or abilities, the supervisor staged a simulated interview or role-play that tested the worker's ability to detect contraindications. Likewise, various women in the community were interviewed who were referred to a clinic to determine whether they actually went to the clinic (a measure of how persuasive the CBD worker was in referring them).

To evaluate the CBD worker's attitudes towards screening for contraindications, and to find out whether the worker is actually performing this task or not, the supervisor interviewed a group of pill users in the community and asked whether they had been screened for contraindications or not before receiving the pill. The supervisor also screened the users, as part of the interview, for contraindications to determine whether the CBD worker had correctly screened them or not. Finally, the supervisor looked for indications from the CBD worker (such as their asking the supervisor about women they had referred to a clinic because of the presence of contraindications) that he or she was truly concerned about screening for contraindications and understood the importance of the task.

This outline provides an example of the various means that can be employed to evaluate the quality of services given by ex-trainees. Knowledge retention tests, role-plays during

supervisory visits, a review of a worker's records, and interviews with program beneficiaries are all useful means for obtaining information to evaluate the quality of services being given. Interviews with a sample of users is a particularly good means of obtaining this kind of information that is often underutilized by training evaluators. The appendix contains an example (#7) of an interview guide used to interview family planning users concerning the quality of service they received.

One other measure of the quality of services provided that was not mentioned in Figure 5 is continuation rates. Measuring how long women continue to use a particular method is a proxy measure for the quality of the program since women often quit using a method when they have complaints about the program's personnel, or experience side-effects that the doctor, nurse, or CBD worker perhaps did not warn them about. Logistical problems that result with a woman being unable to receive new supplies of her method from the program can also cause the woman to give up on family planning or switch to another source. The only problem with using continuation rates as the absolute criterion for measuring a program's quality of services, however, is that women can also cease to remain active users due to reasons entirely unrelated to the program and its personnel.

Determining whether screening for contraindications correctly took place or not is just one example of a service quality issue that can be evaluated in a followup evaluation. Evaluators might be equally concerned about evaluating whether users made an informed choice or not when they "chose" a particular method. Preserving client confidentiality might be another service quality issue that evaluators would want to look at in terms of how ex-trainees are performing and whether they need more training in this area. Ultimately, whatever an institution defines as its norms should be used to define the focus of its evaluation of quality of services.

COST EFFECTIVENESS ANALYSIS

Cost-effectiveness analysis (CEA) is a particularly important decision-making tool for evaluators. Within the context of training evaluation, CEA is a technique for identifying the most effective use of limited training resources. CEA is one of two principal efficiency measures, i.e. measures that analyze the outputs of a given program or training activity in relation to the level of inputs needed to achieve them. The other principal efficiency measure is cost benefit analysis (CBA).

Cost benefit analysis is a tool for measuring the net benefits of program alternatives, even when the alternatives are not aimed at common goals. CBA involves calculating the benefits of a program or activity and the costs of that program or activity, both expressed in monetary terms. The benefits and costs can either by presented as a benefit cost ratio (benefits divided by costs) or as net benefits (benefits minus costs). Expressed in terms of net benefits, CBA can help managers decide between alternative programs by choosing the one that will yield the highest net benefits to the organization. Despite the value of CBA, it is not frequently used in the training evaluation field because of the difficulties in calculating the benefits of training in monetary terms. Calculating the monetary worth of training a nurse-midwife, for example, is a determination that cannot easily be made.

Cost-effectiveness analysis, on the other hand, is much easier to perform and can be used in most training evaluation situations where outcomes can be quantified and where the costs of training can be accurately calculated. CEA is a ratio measure in which the numerator is the cost of a given outcome and the denominator is the outcome in question. While the numerator is expressed in monetary terms, the denominator is not, avoiding the difficulty inherent in cost benefit analyses of trying to place a monetary value on a given outcome.

One purpose of CEA is to help program managers, trainers, and evaluators determine which type of program or training activity accomplishes a given objective or outcome, with the least cost. This is done by keeping the level and type of outcomes constant between two or more programs and comparing the cost differences between the programs. In this way the least expensive approach to obtain a given outcome can be determined. Alternatively, CEA can be used to help determine how much of a given outcome can be obtained by two or more programs or training events for a given cost. In this case, the cost is made constant and the level of outcome is measured to determine which program or type of training is the most effective.

CEA can be performed prospectively, during the needs assessment stage, for the purpose of deciding which course of action to pursue among various alternatives. Likewise, CEA can be performed retrospectively, during the outcome stage of evaluation, for the purpose of evaluating a course of action already taken in order to determine whether it should continue to be carried out in the future. When done retrospectively, CEA is of course more precise, since actual costs and outcomes can be measured. However, it is often more valuable to perform CEA <u>before</u> engaging in a particular type of activity, especially if it requires substantial investments, in order to help the decision-maker determine which among various possible alternatives to the same end should be pursued.

CEA can be performed for any type of outcome sought in a particular training program, including outcomes from all four sub-levels of outcome evaluation: development of knowledge, attitudes and skills, job performance, short-term effects, and long-term impacts. Examples of a CEA using each of these sub-levels include the following:

Development of KAS:	Cost per trainee passing course Cost per 20 point increase in exam scores Cost per trainee learning IUD insertion
Job Performance:	Cost per home visit made by ex-trainees Cost per IEC activity performed by ex-trainees Cost per referral made by ex-trainees
Short-term Effects:	Cost per new user served by ex-trainees Cost per C.Y.P. provided by ex-trainees Cost per "knowledgeable" user served by ex-trainees
Long-term Impacts:	Cost per birth averted Cost per death averted Cost per 1 percentage point decrease in birth rate

The calculation of CEA for the short-term effects and long-term impacts listed above would undoubtedly be a bit artificial when applied to the training function only. Normally these outcomes are compared to the costs of an entire program that supports them, rather than just one of the inputs such as training. Nevertheless, it would be theoretically possible to calculate the CEA for outcomes such as couple-years-protection (C.Y.P.) if it could be determined that training was the only variable that changed when analyzing alternative means of increasing C.Y.P.

Probably the most frequent use of CEA in training evaluation is the calculation of the cost of training per trainee. This involves the most elementary and basic outcome of a training event--the presence of a "trained" person who represents human capital. While the resultant trainee may in fact produce nothing for the agency sponsoring him or her, the trainee at least represents the potential for providing services on behalf of the agency. The purpose of calculating the cost per trainee for specific training events is usually to compare the cost-effectiveness of alternative types of training. Using the example of assessing the cost effectiveness of alternative training types, a hypothetical

CEA will be outlined in the following case study.

CEA Case Study: Alternative Types of Training

Suppose you are the director of training of a large family planning organization that has recently suffered budget reductions and you are told by your executive director that the training department must streamline its budget without reducing the number of training events or trainees previously planned to occur during the year in progress. It would also be unacceptable to the organization to reduce the level of minimal competencies required of each trainee. Consequently, you cannot accept less qualified trainees in order to reduce training costs. Your task is, therefore, to find the least expensive way to train the same number of people, with the same minimal performance criteria applied, with less resources. Since the majority of people trained by your organization are CBD workers, you decide to focus on that level of trainee first. If it is not readily evident to you how you can accomplish this goal of saving money without reducing the number, or quality, of people trained, then you should do some prospective cost-effectiveness analyses comparing various means of providing training.

The steps involved in conducting a CEA are the following:

1. Identify the problem that needs to be addressed by a CEA.
2. Identify alternative means of providing the outcome to be analyzed.
3. Determine the net costs of delivering the outcomes for the various alternatives chosen.
4. Determine the net outcomes provided by the program or training.
5. Determine the cost-effectiveness ratio.
6. Perform a sensitivity analysis.

Problem Identification

In the case study example, the problem that you have been told to address is that the costs of training must be reduced without reducing the number of people trained, nor their level of minimal competencies with respect to the knowledge, attitudes and skills needed for their work. Therefore, you must find a way to produce the same outcomes (well trained people) more efficiently.

Given the nature of the problem, you know that your CEA will be focusing on reducing the numerator of the equation: the costs. The denominator, type of outcome (i.e. number and quality of people trained), should be left constant in this case so as to focus on reducing the costs for a given number of trainees. If the problem had been posed differently, perhaps in terms of the need to improve training while not increasing costs, then the denominator would be manipulated while keeping the numerator (costs) constant.

Occasionally one encounters the unenviable task of trying to reduce costs while at the same time increase the quantity or quality of people trained. If this were the charge given to an evaluator performing a CEA, he or she should really perform two cost effectiveness analyses: the first to determine the least expensive means of producing a given outcome, and the second to determine the level of outcome that could be achieved by various means using an equal amount of money. If the results of the two CEA's agreed in terms of the best alternative, then the correct decision to make would be made easier. Occasionally, however, the two CEA's will disagree in their conclusions. This might occur when certain alternatives appear cost-effective at low levels of output but are relatively less cost-effective at high levels of output due to reduced economies of scale.

Identification of Alternatives

The next step in CEA is identification of various alternatives that could be undertaken to solve the problem identified. In the case study example, you identify all of the possible variables that could potentially affect the cost of training. These are the following:

- training methodology
- training content
- training location
- training materials
- trainer/trainee ratio
- duration of training
- type of trainers
- medium of training (course, manual, video, etc.)

Having determined all the possible influences on cost for training CBD workers, you decide that the three areas most likely to result in cost savings, without greatly affecting the competency levels of the trainees, are the duration of training, the trainer/trainee ratio, and the training location. Although it might be useful to consider alternative mediums for providing the training other than the traditional courses you have been offering, you decide against this alternative for the time being since you lack manuals, videos, or other resources that could possibly substitute for courses.

You next decide how many alternative choices you wish to consider under each of these variables. You decide that you will consider two alternatives for each variable, resulting in a total of eight possible training mixes. The alternatives you decide on are the following:

Duration of training: 5 days vs. 3 days
Trainer/trainee ratio: 1:10 vs. 1:20
Training Location: Centralized vs. decentralized

The eight discreet alternatives you will consider, therefore, are

Duration	Ratio	Location
1. 5 days	1:10	Centralized*
2. 5 days	1:10	Decentralized**
3. 5 days	1:20	Centralized
4. 5 days	1:20	Decentralized
5. 3 days	1:10	Centralized
6. 3 days	1:10	Decentralized
7. 3 days	1:20	Centralized
8. 3 days	1:20	Decentralized

* performing the training in the provincial capital
** performing the training in local communities where CBD workers reside

One point that should be made is that the more alternatives you select for your CEA, the larger your sample size will have to be so that you have at least 20 observations or scores for each alternative.

Determination of Net Costs

Following the elaboration of the various alternative types of training you will consider, you next calculate how much each alternative is likely to cost. It is desirable to calculate as precisely as possible the total costs of training that will likely be incurred for each alternative, including direct and indirect costs, and capital and recurrent costs, if these costs will differ between the alternative training types. If the costs of alternative training types will not differ for certain cost categories, for example in the the amount of capital expenditures they consume, there is no need to calculate them to determine the relative cost differentials between two or more alternatives.

Direct costs are those costs that are clearly attributable to the training event itself such as the cost of staff time and participants' time (if they are paid employees of the organization), the cost of training materials, the cost of per diem, etc. Indirect costs include costs that indirectly apply to the training event such as the cost of a secretary's time required to mail out course invitations, the cost of the executive director's time

when giving the welcoming remarks during a course, or the cost of the utilities that cool, light, or provide water to a classroom. Each of these inputs benefits the training event in question, but since they serve other functions as well and are not exclusively devoted to the training event, they are considered indirect costs.

Recurrent costs are those costs that pay for items that are consumed in a relatively short time period, such as manuals given to course participants, pens and paper, diplomas, staff time, per diem, etc. Capital expenses, on the other hand, are those expenses that are made to purchase equipment, buildings, or other resources whose life will exceed any one training event. Purchasing a film projector is one example of a capital expense that many training departments incur. When capital expenses such as a film projector are made, a determination should be made as to how much of the cost of the capital expense should be allocated to an individual training event. At times these decisions are made arbitrarily, but a reasonable calculation can be made by approximating the total life of the projector (or building, or automobile) and estimating what percentage of it will be "consumed" by an individual training activity.

Table 10 lists most of the possible training expense categories that might be considered when determining the costs of a particular type of training.

TABLE 10
TRAINING EXPENSE CATEGORIES

Salaries and Benefits

- of trainers (including honoraria for lecturers)
- of other organizational staff
- of trainees (if paid employees of the training organization)

Per diem

- of trainers
- of other organizational staff
- of trainees

TABLE 10 (Continued)

Transportation

- of trainers
- of other organizational staff
- of trainees

Office Supplies

This category includes supplies consumed by the organization relating to the training event including stationery, photocopying, postage, telephone expenses, and other miscellaneous office supply expenses.

Training Materials and Supplies

This category includes manuals, paper, pens, diplomas, the cost of developing flipcharts and other training materials, refreshments, and all other supplies and materials, whether recurrent or capital, that pertain to a training event.

Outside Services

This category applies to the cost incurred for consulting fees and expenses of outside consultants and institutions, other than the agency's own in-house staff, who perform special services not covered under a previous category.

Equipment Expense

This category includes both the cost of rental equipment and a portion of the cost of capital expenditures previously made for equipment that were used in the training event in question. In addition, the cost of maintaining and repairing equipment should be proportionately allocated to all training events that utilize such equipment.

Registration Fees

If an organization pays registration fees to have its personnel trained by another agency, those fees are naturally training costs.

Facilities Expense

This category includes the cost of renting facilities used in connection with training, or the cost of using agency-owned facilities for a training event.

General Overhead

This category includes general overhead expenses usually determined as a percentage of the subtotal expenses of a training program.

Other Miscellaneous Expenses

Any expenses not provided for elsewhere are listed under this category.

The list of possible training expenses in Table 10 is a long one and may include many items not normally reported as training expenses. To complicate the picture, many training agencies that receive multiple donor support may receive funds from one donor for the direct costs of training, while receiving funds from another donor for many of the indirect costs of training, such as core administration costs. Usually the training agency will report, and interpret, only the direct costs of training, excluding the indirect costs from the analysis because those costs are covered by other administrative support funds. Regardless of how the expenses of training are reported in accounting statements, or to international donors, it is important to take into account all these possible direct and indirect costs when comparing alternative forms of training, when it is presumed that the amounts of the categories differ between alternatives.

Another aspect of training costs to consider is the cost involved in the analysis, development, delivery and evaluation of training. Often, only the delivery stage, and sometimes the evaluation stage, of training is considered when determining the costs of training. However, as all trainers know, there is (or should be) a considerable amount of time spent in planning and developing curricula, field testing evaluation instruments, performing needs assessments, and evaluating the outcomes of training, all of which constitute training expenses. Many, but not all, of the categories in Table 10 should have four costs associated with them: the cost of training analysis, development, delivery, and evaluation. Again, only the relevant categories that would affect the alternatives being compared need to be analyzed in these terms in order to calculate a CEA that can be used to compare alternative training types.

A final point to remember when calculating training costs is the need to include the cost of voluntary labor and donated goods, if one is to obtain the true cost of training. These represent legitimate costs that should have their market value calculated, particularly when one alternative received such voluntary labor or donated goods, and another alternative did not.

Returning to the example at hand, you determine (with the help of your accountant) that your eight alternatives will have the following costs associated with them for training 100 distributors for each alternative (to simplify the analysis, only a lump cost for each category is presented without subdividing it by the analysis, development, delivery, and evaluation components):

COST OF ALTERNATIVES (Hundreds of $)

	1	2	3	4	5	6	7	8
Salaries	5	5	2.5	2.5	3	3	1.5	1.5
Per diem	20	2	20	1	12	1.2	12	.6
Transportation	4	.5	4	.5	4	.5	4	.5
Office Supplies	5	5	5	5	5	5	5	5
Training Materials	3	3	3	3	3	3	3	3
Facilities	.5	1	.5	1	.5	1	.5	1
Subtotal	37.5	16.5	35	13	27.5	13.7	26	11.6
Overhead	3.8	1.7	3.5	1.3	2.8	1.4	2.6	1.1
Total	41.3	18.2	38.5	14.3	30.3	15.1	28.6	12.7

From the projected costs listed for each alternative, you see that alternative #1 (the type of training you presently perform) is projected to be the most expensive (due to higher per diem, transportation and salary costs), while alternative #8 is projected to be the least expensive. However, it is premature to conclude that alternative #8 is the best alternative to pursue before looking at the outcomes that are likely to result from each alternative.

Determination of Net Outcomes

Determining the proper outcome to focus on in a CEA is not an easy matter. Depending upon what outcome is analyzed, a particular program or type of training could be judged as more or less cost-effective than other alternatives. It greatly simplifies the CEA if only one type of outcome is analyzed. However, at times programs or training events have multiple outcomes that need to be taken into account. A family planning program could, for example, have a measurable impact on both infant mortality and maternal mortality. In this case an evaluator might want to compare alternative programs in terms of their cost per maternal life saved, or their cost per infant life saved. Many evaluators perform multiple CEA's when multiple outcomes are of interest. Alternatively, they may come up with a composite index that attempts to measure both outcomes in one single indicator. The Physical Quality of Life Index[10] is an example of a composite index that takes into account the levels of literacy, infant mortality, and life expectancy in a given population. A CEA could, therefore, be expressed in terms of cost per one point increase in PQLI.

In the current example, you determine that the appropriate outcome to measure is the number of trainees who successfully complete (i.e. meet or exceed all of the minimal competencies required) the training in the amount of time given according to each alternative. Participants who fail to meet the minimal competencies required will be considered to have "failed" the training, but will in fact, be given additional training prior to allowing them to begin their assigned work.

The minimal competencies required for successful completion of the training have already been determined by your department and include requirements such as

10. The Physical Quality of Life Index was developed by the Overseas Development Council, Washington, D.C.

- each trainee will be able to name the 8 absolute contraindications to oral contraceptive use;
- each trainee will be able to name the 5 danger signs that indicate possible serious complications of oral contraceptive use;
- each trainee will be able to give a 5 minute presentation on the correct use of one contraceptive method and mention the key points required to know how to effectively and safely use the method.

etc....

After developing the list of minimal competencies, you instruct your trainers to follow the list strictly to determine the percentage of participants who pass (meeting or exceeding all the minimal competencies) the training in the 3 or 5 days alloted to each alternative. From previous experience and from the predictions of the trainers in your department, you project the following outcomes:

Alternative	Percentage Passing
1	98
2	95
3	93
4	90
5	88
6	85
7	83
8	80

Determination of Cost Effectiveness Ratio

Having calculated the net costs and net outcomes, you then proceed to calculate the cost-effectiveness ratio, which is simply the cost of a given alternative divided by its level of outcome. The corresponding ratios for each of the alternatives are the following:

Alternative	Cost Effectiveness Ratio		
1	4,130 ÷ 98 X 100	trainees	= $42.14 per passing trainee
2	1,820 ÷ 95 X 100	"	= $19.16 "
3	3,850 ÷ 93 X 100	"	= $41.40 "
4	1,430 ÷ 90 X 100	"	= $15.89 "
5	3,030 ÷ 88 X 100	"	= $34.43 "
6	1,510 ÷ 85 X 100	"	= $17.76 "
7	2,860 ÷ 83 X 100	"	= $34.46 "
8	1,270 ÷ 80 X 100	"	= $15.88 "

Analyzing these ratios you find that alternative 8, in addition to having the lowest costs associated with it, has the lowest cost-effectiveness ratio, i.e. it is the most cost-effective. This leads you to believe that you should probably eliminate the alternative you have been using (#1) and adopt alternative #8. Alternative #4 has essentially the same cost-effectiveness ratio which would make it a good alternative to consider as well. Before going ahead with either of these options, however, there is one final step you should perform which is to conduct a sensitivity analysis.

Perform Sensitivity Analysis

A sensitivity analysis in CEA is the process of varying certain assumptions that were made in either the cost or outcome projections. Whenever a CEA is done prospectively, assumptions must be made that can have varying degrees of certainty associated with them. A sensitivity analysis changes the parameters of a CEA about which the evaluator might be relatively unsure and then recomputes the costs and outcomes of each alternative.

In the case study, suppose you were fairly sure that your cost projections were extremely accurate but your projections of the percentage of participants who would pass the course was only an educated guess. You would want to vary the possible outcomes to see if the alternative that appeared the best given your assumptions, remained the best after altering the assumptions within reasonable boundaries. For example, if you

determined that conducting the course in 3 days versus 5 days would result in a 20 percentage point drop in the number of participants passing, rather than a 10 point drop as was assumed originally, the cost-effectiveness ratios would be the following (keeping all other assumptions the same):

Alternative	Percentage Passing	Cost Effectiveness Ratio
1	98	$42.14
2	95	$19.16
3	93	$41.40
4	90	$15.89
5	78	$38.85
6	75	$20.13
7	73	$39.18
8	70	$18.14

With this new assumption, alternative 8 is no longer the most cost-effective but rather alternative #4. If other assumptions were deemed possibly unreliable, they too should be changed and the resultant cost-effectiveness ratio calculated. In reality, it is nearly impossible to calculate precise levels of costs and outcomes before training occurs. In the sensitivity analysis, the evaluator should, therefore, calculate the highest and lowest amounts that could reasonably be expected for both costs and outcomes and then calculate the cost-effectiveness ratios. If the same alternative consistently appears as the most cost-effective, then the choice becomes much easier to make. When differing assumptions create differing conclusions as to the most cost-effective alternative, then the alternative will have to be chosen on other criteria or on the basis of which assumption the evaluator has the most confidence in.

In the case study example, if you were to choose the type of training to pursue based solely on the prospective CEA you performed, you would do best to choose alternative #4 since it tied with alternative #8 for the lowest cost-effectiveness ratio given the first assumption about training outcomes and had the lowest cost-effectiveness ratio given the second assumption about training outcomes.

This case study illustrates the important tool that CEA's can be for training evaluators considering several courses of action to take. CEA's help training evaluators make decisions on the basis of their efficiency, rather than by some other arbitrary or subjective criteria.

However, CEA's can only be as good as the assumptions (for prospective CEA's) or measurements (for retrospective CEA's) that go into making them. In addition, there are many situations where criteria other than cost effectiveness must influence the alternative selected. Oftentimes, the most cost-effective option is not a politically viable option, or might negatively affect other aspects of an agency's services. In short, a wide range of criteria usually go into making a decision about which of several program or training alternatives to pursue, and cost-effectiveness is one of the most important, but not the exclusive, criteria that training evaluators should use.

APPENDIX

EXAMPLES OF TRAINING EVALUATION INSTRUMENTS

With the exception of example 3, the following are unedited examples of training evaluation instruments that have been used by a variety of family planning agencies in the Latin American/Caribbean Region. Although the instruments were chosen because they are good examples overall, they may contain individual questions that are confusing, ambiguous, or in one way or another, need improving upon. They are included only for the purpose of illustrating various styles of developing instruments that meet different needs, and for stimulating ideas for developing similar types of instruments. They are not meant to be instruments that one can use without adaptation to local needs and realities.

Each instrument is presented under the category for which it is primarily intended, i.e. process or outcome evaluation.

Example 1: Process Evaluation Instrument

This instrument is completed by course participants at the end of their training in order to obtain information from them as to their perception of the course and its strengths and weaknesses. Its format contains both a Likert scale, as well as open-ended responses to clarify the respondent's views and elicit a broader range of information.

COURSE EVALUATION

Name of Course: _____

Date: _____

1. Did the course fulfill your expectations?

   Yes ( ) Partially ( ) No ( )

   Why? _____

2. Were the proposed objectives of the course accomplished?

   Yes ( ) Partially ( ) No ( )

   Why? _____

3. Were all doubts clarified?

   Yes ( ) Partially ( ) No ( )

   Why? _____

4. Did the course provide opportunities for discussions?

   Yes ( ) Partially ( ) No ( )

   Why? _____

5. Regarding the organization of the course, did you find it:

   Good ( ) Average ( ) Poor ( )

   Why? _____

6. Regarding the course methodology, did you find it:

   Good ( ) Average ( ) Poor ( )

   Why? _____

7. Describe the course's negative points.

_____

_____

8. Describe the course's positive points.

_____

_____

9. Please provide your suggestions for future courses.

_____

_____

Example 2: Process Evaluation Instrument

This instrument utilizes a rating scale along with two open-ended questions to elucidate comments from training participants regarding their instructors.

EVALUATION OF LECTURERS BY PARTICIPANTS IN COURSE

This evaluation instrument contains a series of questions that cover the performance of the professors (lecturers) during the course. It is filled out by the participants at the end of the course and is based on a numerical scale with a range between a value of one (1) to nine (9), in which one (1) indicates "deficient" and nine (9) is "very good".

1. Preparation of the working materials for the course was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

2. Mastery of the training subject by the lecturer was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

3. Selection of training and audiovisual materials for development of the subject was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

4. The quality of responses to questions submitted by participants was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

5. The lecturer's ability to hold the attention of the participants was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

6. The promotion of active participation by the participants was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

7. The lecturer maintained leadership of the group:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

8. The establishment of good relations with the group was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

9. Use of time by the lecturer was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

10. Fairness in the way participants were treated was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

11. The manner in which the participants' opinions were treated was:

Deficient								Very Good
1	2	3	4	5	6	7	8	9

Remarks: _____

_____

_____

_____

Suggestions for Lecturer: _____

_____

_____

_____

Example 3: Process Evaluation Instrument

This instrument, taken from Mayo and DuBois[11], contains a rating scale to be used by an observer to evaluate the trainer. It is quite comprehensive in nature including questions on various elements of the training session, the trainer's techniques, and the trainees' responses. The form is meant to be shared with the trainer and contains a section for the trainer to write out a "trainer improvement plan".

11. Mayo, G.D.; DuBois, P.H., The Complete Book of Training, San Diego: University Associates, pages 59-60.

TRAINER EVALUATION

Date		
Trainer	Course	Module/Period
Lesson Topic Title		

	Guides for Evaluation P - Poor F - Fair G - Good E - Excellent	P	F	G	E
Elements of a Learning Session Provided by the Trainer	Prepares trainees for learning.				
	Provides learning objectives.				
	Motivates in terms of how the material is to be used.				
	Motivates in terms of why the material needs to be learned.				
	Conveys enthusiasm to the trainees.				
	Establishes and maintains rapport in a professional manner.				
	Senses trainee needs.				
	Holds the respect of the trainees.				
	Allows trainees to communicate with him or her.				
	Provides clarification, amplification, and reinforcement of the learning objectives as necessary for achievement.				
	Handles behavior problems in an effective manner.				
Trainer Techniques	Selects and uses media and/or facilities effectively.				
	Uses training skills effectively.				
	Conducts sessions on the trainees' level.				
	Uses good questioning techniques.				
	Uses a well-modulated voice.				
	Shows care in personal appeareance.				
	Demonstrates flexibility in adjusting to unplanned and extemporaneous learning situations.				
	Displays knowledge of subject matter.				
	Demonstrates adequate ability to use learning materials.				
	Manages time efficiently.				
	Shows evidence of careful planning of presentation.				

Guides for Evaluation P - Poor F - Fair G - Good E - Excellente		P	F	G	E
Trainee Responses	Evidence of interaction between trainer and trainees.				
	Evidence of group involvement.				
	Evidence of attainment of learning objectives through testing				
	Choice and use of media and/or resources/facilities.				
	Demonstration of self-management.				
GENERAL EVALUATION OF TRAINER					

Remarks

I certify that the trainer evaluated was critiqued immediately after evaluation

Signature of Evaluator	Name and Title	Date

To Be Completed by Trainer

I have been critiqued on this evaluation. My Trainer Improvement Plan is as follows:

Signature of Trainer	Name and Title	Date

Example 4: Outcome Evaluation Instrument–Attitude Changes

This instrument is designed to measure attitude changes as a result of training and is given to participants prior to and subsequent to their training. It utilizes a modified checklist format.

EVALUATION OF ATTITUDES REGARDING SEXUALITY AND FAMILY PLANNING

READ EACH OF THE FOLLOWING STATEMENTS CAREFULLY, THEN PLACE AN (X) IN THE BOX WHICH, IN YOUR OPINION, IS THE APPROPRIATE ONE.

SINCE THIS SECTION IS NOT SUBJECT TO SCORING, WE REQUEST THAT YOU PLACE THE (X) IN ACCORDANCE TO HOW YOU FEEL AND NOT ACCORDING TO WHAT YOU THINK WE WANT YOU TO ANSWER.

KEY TO RESPONSES:

CA: Place (X) in this box if you are in complete agreement with the statement.

A: Place (X) in this box if you agree partially with the statement.

NA: Place (X) in this box if you do not know what to answer regarding the statement.

D: Place (X) in this box if you disagree partially with the statement.

CD: Place (X) in this box if you are in complete disagreement with the statement.

	CA	A	NA	D	CD
1. A couple in love, but not married, can have sexual intercourse.					
2. A woman should be a virgin at the time of marriage					
3. Men prefer to marry virgins.					
4. A man should be a virgin at the time of marriage.					
5. Women prefer to marry a virgin man.					
6. Women may have extra-marital sexual intercourse.					

	CA	A	NA	D	CD
7. Men may have extra-marital sexual intercourse.					
8. Men must possess greater sexual experience than women.					
9. Only women should worry about family planning.					
10. The man should take the initiative regarding sexual intercourse.					
11. Adolescents should use birth control methods.					
12. Women should enjoy the same freedom that men do.					
13. The use of birth control methods results in sexual licentiousness.					
14. The woman should take the initiative regarding sexual intercourse.					
15. Sex education for adolescents promotes sexual intercourse among them.					
16. Men should have the same sexual liberties and restrictions as women.					
17. Only men should worry about family planning					
18. During sexual intercourse it is the woman who should try to avoid an unwanted pregnancy.					
19. The basic purpose of marriage is to have children.					
20. The basic purpose of family planning is to avoid more births.					

Example 5: Outcome Evaluation Instrument–Knowledge Changes

The purpose of this instrument is to test the amount of increase in participants' knowledge relating to contraceptive methodology as a result of training. It is given as both a pretest and a posttest. Its format includes true/false and multiple-choice questions.

KNOWLEDGE ABOUT FAMILY PLANNING

MARK THE BOX WITH (T) IF THE STATEMENT IS TRUE OR WITH (F) IF THE STATEMENT IS FALSE.

1. ( ) Temporary birth control methods are those that, when discontinued, allow you to have children again.

2. ( ) Female sterilization is a temporary birth control method.

3. ( ) The pill is the least reliable of the temporary methods.

4. ( ) The most reliable methods are female sterilization and vasectomy.

5. ( ) The pill acts by preventing ovulation.

6. ( ) Women with serious circulatory problems, such as phlebitis, may use the pill.

7. ( ) Women with an infection of the vagina and/or the cervix, may use an IUD.

8. ( ) A condom may be used several times.

9. ( ) Protection during sexual intercourse will be greater if a condom is used together with vaginal foam or suppositories.

10. ( ) Use of a condom helps prevent sexually transmitted diseases (venereal diseases).

11. ( ) The rhythm method is based on the approximate knowledge of the date of ovulation.

12. ( ) The operation known as vasectomy consists of the severing and closure of the vas deferens.

13. ( ) Vasectomy results in sexual impotence in the male.

14. ( ) Women with very irregular menstrual periods may protect themselves very well by using the rythm method.

SELECT THE CORRECT ANSWER BY ENCLOSING IT IN A CIRCLE. THERE IS ONLY ONE CORRECT ANSWER PER QUESTION:

15. Barrier methods are:

    a. Final
    b. Temporary

16. Condoms act by:

    a. Preventing ovulation
    b. Preventing the implantation of the fertilized ovum
    c. Destroying the spermatozoa
    d. All of the foregoing
    e. None of the foregoing

17. Use of the pill must begin on:

    a. The first day of the menstrual flow
    b. The last day of the menstrual flow
    c. The fifth day of the menstrual flow
    d. All of the foregoing
    e. None of the foregoing

18. The woman who is using the pill:

    a. Must take it only when having sexual intercourse
    b. Must give it to her husband
    c. Must take it every day (one per day)
    d. All of the foregoing
    e. None of the foregoing

19. If a woman using the pill forgets to take it one day, she:

    a. Must stop taking the pill and use a different method.
    b. Must take two pills: one as soon as she remembers and one at bedtime of that day and continue to take one every night at bedtime.
    c. Continue to take them as if nothing had happened (one per day).
    d. All of the foregoing.
    e. None of the foregoing.

20. If a woman using the pill forgets to take it during two consecutive days, she:

    a. Must stop taking the pill and use a different method.
    b. Must take two pills.
    c. Must continue to take it as if nothing had happened.
    d. All of the foregoing.
    e. None of the foregoing.

21. The IUD is placed in:

    a. The abdomen
    b. The vagina
    c. The uterus
    d. All of the foregoing
    e. None of the foregoing

22. The best time to insert the IUD is:

    a. Before the menstrual flow begins.
    b. Half way through the menstrual cycle.
    c. Immediately after the menstrual flow ends.
    d. All of the foregoing.
    e. None of the foregoing.

23. Female sterilization works as a method to prevent pregnancy because it:

    a. Prevents ovulation.
    b. Interrupts passage of the ovum through the fallopian tubes.
    c. Destroys the spermatozoa.
    d. All of the foregoing.
    e. None of the foregoing.

24. Vasectomy works as a method to prevent pregnancy because it:

    a. Prevents production of spermatozoa.
    b. Interrupts passage of spermatozoa by severing the vas deferens.
    c. Prevents ejaculation.
    d. All of the foregoing.
    e. None of the foregoing.

25. When a man undergoes a vasectomy:

    a. He can no longer ejaculate.
    b. He becomes impotent.
    c. He ejaculates but his sperm contains no spermatozoa.
    d. All of the foregoing.
    e. None of the foregoing.

Example 6: Outcome Evaluation Instrument–Job Performance

This instrument is presented in two parts, both of which are designed to evaluate the effect of training on the job performance of family planning promoters. The first part is a short questionnaire given to the promoter to fill out regarding the volume of project related activities he or she has undertaken since training, or the reasons for lack of such activities.

The second part is a monthly activity sheet filled out by the promoter and delivered to the supervisor in charge. It is an example of an existing record that the institution maintains that can be used to evaluate the effect of training on the promotor's output of certain project activities. Since it asks for nearly the same amount of information as the questionnaire, and is collected on a monthly basis, it can be used to compare trends in the volume of activities undertaken by trainees before and after their training.

PROMOTION ACTIVITIES

(EVALUATION)

1. After completing your training, have you carried out any activities in your community?

   YES _____ NO _____

2. Check the type of activities you have carried out and indicate how many of each you have performed:

   Individual conversations: # _____
   Home visits: # _____
   Referral of persons to a physician or family planning clinic. # _____
   Talks given to groups # _____
   Other # _____

3. If you have not begun promotion activities in your community, what have been the reasons? (indicate the appropriate ones).

   Lack of resources _____
   Lack of opportunities _____
   Lack of knowledge _____
   Lack of training _____
   Lack of interest _____
   Lack of confidence _____

Date: _____
Date of Training: _____
Name of Evaluator: _____

FORM USED BY PROMOTERS

FOR FIELD WORK

Record of Promoter's Activities

Name of Promoter: _____ Community: _____

Date	Community	Subject	ACTIVITIES				
			Individual Conversation	Home Visits	Organization of Groups	Medical Referral	Group Talks

Delivery Date: _____

_____
Signature of Promoter

Example 7: Outcome Evaluation Instrument–Job Performance

This instrument is an interview guide designed to be used when interviewing family planning users about the quality of service they received from CBD workers. Its limitation is that it relies on the users' recall of what the CBD worker said and did during his or her initial contact with the user. For this reason this kind of instrument should not be used when a long period of time has elapsed between the time of the interview and the time of the behavior being evaluated. The format of this instrument includes two-way questions and checklists.

INTERVIEW GUIDE FOR NEW USERS

INSTRUCTIONS: Below you will find a series of questions. Please mark with an X the correct answer.

1. When you first went to request a contraceptive method, were you informed about all the services available?

   (If the answer is yes, please go to question No. 2 . If the answer is no, go to question No. 3.)

2. Do you remember if any of these services were mentioned to you?

   _____ Medical exams

   _____ Contraceptive methods

   _____ Papanicolaou exam

   _____ Voluntary sterilization

3. Before you were given a contraceptive method, were you asked if you suffer from any illness.

   Yes ( ) No ( )

4. Before the method was given to you, were you told how it should be used?

   Yes ( ) No ( )

5. Were you given any written instructions when the method was given to you?

   Yes ( ) No ( )

6. How are you using the method?
   (please mark only the ones the user mentions)

   _____ She explained correctly how the method should be used.

   _____ She gave one or more incorrect statements on how the method should be used.

_____ she stated that she received written and verbal instructions on how to use the method.

_____ she stated that she only received written instructions.

7. Were you referred to a clinic for an exam? (Ask only in the case of pill users)

Yes ( ) No ( )

8. Were you told that some women may have some discomforts when using the pill?

Yes ( ) No ( )

9. If you had any problems with your method what would you do? (mark only the answers given by the user)

_____ would attend a clinic

_____ would see the promoter

_____ would see a doctor

_____ would stop using the pill

_____ other

10. Did the promoter clear up any doubts that you may have had?

Yes ( ) No ( )

11. Did the promoter treat you in a polite manner?

Yes ( ) No ( )

12. Are you satisfied with the services you received?

Yes ( ) No ( )

Name of Supervisor:_____

Name of Promoter:_____

Example 8: Outcome Evaluation Instrument–Job Performance

This instrument is used as a questionnaire to assess the effect of training on participants' job performance in three areas: clinical service delivery, training activities, and supervision/administration activities. It is used primarily to evaluate training given to nurses and nurse-midwives. Its format includes both open-ended questions and checklists.

186

TRAINING SURVEY
DEVELOPMENT ASSOCIATES, INC.

Name: _____

Part I: General Information: We need information about your current work and the family planning training you have received.

A. What is your current job title and the name and address of your institution? (If you work in family planning in more than 1 institution, please use the additional spaces for this information. Include institutions or agencies in which you do volunteer work.)

1. Job title: _____

   Institution: _____

   Address: _____

   _____

   _____

   Is this institution a government agency or a private institution?

   government _____ private _____

2. Job title: _____

   Institution: _____

   Address: _____

   _____

   _____

   Is this institution a government agency or a private institution?

   government _____ private _____

3. Job title: _____

   Institution: _____

   Address: _____

   _____

   _____

Is this institution a government agency or a private institution?

government _____ private _____

B. At which of the following institutions did your receive training in family planning sponsored by Development Associates?

1. Centro Docente (Chile) _____
2. Profamilia (Colombia) _____
3. Universidad de Durango (México) _____

C. How long has it been since you completed your training?

1. Less than 6 months _____
2. 6-12 months _____
3. 13-24 months _____
4. More than 2 years _____

D. Have you received any training in family planning other than that sponsored by Development Associates at Centro Docente, Profamilia, or Durango? (Please check all the appropriate boxes).

Type of Training	Before DA Training	After DA Training	Length of Training
1. Formal training in nursing or other medical school			
2. Inservice training (clinical)			
3. Inservice courses or seminars (theory)			

Please describe this training: _____

_____

_____

_____

PART II: Clinical Family Planning Activities: We need information about your clinical family planning activities, both before and after your Development Associates' sponsored training.

A. Do you provide any direct clinical family planning services to patients? (Please check all the appropriate boxes. "Before Training" refers to your activities during the year prior to the DA training; "After Training" refers to any time after the DA training; "Currently" refers to this current month.)

	Before Training	After Training	Currently
1. Yes			
2. No			

If you have not provided any clinical family planning service either before or after your training, go on to Part III. If you have provided clinical family planning services, please continue with Part II.

B. If you do provide or have provided direct clinical family planning services, please place a check ( ) in the appropriate boxes to indicate those you provide before your training and those you have provided since.

Service	Provided within the year before DA training	Provided after DA training	Provided currently	This service is not provided by my institution
1. Prescribe oral contraceptives				
2. Re-suppy oral contraceptives				
3. Insert IUDs				
4. Distribute other contraceptives (foam, condoms)				
5. Perform pelvic exams				
6. Perform breast exams				
7. Treat vaginal infections				
8. Perform PAP tests				
9. Assist with sterilizations				
10. Provide family planning education to patients				
11. Provide injectable contraceptives				
12. Deal with patients' problems or complaints related to their contraceptive method				

Comments: _____

_____

_____

C. If you are <u>not</u> <u>permitted</u> to provide any of these services by your institution, the supervising physician, or the

supervising nurse, please place a check ( ) in the appropriate boxes.

Service	Not permitted by institution	Not permitted by supervising physician	Not permitted by supervising nurse
1. Prescribe oral contraceptives			
2. Re-suppy oral contraceptives			
3. Insert IUDs			
4. Distribute other contraceptives			
5. Perform pelvic exams			
6. Perform breast exams			
7. Treat vaginal infections			
8. Perform PAP tests			
9. Assist with sterilizations			
10. Provide family planning education to patients			
11. Provide injectable contraceptives			
12. Deal with patients' problems or complaints related to their contraceptive method			

Please explain <u>why</u> you are not permitted to provide the services you checked above, e.g., it is against institutional policy for nurses (obstetrices, matronas) to provide these services; your supervising physician or nurse does not believe nurses (obstetrices, matronas) should provide these services, etc.

_____

_____

_____

D. Are the supplies or material that you need to provide clinical family planning services available in adequate supply? Please place a check ( ) in the appropriate boxes.

Supply/Material	Always Available	Sometimes Available	Never Available	Supply is not needed in current clinical work
1. Contraceptive supplies				
2. Instruments (Equipo)				
3. Patient Education Materials				
4. Other materials or supplies (Please specify)				

Comments: _____

_____

E. What percentage of your time is devoted to providing clinical family planning services to patients? (Please check all the appropriate boxes. There are 2 other questions in this survey which ask you for the percentage of your time devoted to certain tasks. Please be sure these do not total more than 100%)

Percent of Time	Before Training	After Training	Currently
1. 75-100%			
2. 50-74%			
3. 25-49%			
4. 10-24%			
5. Less than 10%			

F. Which of the following statements most accurately describes the clinical family planning services which you have provided since your training compared with those you provided before your training? (Please place a check ( ) beside all the appropriate statements).

1. I provide much better services now than I did before my training. _____
2. My training did not substantially increase my clinical family planning skills. _____
3. I feel much more confident in my family planning skills since my training. _____
4. I feel no more confident of my family planning skills now than I did before my training. _____
5. Since I completed my training, my supervisor has encouraged me to provide more clerical family planning than before. _____
6. Since I completed my training, I provide fewer clinical family planning services than I did before. _____
7. I did not provide any clinical family planning services before my training. _____
8. I do not currently provide clinical family planning services, although I have provided them since completing my training. _____

Comments: _____

_____

_____

G. Approximately how many patients do you personally provide clinical family planning services to in a typical week? (Please check all appropriate boxes).

Number of Patients	Before Training	After Training	Currently
1. More than 50			
2. 24-50			
3. 10-24			
4. Less than 10			

PART III: Training/Teaching Activities: We need information about your family planning teaching and training activities.

A. Do you conduct any training or academic teaching activities in family planning? (Please check all appropriate boxes. "Before Training" refers to the year prior to the DA trainin "After Training" refers to any time after the DA training; "Currently" refers to this current month.

	Before Training	After Training	Currently
1. Yes			
2. No			

If you have not conducted any training or teaching activities in family planning, either before or after your DA training, go on to Part IV. If you have conducted such activities, please continue with Part III.

B. If you do conduct or have conducted family planning training or teaching activities, please place a check ( ) in the appropriate boxes to indicate your activities both before and after your training.

Training/Teaching Activity		Conducted within the year before DA training	Conducted after DA training	Conducted currently	Total number of individuals taught or trained since you completed your training
1. Docencia clínica o teórica en una escuela para:	A: enfermeras				
	B. obstetrices o matronas				
	C. auxiliares				
2. Cursos o seminarios en servicios para:	A. enfermeras				
	B. obstetrices o matronas				
	C. auxiliares				
3. Cursos o seminarios en la comunidad para:	A. maestros de educación primaria o secundaria				
	B. líderes de comunidades				
	C. promotores				
	D. TBAs				
4. Other types of training (please specify)					

Comments: _____

C. Are the supplies and materials that you need to conduct teachi training activities available in adequate quantities? (Please place a check ( ) in the appropriate boxes).

Supply/Material	Always Available	Sometimes Available	Never Available	Not needed for current teaching or training activities
1. Educational materials for students				
2. Manuals for community workers				
3. Films and other audio visuals				
4. Instruments and supplies for clinical instruction				
5. Other (Please specify)				

Comments: _____

_____

_____

D. What percent of your time is devoted to conducting family planning training/teaching activities? (Please check the appropriate boxes. There are 2 other questions in this survey which ask you for the percentage of your time devoted to certain tasks. Please be sure these do not total more than 100%).

Percent of Time	Before Training	After Training
1. 75-100%		
2. 50-74%		
3. 25-49%		
4. 10-24%		
5. Less than 10%		

E. Which of the following statements most accurately describe the family planning training/teaching activities which you have conducted since your training in comparison with those which you conducted before your training? (Please place a check ( ) beside all the appropriate statements).

1. I am much better able to design a family planning curriculum now than I was before my training. _____

2. My training did not substantially increase my ability to design a family planning curriculum. _____

3. I can design and conduct much better clinical evaluations of family planning trainees/students now than I was before my training. _____

4. My training did not substantially increase my family planning evaluation skills. _____

5. I feel more confident of my family planning training abilities now than I did before my training. _____

6. I feel no more confident of my family planning training skills now than I did before my training. _____

7. I did not provide any family planning training before I received my training. _____

8. I do not currently provide any family planning teaching on training, although I have provided them since completing my training. _____

Comments: _____

_____

_____

Part IV: Supervision/Administration: We need information about your supervisory and administrative activities.

A. Do you have supervisory or administrative responsibilities in family planning? (Please check appropriate boxes. "Before Training" refers to the year prior to the DA training; "After Training" refers to any time after the DA training; "Currently" refers to this current month.)

	Before Training	After Training	Currently
1. Yes			
2. No			

If you have not had supervisory or administrative responsibilities in family planning, either before or after your DA training go on to Part V. If you have or have had such responsibilities, please continue with Part IV.

B. If yes, please place a check ( ) in the appropriate boxes to indicate your responsibilities both before and after your training.

Area of Responsibility	Responsible within the year before DA training	Responsible after DA training	Responsible currently
1. Supervise other family planning health personnel in a clinic or hospital unit			
2. Supervise or administer family planning services on an area level			
3. Supervise or administer family planning services on a regional level			
4. Supervise or administer family planning services on a national level			
5. Other administrative or supervisory responsibilities in family planning (Please specify)			

Comments: _____

_____

C. How many family planning workers do you supervise?

Number of Workers	Before Training	After Training	Currently
1. More than 50			
2. 24-49			
3. 10-24			
4. Less than 10			

D. Do you have the support you need from your institution to carry out your supervisory or administrative responsibilities? (Please check all appropriate boxes).

Type of Support	Always Available	Sometimes Available	Not needed for supervisory/ administrative activities
Transportation (car, driver, etc.)			
Per diem for supervision visits			
Supplies and materials for workers supervised			

Comments: _____

_____

E. What percentage of your time is devoted to performing supervisory/administrative activities in family planning? (Please check all appropriate boxes. There are 2 other questions in this survey that ask you for the percentage of your time devoted to certain tasks. Please be sure these do not total more than 100%.)

Percent of Time	Before Training	After Training	Currently
1. 75-100%			
2. 50-74%			
3. 25-49%			
4. 10-24%			
5. Less than 10%			

F. Which of the following statments most accurately describe your current supervisory/administrative duties compared with those you performed before your training? (Place a check beside each appropriate statement).

1. I am much better able to supervise and administer family planning activities now than I was before my training

2. My training did not substantially increase any ability to supervise and administer family planning activities. _____

3. I feel much more confident of my family planning supervisory/administrative skills now than I did before my training. _____

4. I feel no more confident of my family planning supervisory/administrative skills now than I did before my training. _____

5. I did not have any supervisory or administrative responsibilities in family planning before my training. _____

6. I do not currently have any supervisory or administrative responsibilities in family planning, although I have had such responsibilities since my training. _____

Comments: _____

_____

_____

Part V: Institutional Family Planning: We need information about family planning training and other people in your institution.

A. Who recommended that you receive your training?

1. Your immediate supervisor _____
2. Your area, regional, or national supervisor _____
3. Other (Please specify) _____

_____

_____

B. What kind of family planning training have others in your institution received? (Please check each appropriate box).

Type of Training	Your current immediate Supervisor	Your current co-workers (at your same professional level)	Your subordinates that you directly supervise
1. Extensive formal training (more than 1 month)			
2. Some formal training (1 month or less)			
3. Some informal training			
4. No family planning training			

C. Please describe how the family planning training, or lack of such training, that others in your institution have received has objected your ability to provide family planning services or training.)

_____

_____

_____

PART VI: Assessment: We are very interested in knowing if you feel that additional training or technical assistance would be beneficial to you in your family planning work. Please answer the following questions to help us understand your needs:

A. Are there any additional family planning clinical skills which you feel should have been included in your training but were not?

_____

_____

B. Are there any additional family planning teaching/training skills which you feel should have been included in your training but were not?

_____

_____

C. Are there any additional family planning supervisory/administrative skills which you feel should have been included in the training but were not?

_____

_____

In order to make our future courses more responsive to institutional needs, we would like to contact your immediate supervisor for information and suggestions. Please indicate his/her name, title, and mailing address:

_____

_____

_____

THANK YOU VERY MUCH FOR YOUR HELP. WHEN YOU HAVE COMPLETED THIS QUESTIONNAIRE, PLEASE RETURN IT TO DEVELOPMENT ASSOCIATES, INC., IN THE ENCLOSED ENVELOPE.

BIBLIOGRAPHY

Andrade, S.; Shedlin, M.; Bonilla, E. 1987. Metodos Cualitativos para la Evaluacion de Programas: Un Manual Para Programas de Salud, Planificacion Familiar y Servicios Sociales. Watertown, MA: The Pathfinder Fund.

Bhola H. 1982. Evaluating Development Training Programs. German Foundation for International Development.

Blumenfeld, S. 1985. Operations Research Methods: A General Approach in Primary Health Care. Chevy Chase, MD: PRICOR.

Brindis, C.; Korenbrot, C.; Brown, P. 1986. Evaluation Guidebook for Family Planning Information and Education Projects. San Francisco: Center for Population and Reproductive Health Policy, U.C.S.F.

Fisher, A.; Laing, J.; Stoeckel, J. 1983. Handbook for Family Planning Operations Research Design. New York: The Population Council.

Fitz-Gibbon, C.; Morris, L. 1983. How to Design a Program Evaluation. Beverly Hills: Sage Publications.

Green, L.; Lewis, F. 1986. Measurement and Evaluation in Health Education and Health Promotion. Palo Alto, CA: Mayfield Publishing Company.

Issac, S.; Michael, W. 1981 Second Edition. Handbook in Research and Evaluation. San Diego: Edits Publishers.

Kirkpatrick, J. 1981. The Magic of Structure: A Manual for Planning and Evaluating Family Planning Education Programs. New York: Planned Parenthood Federation of America.

Kosecoff, J.; Fink, A. 1982. Evaluation Basics: A Practitioner's Manual. Beverly Hills: Sage Publications.

Mayo, G.D.; DuBois, P.H. 1987. The Complete Book of Training: Theory, Principles, and Techniques. San Diego: University Associates.

Merwin, S. 1986. Effective Evaluation Strategies and Techniques: A Key to Successful Training. San Diego: University Associates.

Morris, L.; Fitz-Gibbon, C. 1978. How to Measure Achievement. Los Angeles: Sage Publications.

Phillips, J. 1983. Handbook of Training Evaluation and Measurement Methods. Houston: Gulf Publishing Company.

Reynolds, J. et al; 1978. Evaluation Handbook for Family Planning Programs. Rockville, MD: Department of HEW, PHS, HSA, Bureau of Community Health Services.

Reynolds, J.; Gaspari K. 1985. Operations Research Methods: Cost-Effectiveness Analysis. Chevy Chase, MD: PRICOR.

Windsor, R. et al; 1984. Evaluation of Health Promotion and Education Programs. Palo Alto: Mayfield Publishing Company.