CHAPTER 6

MEASUREMENT AND EVALUATION METHODS FOR WORKSITE STRESS MANAGEMENT PROGRAMS

Gene L. Stainbrook and Lawrence W. Green

INTRODUCTION

This review summarizes the main features of the cumulative development of measurement and evaluation in stress management programs in working settings.

Definitions of Measurement and Evaluation

Evaluation has been defined variously. Jemelka and Borich (1979) defined it as a process for decision making, Nutt (1981) as a measure of the degree to which objectives have been achieved, and Green (1974) as the comparison of an object of interest against a standard of acceptability. In contrast to basic research, evaluation implies and requires from the onset criteria and procedures for making judgments of merit, value, or worth (Scriven, 1967). Measurement represents the systematic application of procedures for assessing quantities and qualities, whether for purposes of planning or of evaluation.

Purposes of Evaluation and Measurement

As a systematic endeavor, evaluation serves two general purposes. One purpose is to assess the impact or effectiveness of products and services in achieving pre-determined objectives. A second purpose is to assess the efficiency of products and services in bringing about any change, but more commonly in achieving pre-established objectives. The assessment of effectiveness requires the detection of a change or effect compared with some absolute criterion or standard. In contrast, the assessment of efficiency requires the detection of change relative to some comparable product or service. Other applications of measurement, besides evaluation, include the assessment of employee needs, experiences, and interests prior to their recruitment into a program.

A common purpose of most program evaluations is to determine effectiveness; specifically whether the program objectives are being met. A second common purpose is to determine the efficiency or comparative effectiveness of two or more programs or methods within a

---

Note: The authors are indebted to Richard A. McCuan, M.S., Pre-doctoral Fellow, Chris Lovato, Ph.D., Faculty Associate, and Patricia Mullen, Dr. P.H., Associate Director, Center for Health Promotion Research and Development, The University of Texas Health Science Center at Houston.

program. A third purpose is often to assess the cost-benefit ratio or the cost-effectiveness of the program. A list of general reasons for evaluation adapted from prior summaries (Rossi and Freeman, 1982; Shortell and Richardson 1978; and Weiss, 1972) is displayed below:

- o To determine how effective a program has been in achieving its goals.
- o To examine how efficient a program has been in achieving its goals.
- o To determine the success of a program with different target groups.
- o To study the cost-benefit of a program.
- o To determine the cost-effectiveness of a program.
- o To justify past or projected expenditures.
- o To gain greater control over a program.
- o To determine future courses of action.
- o To contribute to the fields of applied and basic knowledge.

The priority given to particular reasons for evaluation usually depends on the perspective of the program sponsor. For example, executives may be concerned primarily with outcomes and costs, program managers may be interested in program utilization and impact, and participants may be primarily concerned with their own personal interests and satisfaction.

Program Planning

The first stage in the development of a stress management program is planning and the first step in planning is the assessment of needs. The care with which a program is planned often determines the quality of the evaluation that can be done. Ideally, considerations of measurement and evaluation should be an integral part of the planning process.

Three basic steps should be used in the planning and determination of the scope and specific direction of stress management programs. These steps are (1) conducting a needs assessment, (2) establishing priorities, and (3) specifying, goals and objectives. Each of these steps will be discussed briefly.

Needs Assessment. The first step in program planning should be to conduct a needs assessment. While the importance of thorough needs assessments to the success of programs often has been emphasized, it still remains a weak component of most programs. The rationale and methods of needs assessments have been detailed by many writers (French and Kaufman, 1983; Rossi and Freeman, 1982; Siegel et al., 1977; Warheit et al., 1977). This topic is treated specifically as it relates to health education and health promotion programs by Green et al. (1980); Parkinson et al. (1982); and Green and Lewis (1986); and as it relates more to mental health programs by Siegel et al. (1977); and Warheit et al. (1977).

The purpose of a needs assessment is to identify and document the type and severity of problems in particular populations. Six objectives of needs assessments have been identified by Green et al. (1980). These are presented in below they apply to worksite programs:

1. To determine the subjective concerns with quality of life in the employee population and with productivity in the employer population.

2. To verify and clarify these concerns with analyses of existing business and social indicators and other available information sources.

3. To document the status of the employee and employer groups in relation to those priority concerns for which there is a health component or cause.

4. To make explicit the rationale for the selection of priority problems.

5. To use the documentation and rationale to justify the further expenditure of resources for the selected problems.

6. Ultimately, to use the documentation and rationale as the bases on which to set objectives and to evaluate the program in cost-effectiveness or benefit terms.

Several strategies are available to determine the type and extent of problems that exist in particular jobs and work settings. Initial steps of a needs assessment include a social or economic diagnosis, an epidemiological diagnosis, and a behavioral diagnosis (Green et al., 1980). The first assesses data on social or economic problems of the company or employees. The epidemiological diagnosis assesses data on the presence of work-related problems, and on the incidence and prevalence of physical and mental health problems contributing to the

social or economic problems in the specific employee population or firm. Another method consists of making comparisons between rates of problem indicators in different work populations.

Comparative data from sources such as the National Center for Health Statistics, the National Institute for Occupational Safety and Health, other agencies of the Department of Health and Human Services, local and state health departments, and other planning agencies are useful in documenting a particular worksite problem. Also, social indicators for particular occupations and communities can be used (Attkisson et al., 1978; Sheldon and Parke, 1975). Soliciting employee views through small group techniques like the Nominal Group Technique (Delbecq, et al., 1971) and the Delphi Technique (Dalkey and Helmer, 1969) also may be helpful.

The use of direct archival data from company records or files provides another source of information. Data on attendance, rates of accidents and injuries, and use of mental and physical health services are useful in establishing program needs. Data from records on use of services, however, must be used cautiously as a basis for an epidemiological diagnosis. Often strong biases exist in the use of services by subgroups of employees and in the reporting of service use (Attkisson et al., 1978). When used alone for epidemiological diagnosis, this information could seriously distort the needs assessment. Use of services does provide a good measure of behaviors associated with health problems, however, and therefore contributes to the behavioral diagnosis.

Key informant surveys can be used to collect information from employees known to have knowledge about problems and needed services. This method can provide valuable insights on both behavior and the next step following behavioral diagnosis--the educational diagnosis. This method provides a more balanced perspective when views are obtained from sectors of the company likely to have contrasting views, such as management and labor (Neale et al., 1983; Martin, 1983).

After establishing the presence of specific behavioral problems, the next major step of a needs assessment is to determine the interest and willingness of employees in the target population to use certain services or to participate in particular programs. This is often accomplished through surveys of employees' prior experiences, current attitudes, and future intentions. As was the case for identifying behavioral problems, techniques such as employee sample surveys, key informant surveys, and the Nominal Group Technique can be useful in determining the interest of employees in specific services or programs. There is no simple formula or ideal way to carry out needs assessments. Perhaps, the best overall strategy for conducting a needs assessment is to use multiple sources for data collection and the method of triangulation to reach final conclusions (Attkisson et al., 1978; Campbell and Fiske, 1959).

Setting Priorities. A second step in the program planning process that is critical to subsequent evaluation is setting the priorities among needs to be addressed. Several formal methods exist for clarifying and prioritizing needs. One of these is multi-attribute utility analysis, which is based on a decision theoretic approach to evaluation (Edwards, et al., 1975).

This approach allows the formal explication and ranking of the objectives of different groups. Each group first defines and ranks its objectives and provides information on those that it considers most useful. Then through the use of Bayesian statistics, the choices are analyzed and reported back to the groups. On this basis, the priorities are reordered. The process of providing information, linking objectives to inferences, and reordering objectives is continued until the groups have taken into account their diverse views. The decision theoretic approach is especially useful when the different stakeholders hold sharply conflicting views and the pool of potential objectives is beyond informal reconciliation.

Goals and Objectives. When the main problems and priorities have been defined, the next step is to develop formally the program goals and objectives. A statement of clear and concise objectives is critical both to the implementation and to subsequent evaluation of the program (McLeroy et al., 1984). It is almost axiomatic that evaluation cannot be conducted objectively unless adequate objectives have been developed.

Methods of specifying goals and objectives have been addressed for many years in educational and service programs (Green et al., 1980; Mager, 1962; Rossi and Freeman, 1982; and Weiss, 1972). The basic purpose of goals is to provide the general direction or orientation of the program and that of objectives to map out the specific procedures and methods. Thus, goals typically are stated in general terms while objectives provide details. The amount of detail provided in objectives often sets limits on the quality of evaluation that can be done.

Objectives should be developed both at the program level and at the individual level. At the program level, objectives should specify who (the target population) will change or will receive how much of what health program, behavior or services; and, by when (the expected date or elapsed time required for measurement of the impact or outcome of the services). The specification of the characteristics of the target population and the program should be very routine but often insufficient details are given. Minimal information about target groups should include basic demographics; age, sex, level of education, income levels, and job types, etc. In worksite programs, facts about the specific characteristics of work also should be provided.

Details about programs and services should include the times and places, and the frequency, intensity, and duration of activities. Information also should be provided on the type of personnel or staff involved in the programs. Objectives for impacts and outcomes often are expressed in terms of expected changes in knowledge, attitudes, behaviors, and physiological or biochemical changes. Facts provided on impact and outcome objectives should include specifics on the amount of change expected, the time when the change is projected to occur, and the expected duration or durability of the change.

An important factor in setting behavioral objectives is the choice of quantifiable outcome measures (Green et al., 1980; Sechrest and Cohen, 1980). It often is necessary for evaluators to spend time with program planners in the early stages of the program development to assist them in articulating objectives that are clear, specific, and measurable. Skilled evaluators with a knowledge of the stress field can help in the selection of impact and outcome objectives that meet these requirements.

Most objectives are stated in terms of the "average" change that is expected to occur in the target group. Sometimes it also is useful to complement the statement of objectives for groups by specifying a set of objectives for individuals. The technique of Goal Attainment Scaling (Kiresuk, 1973) allows goals and objectives to be tailored for individuals. It uses relative rather than absolute measures and allows the progress of individuals to be tracked against their own baselines on a number of variables and thus provides a personalized profile. The results of individuals then can be summed to provide a composite estimate of the program impact.

Standards of Acceptability in Program Evaluation

An important early step in planning an evaluation is to consider and decide upon standards of acceptability. In evaluating a program, an object of interest, (usually an impact or outcome measure based on a program objective), is compared to a standard of acceptability. The method of determining whether the object of interest has met a predetermined standard depends on the standard of acceptability selected. Different standards exist against which program effects can be judged. There are both individual and aggregate or group standards.

At the individual level, the acceptable standard of change may be personally defined or defined by professionals. For example, an individual may wish to lower diastolic blood pressure by 5 mmHg without drugs; or the doctor may recommend that a patient must lower diastolic blood pressure by 5 mmHg or must take medications. In either case, the target level of change can be set for the individual and the actual change, within certain time limits, can be evaluated against the personalized standard of acceptability. Also, the technique of goal attainment scaling may be a useful adjunct in establishing individual standards.

At the aggregate level, one or more of five standards of acceptability may be used. A basic description and examples of each of these follows (Green, 1974).

Historical Standards. Current program outcomes are compared to prior program results for comparable persons during a similar time period. For example, if last years' stress management program yielded a 20% reduction in stress-related complaints and symptoms in selected participants, a historical standard of acceptability can be obtained by comparing the results of subsequent programs with last year's 20% reduction.

Normative Standards. Current program effects are sometimes compared with the levels of performance or achievement against regional, national, or international standards. For example, if the object of interest were decreased stress-related symptoms, a suitable standard of acceptability could be a 30% decrease in symptoms reported by employees participating in that program. If this has been shown to be a typical rate of decrease in other stress management programs in industry, it could then be considered a normative standard of acceptability.

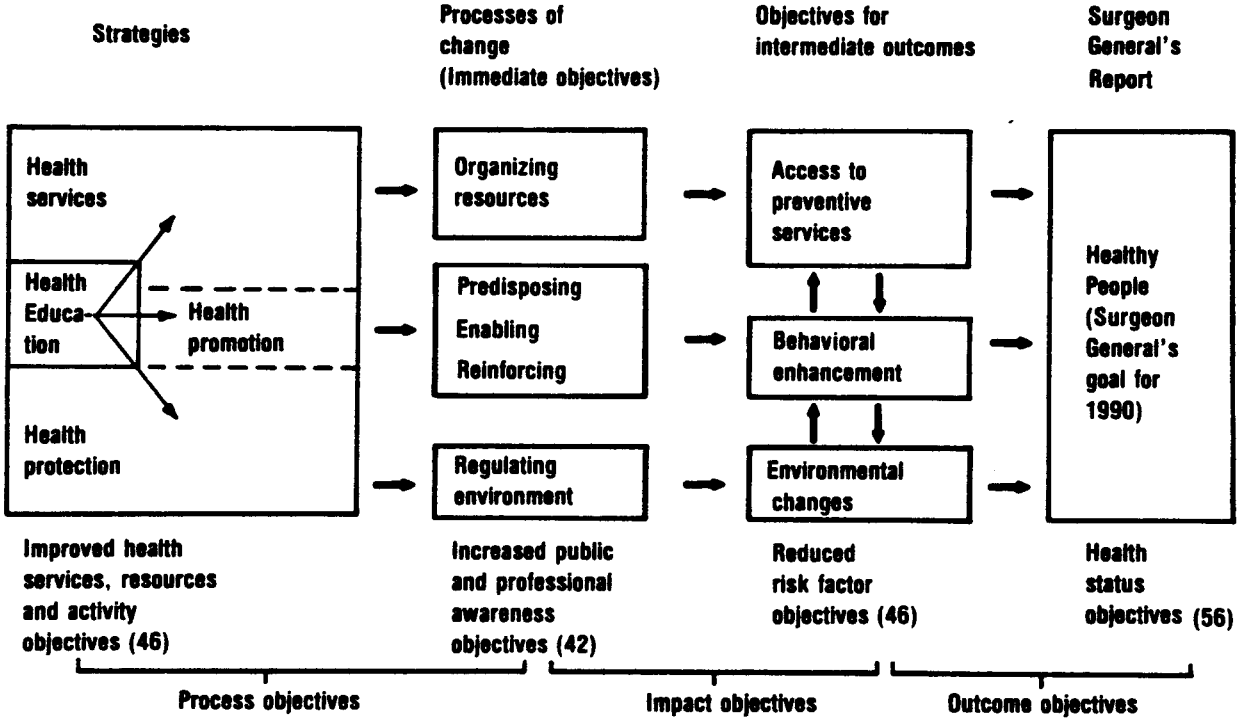
Theoretical Standards. Program outcomes can be compared to a theoretical level expected if everything were to go exactly as planned. A theoretical standard is often based on the results of previous research in which interventions have been tested in controlled laboratory or clinical situations. For example, if a demonstration stress management program, conducted by a university-based team of behavioral scientists using state-of-the-art methods yielded a 50% reduction in stress-related symptoms in a group of management level employees, this could serve as the theoretical standard of acceptability for application of the same stress management methods in the "real world" with other management groups and possibly other employee groups.

Absolute Standards. Program outcomes are sometimes compared to the highest possible level attainable. Whereas theoretical standards are based on the premise that everything will go as planned, absolute standards are often even more unrealistic and may never be possible to attain. For example, a 100% reduction of stress-related symptoms among employees, an example of an absolute standard, is neither realistic nor feasible, and probably even undesirable.

Negotiated Standards. Program criteria usually emerge from the compromise and negotiation of several possible standards. A negotiated standard is frequently an average of the preceding standards of acceptability. For example, if other stress management programs yield a 30% reduction of stress-related symptoms (normative), historical standards for this company are approximately 20%, the theoretical symptom reduction for your population is 50%, and the

Figure 6.1

Figure 1. Structure and logical relationships of the Objectives for the Nation in disease prevention and health promotion



From Green, Wilson, and Bauer, 1983, p. 19

Outcome Evaluation. In the outcome evaluation of preventive medicine programs the main objects of interest usually are morbidity and mortality. In the case of morbidity, the years of productive life and length of survival following detection and the treatment of the conditions also are important variables. Stress management program outcomes also may be expressed in terms of work-related variables. Stated in their most succinct form, the standards of acceptability are cost/benefit estimates, where the benefits may be stated in company savings or profits, or stated in their most humanistic form, improved quality of life of workers.

Currently, improved evaluation of stress management programs is needed at all three of these levels. Unfortunately, generally accepted criteria and standards do not exist against which to judge the qualifications of providers. Also, standards for assessing the methods and procedures of programs have not been developed. Thus, considerable work needs to be carried out to strengthen the measurement of process and to obtain consensus on the standards of acceptability.

A substantial amount of work has been done in the assessment of the impact, i.e., short-term effects, of programs (Murphy, 1984; McLeroy et al., 1984). Nevertheless, several problematic issues plague evaluation at this level. Some of these will be discussed at greater length in the section on measurement. The relationship of the short-term effects have not been related clearly to more long-term indicators. Very little work has been done on estimating cost-benefits and cost-effectiveness. Improved measurement and documentation of impact on knowledge, attitudes, and beliefs, and especially behavioral and environmental changes related to outcomes is necessary. Thus, more work on impact evaluation is needed.

Both scientific and financial barriers limit the likelihood of good outcome evaluation of stress management programs in the near future. The evidence linking particular sources of environmental stress and personal coping behaviors to short- and long-term indicators of work performance and health is not strong. Also, very few reports on good comparative short-range or impact studies have been published. Finally, clinical trials large enough to link stress and stress reduction interventions to work performance, and especially to morbidity and mortality, would be expensive. No funding mechanism has offered to support such costly, large-scale, long-term studies.

Evaluation Designs

Many designs are available for use in program evaluation. The most appropriate design depends on the logistical circumstances of the program and the available resources. The decision to use a specific design for evaluation should be based on several considerations. An estimation both of practical importance and potential scientific value

absolute standard is complete reduction of symptoms (100%), then a negotiated standard could be 35%. That is a weighted average of the other standards that gives greater weight to historical and normative standards than to theoretical and absolute standards.

Model for Planning Program Evaluations

A model that can be used to assist in the planning and evaluation of worksite health promotion and stress management programs is presented in Figure 6.1 This model was developed to help conceptualize the overall plan and main strategies for achieving and evaluating progress toward the objectives for the nation in disease prevention and health promotion (Green, et al., 1983). In the model, three levels of objective--process, impact, and outcome--are specified. There are levels of evaluation that correspond to these three levels of objectives.

Evaluation models, monitoring systems, and specific data collection techniques, for each of these levels must be selected and implemented in order to track progress toward the final or outcome objectives. In subsequent sections, levels of evaluation, general evaluation models, and measurement methods appropriate for stress-reduction and management programs in worksettings will be discussed.

Evaluation efforts can be focused on one or several levels of program objectives. By convention there are three basic levels of evaluation. These are process, impact, and outcome, and each one is treated briefly in the following discussion.

Process Evaluation. In process evaluation, the object of interest is professional or management practice and the delivery of services. The standard of acceptability is appropriate conduct of practice. Common methods of evaluation include quality assurance mechanisms such as peer review, audit, accreditation, certification, and government or administrative surveillance. Informal and formal feedback from service providers and participants also are used.

Impact Evaluation. Impact evaluation typically focuses on the immediate effects of the program on knowledge, attitudes, and behavior of participants. This evaluation is concerned, then, with the more immediate, short-term, goals of the program. Knowledge, attitudes, and other predisposing, enabling, and reinforcing factors influence behavior that relates to reduced exposure to risks, reduced delay in use of preventive health services, and decreased lead time in diagnosing and treating disease. The most widely comparable standard of acceptability against which to evaluate an impact is cost-effectiveness because it uses a common metric in the numerator (dollars) and a common denominator (unit of impact) (Green et al., 1980).

of the study should be made. When research is a major emphasis the main threats to the validity of the conclusions that might be drawn from the design must be considered (Cook and Campbell, 1979; Green and Lewis, 1986). A sound choice of designs also requires recognition of the practical, ethical, and financial constraints on the conduct of the study. For example, issues of informed consent and denial of services to control groups put constraints on many potential evaluation activities.

There are some basic procedures that can always be used. Other very elaborate designs can be used only when logistics are favorable and substantial funding is available. By adding successive elements to the basic procedures, it is possible to increase progressively the level of internal validity or rigor and also the level of external validity or generality. Elements of evaluation designs will be discussed in the next section.

Six different evaluation designs are listed below. These designs increase in complexity and cost of implementation from 1-6. The historical and inventory approaches are basically bookkeeping techniques, the comparative and controlled comparison approaches allow effectiveness estimates, and the controlled experimental and full-blown evaluative research project allow causal inferences and generalizations to be made with maximum assurance. Examples of worksite based health programs that were evaluated through the inventory and other approaches may be found in Green and Lewis (1986) and Parkinson et al. (1982).

1. Historical, Record Keeping Approach

2. Inventory Approach

3. Comparative Approach

4. Controlled Comparison, or Quasi-experimental Approach

5. Controlled Experimental Approach

6. Full-Blown Evaluative Research Project

Historical Approach. When an evaluator sets up a continuous record-keeping procedure to accumulate data and then periodically charts the data to determine if change is occurring, a historical standard of acceptability has been applied. The frequency of data collection depends on how often the events that are being recorded occur. This very basic approach generates data that can be presented in charts and graphs to demonstrate how the program is doing. Collecting and charting data in this way provides periodic benchmarks

against which to compare both previous and future program efforts. The rates of problem indicators can be plotted against program inputs over time and be presented as time-series graphs or frequency polygons.

Inventory Approach. Source data cannot be collected continuously. An evaluator must collect data at specific intervals and compile them at specific points in time--at least at the beginning and end of the program. Target dates for interim assessments can be set, expected outcome levels must be identified, and observations made or sample surveys performed. For some type of programs, the critical measurement points have been standardized (e.g., smoking cessation at 1 1/2, 3, 6, and 12 months). These intervals also would be applicable to most stress management programs.

Comparative Approach. The standard of comparison can be the results of programs completed in other settings. It is therefore necessary that the evaluator identify similar programs carried out in other settings and then borrow or buy the standardized instruments for collecting data. Comparative evaluations between companies also can be done if standardized methods and procedures are adopted. Thus, use of standardized procedures allows comparisons both of results obtained in other settings and with future results in the same company. Data from a particular program also can be compared with national data. Again, such normative comparisons are greatly facilitated if standardized instruments are used for data collection whenever possible.

Controlled Comparison, or Quasi-experimental Approach. When the evaluator identifies a population for comparison that is similar to the target population but is not receiving a stress management program, the quasi-experimental design is applied. The historical or inventory method is then applied both to the target population and to the comparison population, which are then periodically compared. This approach reduces some of the threats to internal validity that weaken the two prior designs.

Controlled Experimental Approach. This approach is comparable to the clinical trial in medical research. The evaluator establishes a formal procedure for randomly selecting the persons within the study population who will participate in the experimental stress management program and those who will not, a control group. Use of this approach requires a situation in which it is possible to deny the program to some individuals. The evaluator collects identical data at similar intervals in both the experimental and control groups and tracks their progress over time.

Full-blown Evaluative Research Project. This approach is not feasible for most worksite based stress management programs. In this design the strategies from the controlled experimental approach are applied

within one worksite population. Two or more groups are randomized to systematically varied combinations of program elements, and multiple measurements are obtained. Each group receives a different mix of stress management interventions (e.g., group A, relaxation training alone; group B, relaxation training plus biofeedback; group C, biofeedback training alone; group D, no program). Such designs have been used in the evaluation of several stress management programs (Murphy, 1984; McLeroy, et al., 1984).

Selection of Evaluation Measures

The selection of specific measures both of individual and of organizational characteristics is a critical step in program evaluation. Making decisions about what to measure is often neither simple nor straightforward. There is a large increase in the number of potential, relevant, variables when one moves from the field of basic research to that of program evaluation. Furthermore, the selection of variables and measurement strategies in the stress field is particularly difficult since there is a very large pool or potential measures to choose from. Two general criteria, relevance and feasibility always should be considered in the selection of measures.

Relevance. Relevance is the first factor that should be considered in the selection of measures. A measure can be considered relevant to the extent that it either measures directly a specific object of interest or behavioral objective or provides a good approximation of it.

The selection of a particular measure or set of measures should reflect a balance between the major objects of interest of the sponsors and recipients of the program and evaluation, the standards of acceptability that they are willing to apply to those objects, and the criteria of ethical and scientific merit that can be applied to the objects. The objects of interest may be one or more elements of process, impact, or outcome as shown in Figure 3.

The issue of relevance often is decided in the needs assessment and objective setting phases of a program. The specific objectives of programs often largely determine what is measured. When objectives are poorly conceived and loosely stated, they provide little guidance for the selection of measures. Thus, time and money can be wasted by placing emphasis on the detailed measurement of variables that have little relevance to goals and objectives. However, if program objectives are well developed and clearly stated they usually direct attention to the general factors and sometimes the specific variables of greatest importance. Therefore, the precise statement of objectives is critical to the selection of the variables to measure and monitor in programs.

Relevance, however, is a highly subjective factor and depends on the views of major shareholders or stakeholders in the program. The issue of the relevance of measures has been analyzed in terms of the different needs and priorities of administrators, researchers and clinical perspective (Green et al., 1980).

Often, there are several identifiable groups, sometimes with conflicting views, who have an interest in program design and outcomes. In the case of stress management programs, management, labor unions, clinical practitioners, and researchers or evaluators, all have different interests and sets of priorities. Therefore, what is relevant to one group may be much less relevant to another. Failure to consider the relevance of measures for different groups affected by a program can seriously compromise the program outcomes and usefulness of the evaluation results.

Thus, in addition to the scrutiny of program objectives, it is sometimes important for evaluators to distance themselves from the major assumptions of the program sponsors and to analyze the theoretical or conceptual framework and the particular biases that guided the program development. A critical analysis of the theoretical framework, and political-economic rationale for a program can suggest additional variables that may not have been specified in the objectives. These may be highly relevant when considered in a broader social and ethical framework.

Feasibility. Feasibility refers to the practical issues of making measurements and obtaining data. Some of the basic factors that affect feasibility are access to the data, technical expertise (ability to make measures), cost of making the measurements (equipment, personnel and timecosts to company and employees), and ability to track participants over time. All these factors need to be given some consideration in selecting measures. Feasibility should not, however, be equated with appropriateness of measurement. Unfortunately, in many stress management programs, measures have been chosen primarily because they are inexpensive and easy to use. Changes in these variables may have very little clinical, economic, or scientific significance.

Technical Features

In addition to the previously discussed general criteria, there also are several technical features of measurement techniques that should be considered. Three important criteria are level of measurement, reliability, and validity.

Measurement, by definition, is the assignment of labels or numbers to objects, events, or persons according to specified rules. Measurements require first, specification of the objects to be measured, second, the labels or numbers to use, and third the rules by

which the labels or numbers are assigned to objects. In program evaluation, measurement refers to the systematic procedures applied to the objective quantification of needs, processes, impacts, and outcomes.

Levels of Measurement. An understanding of levels of measurement is necessary to determine how the various forms of measurement set limits on the statistical procedures that can be used in the data analysis.

By convention, there are four levels of measurement: nominal, ordinal, interval, and ratio. The nominal level is considered the lowest, ordinal and interval intermediate, and ratio the highest. These four levels of measurement along with their definitions, and a summary of some of the statistical tests that can be applied at each level are presented in Table 6.1. A number of books provide detailed discussions of levels of measurement and their characteristics (Green and Lewis, 1986; Siegel, 1956; Windsor et al., 1984).

From a technical perspective, it is preferable to select data collection techniques that allow the ratio, or highest, level of measurement to be used. This maximizes the ability to distinguish between background noise or variance and specific treatment effects. Higher levels also permit use of a wider range of statistical tests and more powerful statistical procedures in the data analyses. Using sophisticated statistical tests enhances the likelihood of detecting program-specific effects and distinguishing them from non-program effects.

Regardless of how creatively designed, well controlled, and smoothly executed an evaluation design is, it is only as good as the measures from which data are derived. Inappropriate and inadequate or "noisy" measures will impair and can totally compromise the quality of the most elaborate and expensive evaluation. Thus, careful attention should be given to measurement instruments and techniques.

Reliability and Validity

Accuracy in measurement is traditionally viewed as a combination of two separate issues, reliability and validity (Bernstein, 1976).

Reliability. As generally used, reliability refers to the extent to which an instrument is consistent. It is important, however, to distinguish clearly between reliability as relative freedom from error and stability.

Reliability coefficients are affected by the variance of the scores upon which the correlation coefficient is based, and the reliability of an instrument typically increases with the homogeneity of scores. However, a reliability coefficient is as much a function of the population being assessed and the conditions under which the

TABLE 6.1 LEVELS OF MEASUREMENT AND EXAMPLES OF NUMERICAL PROCEDURES APPROPRIATE TO EACH.

Levels of Measurement	Comments	Permissible Numerical Procedures and Statistics
Nominal	Categorizes subjects, events, or objects. Numbers assigned have no numerical meaning - i.e., cannot be added or rank ordered.	1. Frequency counts 2. Frequency statistics - e.g., chi square, percentages, contingency coefficients[b]
Ordinal	Rank orders subjects, objects, or events. Numbers indicate rank but not absolute numerical quantity. As such, numbers cannot be added or subtracted.	1. Frequency counts 2. Frequency statistics - e.g., chi square percentages, contingency coefficients 3. Ranks determined 4. Rank-order measures - e.g., rank-order correlation coefficient; Kendall's W
Interval	Distances between number are assumed to be equal in size. Intervals can be added or subtracted but only with understanding that it is intervals, not absolute numbers, that are involved in computations.	1. Frequency counts 2. Frequency statistics - e.g., chi square, percentages, contingency coefficients 3. Ranks determined 4. Rank-order measures - e.g., rank-order correlation coefficient; Kendall's W 5. Summated ratings 6. Mean, t-tests
Ratio	Scale has an absolute or natural zero point that possesses empirical meaning. Numbers on scale represent the actual amount of property measured. Numbers can be added and subtracted as well as divided.	1. Frequency counts 2. Frequency statistics - e.g., chi square, percentages, contingency coefficients 3. Ranks determined 4. Rank-order measures - e.g., rank-order correlation coefficient; Kendall's W 5. Summated ratings 6. Mean, t-tests 7. Analysis of variance 8. Correlation analysis and all other parametric tests

[a] Levels are ordered from the lowest (nominal) to the highest (ratio) level of measurement.
[b] For a detailed discussion of statistical assumptions underlying nonparametic statistics see Siegel (1956).
[c] The list of statistical operations is suggestive, not exhaustive.

From Green and Lewis (1985).

instrument is administered as it is a function of the psychometric qualities of the instrument. Such variations in testing or measuring as enthusiasm of the tester, motivation of the respondents, and even room characteristics such as temperature and humidity can all affect reliability coefficients. Reliability is most usefully conceptualized as a set of statistical, and situational conditions which affect the error in the stability of data gathered by a given instrument (French and Kaufman, 1983).

Validity. On the other hand, validity is the accuracy with which a measurement instrument or procedure measures what it was intended to measure. It is possible to have a highly reliable instrument that is measuring the wrong impact or outcome. Validity only can be determined by obtaining independent measures of the same impact or outcome and comparing the results (Campbell and Fiske, 1959; Green and Lewis, 1986; Windsor et al., 1984).

Measures of Stress

In the stress field, it is a much easier task to provide general criteria for measurements than it is to suggest specific variables that should be chosen. Before providing suggestions for the selection of specific measures of stress at both the individual and organizational level, it may be instructive to look briefly at why this process is so complex.

First, no satisfactory definition of stress or specific goals and objectives for stress identification and reduction were proposed in Healthy People: The Surgeon General's Report on Health Promotion and Disease Prevention (USDHEW, 1979), or in Promoting Health/Preventing Disease: Objectives for the Nation (USDHHS, 1980b). The Institute of Medicine (IOM) of the National Academy of Sciences subsequently completed a status report on the relationship of stress to health (Elliott and Eisdorfer, 1982). On the basis of this comprehensive task force report, it was decided that it was not feasible to get a concensus on a general definition of stress. Instead, a conceptual model for the study of stress was proposed. In this model, the stress concept was divided into four major domains: 1) stressors or sources of stress, 2) reactions, 3) consequences, and 4) mediators.

The model and each of the four domains of variables is discussed in detail in the IOM Report. In principle, variables should be looked at in each of the domains. This would provide the most complete perspective on the nature of stress-related problems and the dynamics of intervention program process, impact, and outcomes. However, in practice this generally is not feasible as most stress management programs rarely have large budgets for research and evaluation. Therefore, a great amount of selectivity must be used in choosing variables. The following discussion will focus on measures of sources of stress and reactions to stress since these have been studied most carefully and are feasible to measure in low budget programs.

Sources of Stress. Sources of stress and potential solutions to stress-related problems have been identified at both the organizational and individual levels. Sources of stress at the organizational level can be divided broadly into categories of physical and psychosocial stressors.

Organizational Level. Different types of physical stressors have been studied carefully in the laboratory and have been identified in work settings. They have been discussed in many reviews (USDHEW, 1978a, b; USDHHS, 1980a; Holt, 1982; Neale et al., 1983).

Holt has listed five categories of physical properties of working environments that can be sources of stress. These are 1) physical hazards, chronic dangers, 2) pollution, less immediate dangers, 3) extremes of heat, cold, humidity, and pressure, etc., 4) noise, and 5) bad man-machine design. Shift work is another physical property of jobs that may be a source of stress for many workers.

Most of these sources of physical stress can be measured objectively, and with a high degree of reliability and validity. In some cases, standards exist which can be used in the regulation of the levels of some of these stressors. However, individual tolerances vary widely, so the performance of employees should be observed, and subjective, self-reports of workers about the aversiveness of these factors should be obtained. Given the initial differences in tolerances among individuals for potential sources of physical stress, and differential abilities to adapt, multiple sources of input must be obtained to establish the degree of stressfulness of these physical factors.

Over the past 20 years, much progress has been made in identifying sources of psychosocial stress in work environments. A number of properties of system design and job content appear related both to job satisfaction and to health (Elliott and Eisdorfer, 1982). Those that have been studied most carefully include the following:

- o. Quantitative overload: too much to do, excessive time pressure, or repetitious work flow in combination with one-sided job demands and superficial attention.

- o. Qualitative underload: too narrow and one-sided job content, lack of stimulus variation, no demands on creativity or problem solving, and low opportunities for social interaction.

- o. Lack of control: especially in relation to pace of work and working methods.

- o. Lack of social support: inadequate social networks with fellow workers and lack of support from supervisors.

Several of these organizational characteristics appear to interact synergistically to impair mental and physical health. For example, Swedish workers with high work loads and low control over the work were found to have higher rates of morbidity and mortality than workers with moderate loads and higher control over the work situation (Ahlbom et al., 1977; Karasek, 1979; 1981). The high-load, low-control workers showed more symptoms of excessive fatigue and depression, and higher rates of cardiovascular disease and overall mortality.

Besides the fact that these four job dimensions have been linked through epidemiologic studies to mental and physical health problems, another advantage in their use is that they can be measured both objectively and subjectively. Most assembly line or production jobs, and many blue collar jobs, can be rated independently by outside observers for characteristics of overload and underload and level of control. Thus, rough "objective indices" of the stressfulness and the relative risk for health problems of different occupations can be developed. However, since the demands of many jobs are dynamic rather than static, these general indices should be supplemented with surveys of the ratings of employees in specific companies on these job characteristics.

Other potential sources of psychosocial stress in work settings are role related factors such as role-ambiguity, conflict, and strain. Poor person-environment fit (PE-fit) is an additional possible source of stress. While these measures have been used frequently in the past, information on them can be collected only through subjective, self-reports. Thus, job types cannot be classified independently in terms of these factors. Furthermore, while these factors have been associated with conditions such as job dissatisfaction, they have not been found to be strongly predictive of mental or physical illness. Despite recent criticisms of the utility of these variables in stress research (Baker, 1985; Kasl, 1984), they still may be useful if their limitations are recognized. Jenkins et al. (1984) have provided a review and discussion of many of the available instruments to measure role-related factors and person-environment fit. This review provides a brief description of the scales and information on their reliability and validity.

Individual Level. Potential sources of stress also can be identified at the individual level through the study of personal characteristics and patterns of social interaction. Personal factors that may increase the likelihood of stress at work include, anxious-tense personality, low self-esteem, Type A personality, poor communication skills, poor assertiveness skills, and minor and major forms of psychopathology that interfere with technical work performance and social interactions. Factors outside of work that may increase the likelihood of job stress include alcohol and other drug problems, poor nutrition and lack of exercise, financial and legal problems, and social and family problems.

Given the range and complexity of many of these personal factors that may increase the likelihood of work-related stress, there is no easy way to define and isolate individuals' susceptibility. Thus, several different approaches should be used to identify the type and severity of stress producing factors at the individual level.

When the purpose is to screen a relatively large number of employees who are functioning reasonably well, self-report inventories may be used. A number of life events scales are available which allow estimates to be made of the amount of stress that persons are under. Some of these scales are reviewed by Jenkins et al. (1984). Also, an instrument is available to measure more proximal and frequently occurring daily hassles (Kanner et al., 1981).

Since the specific types and general pattern of major life changes and hassles that can occur at work may differ greatly for different occupations and companies, questions may need to be specifically tailored. Often this will require the development and use of semi- or unstructured techniques. Martin (1983) has reported on the use of semi-structural techniques in obtaining data on stress in the graphic arts industry. The advantage of these techniques are their flexibility and ability to provide details on specific problems that are sources of stress to employees on particular jobs.

Baseline information on physical health habits can be collected conveniently and relatively expensively through the use of self-report, health-risk appraisals. Information on the availability and characteristics of a large number of health risk appraisals has been summarized in a recent publication (Green and Lewis, 1986). Details and a discussion of the prospects for the use of health hazard appraisals are provided in a recent technical report (Breslow et al., 1985).

Reactions to Stress. Sources of stress have been implicated in psychological, behavioral, and physiological/biochemical changes. There is strong empirical documentation and a voluminous literature on short-term reactions to sources of stress (Cincirpini et al., 1984; Stainbrook and Green, 1983). However, while it has been suggested that stress contributes to long-term mental and physical health problems the strength of the relationship between individual stressors or collective indices of stress and these health status indicators is not strong (Baker, 1985; Kasl, 1980, 1984).

Some of the psychological, behavioral, and psychosomatic problems that commonly have been associated with stress are presented in below:

Psychological	Behavioral	Psychosomatic
Anxiety	Smoking	High blood pressure
Depression	Alcohol use	Tachycardia
Anger	Drug use	Headaches
Low job satisfaction	Disturbed relationships	Ulcers
Low self-esteem	Violent behavior	Sleep problems

Several different psychometric instruments have been used to assess the levels of stress-related psychological symptoms in populations of workers. The Symptom Checklist 90 or SCL-90R often has been used and has scales for anxiety, depression, and anger and is a good general screening instrument (Derogatis, 1975). Anxiety frequently has been measured with the Spielberger State-Trait Anxiety Scale (Spielberger et al., 1968). This instrument, particularly the state component of the scale has been found to be a sensitive indicator of stress in non-clinical and work populations. Many self-report schedules can be used to measure job-satisfaction. Jenkins et al. (1984) have reviewed those instruments most commonly used to measure job satisfaction. Self-esteem also can be measured with a number of different questionnaires (Gilberts, 1983).

Behaviors that may be stress related can be screened with health hazard appraisals or with instruments designed to obtain extensive information on the specific behaviors, e.g., surveys of smoking, alcohol use, and drug use, etc. Data on disturbed relationships and violent or aggressive behavior can be collected through interviews with the employee, fellow workers, and family members, and sometimes through observations.

Information on many psychosomatic problems can be collected with self-report instruments like the SCL-90R or more informal checklists. If resources are available, the validity of self-report information can be checked through medical record searches. Also, in the case of tachycardia and high blood pressure direct measurements can be made.

It also can be helpful in the selection of measures to examine carefully the indicators of stress that have been used in prior stress management programs. The impact and outcome measures used in many

prior studies have been summarized in three recent reports (Chen, 1984; McLeroy, et al., 1984; and Murphy, 1984). The dependent variables that were measured most frequently in the 19 prior stress management programs reviewed by McLeroy et al. (1984) are presented in below:

Variable	Frequency
Anxiety (trait=5, state=1)	6
Muscle tension	5
Stress symptoms	5
Perceived job stress	5
Blood pressure	4
Hand temperature	3
Job satisfaction	3

Of these seven variables, four (anxiety, stress symptoms, perceived job stress, and job satisfaction) are subjective measures that depend entirely on self-reports. The three other variables (muscle tension, blood pressure, and hand temperature) all can be measured objectively by independent observers with monitoring equipment. Typically, it is preferable to choose variables that can be measured objectively. However, the other general criteria that were discussed previously, namely, relevance and feasibility also should be given careful consideration, as well as two other factors, sensitivity and representativeness.

Sensitivity to Change. An additional issue that should be considered is the likelihood of effecting reductions in the levels of the impact and outcome indicators through the types of programs being offered. The results of the stress management studies reviewed by Murphy (1984) and McLeroy et al. (1984), suggest that most of these variables are sensitive and can be reduced acutely by relatively low-cost programs. However, the long-term health implications of these changes and the cost-benefit and cost-effectiveness of most stress management programs have not been studied.

There are other variables that would be relevant outcome measures for some stress management programs and would allow better cost-benefit and cost-effectiveness estimates than most of the previously discussed measures. These include absenteeism, turnover, health services expenditures, health insurance claims, and disability and workers' compensation payments. To date, very little study of these variables has been done in relation to stress management programs. The issues of cost-benefit and cost-effectiveness analysis will be considered in the next section.

Representativeness. Another issue that should be looked at carefully in the selection of measures is that of representativeness. Representativeness is a term that typically is applied to the type of sampling techniques that are used, but it also applies in the selection and use of measures. Measurements of some variables have limited clinical significance unless they are representative, or sample important domains of the environment and individual behavior. Sometimes increasing the representativeness of measures requires substituting informal data collection techniques for more formal or standardized measures. This means that some reliability may have to be sacrificed. However, there can be significant gains in validity.

Standardized psychometric scales often are given only to persons before and after programs. While they are useful for general screening purposes, they often do not provide accurate data on the frequency and intensity of moods or symptoms at work and do not reflect how much they interfere with work performance. Since most standardized scales cannot be given frequently or during work, they should be or supplemented by less formal but more frequently administered measurement techniques such as daily stress logs. The use of daily stress logs or diaries allow employees to record sources of stress and their reactions to them both at work and at home. Persons also can rate the severity of symptoms and estimate the amount of work time lost for specific stressors and their reactions. This allows rough estimates of the costs of stressors and stress reactions and thus provides a baseline against which the cost effectiveness of programs can be calculated (Manuso, 1983, 1984).

While physiologic variables often can be measured with a high degree of accuracy and precision, they frequently are not representative. Thus, their utility as clinical predictors of impact and outcome is highly limited. For example, blood-pressure level often is used as a dependent variable in stress management programs. Usually, blood pressure is measured only in a clinic and not in the work environment. The measures in the clinic may not accurately reflect the blood pressures at work and may be less predictive of hypertension (Sokolow et al., 1980; Pickering et al., 1982). Failure to obtain representative measures may result in false positives or false negatives -- both diagnostic errors -- and inaccurate estimates of program effectiveness. Thus, while stress management programs may lead to reductions in pressures taken in the clinic they may have little impact on the pressures during work. Many of the stress management training studies on treatment of essential hypertension have been criticized for their failure to demonstrate the generalizability of blood pressure lowering. The representativeness of blood pressure measurements can be improved by training individuals to take their own blood pressures in home and work settings and through the use of automatic-monitoring devices (e.g., Bertera and Cuthie, 1984; Pickering et al., 1982).

While knowledge of previously used variables and the impact of programs on these indicators is useful, these variables should not be tacitly chosen in subsequent studies. Frequently, systematic biases exist in the ways that program objectives are set and measures are selected. For example, most of the programs reviewed by McLeroy et al., and Murphy strongly reflect the biases of clinicians, behavioral researchers, and management. Most of these programs might best be termed symptom-reduction programs. In many cases, the variables that are measured are not relevant to participants. Furthermore, the exclusive emphasis of these programs on the reductions of physiologic reactions to stressors and stress-related symptoms has recently come under criticism from organized labor. The primary complaint has been that the exclusive emphasis on individuals represents a strong psychological bias and avoids management responsibility for environmental factors. It is pointed out that in most cases no attempts are made to identify and minimize sources of stress in the workplace (Lerner and Shore, 1982; Neale et al., 1983; Tesh, 1983).

The preference for physiological indices reflects other biases. First, many of the measures have been standard measures in laboratory-based research; thus, they have an aura of credibility or scientific merit. Second, most of them can be measured with physical techniques and expressed as interval or ratio measures. Third, most of the measures are highly reactive and are subject to being reduced with short-term interventions. Therefore, the chances of getting a positive effect are good which increases the chances for publication of the findings. The relationship of short-term reactions to stressors to longer-term, outcome measures has not been firmly established. The findings thus may be of interest and may serve a purpose in a narrow research or academic context but may have little practical, clinical, or administrative meaning outside these contexts.

Critical Issues in Program Evaluation

Cost-Benefit and Cost-Effectiveness. Cost-benefit analysis and cost-effectiveness analysis are logical extensions of evaluation research. The procedures are based on the assumption that judgments about either costs or benefits of programs cannot be made without relating them to each other. In order for the value or merit of a program to be determined either in dollar amounts (cost-benefit analysis) or in relation to available alternatives (cost-effectiveness analysis) there must be some evaluative evidence, i.e., impact or outcome results (Weiss, 1972; Rossi and Freeman, 1982; French and Kaufman, 1981).

These analyses can provide valuable information to program managers and policy makers in a variety of ways. Some of their major uses are summarized below:

- o To account for the use of public and private funds
- o To compare the efficiency of the operation
- o To compare the cost of alternative services or programs
- o To determine allocation or reallocate of resources.

In some cases government organizations are the primary source of funds for prevention programs. In most cases, the costs of stress management programs have been borne by individual companies. Cost analysis can answer questions regarding the efficient use of resources and the optimum size of a program, etc. Cost analysis also can be used to compare alternative methods of providing services or programs. With cost data on alternate methods for providing prevention services, the analysis seeks to identify the least costly program alternative that can accomplish the desired objective. Information on the costs and effectiveness of alternative programs or methods within a program can help program managers to modify or improve the process of the program or to reallocate resources to alternative programs.

Cost-benefit Analysis. Cost-benefit analysis reduces all outcomes to monetary terms. This allows the direct comparison of programs with different outcomes. In order to conduct cost-benefit analyses several things must be known. First, it must be possible to estimate the effectiveness of the programs. This step requires estimating the reduction in morbidity and mortality; calculating the direct cost of the treatment or other services that may be averted; figuring the indirect value of income which would be lost if the person continued to exhibit certain problems; and estimating the money saved through the reduction of the problems. Second, the operational costs of running the program must be determined. Third, a monetary value must be placed on the expected outcomes. When information is available on two or more programs, cost-benefit data allows comparisons between programs. Given the right set of conditions, the use of cost-benefit analysis is very useful. However, in many disease prevention/health promotion programs there is insufficient information available to carry out this type of analysis. Methods and procedures for conducting cost-benefit analyses have been discussed in detail (Green and Lewis, 1986; Rossi and Freeman, 1982; Shepard and Thompson, 1979).

Limiting Conditions. Often the access to medical records and insurance claims forms is limited; the benefits of health programs cannot be defined easily in concrete terms; and even when outcomes can be clearly defined and measured, and usually it is difficult to place dollar amounts on them. In the case of stress, the epidemiologic evidence linking specific sources of stress to particular long-term effects or outcomes is not very strong. The best case has been made for its association with elevated blood pressure. While there is evidence that stress management programs can have short-term impacts on blood-pressure and other physiologic variables, there are few long-term data. Thus, their effect on morbidity and mortality is not known.

Another complication is that ethical problems may arise in the use of cost-benefit analyses. For example, when employee preferences are weighted heavily in the calculation of value, more persons may be willing to pay for and/or use personal time to attend weight loss classes rather than to participate in smoking cessation and alcohol treatment programs. This presents a problem in decision making because the likelihood of success and potential payoffs of the latter are greater. Also, when earnings are used to calculate the indirect costs of absenteeism and lower productivity related to symptoms of stress, those with the highest incomes such as top management and professionals will be assigned a higher value than clerical and blue collar workers. Typically, most stress management programs have been offered first to management and possibly later to clerical and blue collar workers. Thus, cost-benefit or pay-off projections already have been operating in decisions about allocation of resources.

Given the number of limitations in conducting cost-benefit analyses of most health promotion programs, it has been suggested that cost-effectiveness analyses may be more appropriate for most of these programs at their current stage of development and cost-benefit analysis should be limited to valuation of immediate outcomes (Green and Lewis, 1986).

Cost Effectiveness. In contrast to cost-benefit analysis, cost-effectiveness analysis does not require that program impacts or outcomes be reduced to monetary terms. Instead, the respective costs of a number of alternative strategies or programs for achieving a desired end are compared. Therefore, cost-effectiveness analysis generally is an easier approach to use in health promotion program evaluation. Its main limitation is that it does not allow the comparison of programs with different outcomes.

There are four basic steps in most cost-effectiveness analyses: 1) definition of program objectives; 2) computation of the program's net monetary costs; 3) definition of program outcomes; and 4) conduct of a sensitivity analysis. As this process has been discussed in detail by

several authors (Green and Lewis, 1986; Rossi and Freeman, 1982; Shepard and Thompson, 1979), it will be discussed only briefly in this chapter. The processes of setting specific objectives and selecting measurable outcomes have been discussed in an earlier section, so will not be dealt with here. Steps two and four will be considered briefly.

Computation of the monetary costs of programs requires that costs to both sponsors and participants be considered. Sponsor costs include staff, space, materials, telephone, postage, etc. Participant costs often include time, transportation, child care, etc. Potential indirect or side effects with associated costs also should be considered. For example, exercise programs carry a small risk of injury that can be estimated from experience. If a program results in such costs they should be included in the calculations. This procedure is sometimes called "risk-benefit analysis."

The final step of a cost-effectiveness analysis is to run a sensitivity analysis. The purpose of this procedure is to vary the basic assumptions, e.g., participation rates, compliance rates, etc., and to figure the worst as well as the best possible cases. This process is discussed in detail by Shepard and Thompson (1979).

Confidentiality. The confidentiality of information in worksite health promotion programs should be given careful attention (Alderman, 1984; French and Kaufman, 1981; Kiefhaber and Goldbeck, 1984). In all worksite based preventive medicine and health promotion programs confidentiality and job security and sensitive issues. Data from insurance claims, medical examinations, employee assistance programs, mental health programs, and health risk assessments are subject to many uses. Some of these may not be in the best interests of employees. Both individual employees and union leaders have expressed concern about the possible misuses of health data made available to employers. There is fear that such information could secretly be used in decisions about job assignment, promotion, or termination. Certain diagnostic labels present more problems than others. For example, drug abuse, alcoholism, and epilepsy, and are feared to be grounds for not hiring, denying promotions, and forcing early retirement or even dismissal.

The acceptance and success of some worksite based screening and intervention programs has depended on the degree to which confidentiality and job security were guaranteed (Alderman, 1984; Masi and Teems, 1983). If guarantees of confidentiality are not provided often programs will not be accepted.

Research and evaluation activities should be guided by high integrity and a strong respect for human dignity. Thus, evaluators should not engage in activities which compromise or infringe on individual rights. In addition to the considerations of ethics at the individual

level, ethical issues also should be considered at the organizational level. Ethical issues in organizational research have been treated in two recent publications (Mirvis and Seashore, 1979; Evaluation Research Society, 1980).

Informed Consent. Prominent among the principles that should guide evaluators is that of informed consent. The principle of informed consent requires that an evaluator secure in advance of the study agreement of all participants in an investigation. This consent is obtained after the potential participants have learned about the nature of the investigation. The issue of informed consent has been addressed previously (ADAMHA, 1975; APA, 1973).

Informed consent has been defined as:

> The knowing consent of an individual or his/her legally authorized representative, so situated as to be able to exercise free power of choice without undue inducement or any element of force, fraud, deceit, duress, or other forms of constraint or coercion (ADAMHA, 1975).

Several basic elements of information necessary to informed consent are:

- A clear explanation of the program, its objectives, and procedures. Identification of any procedures that are experimental.
- An explanation of any risks or discomfort that might occur.
- A description of the benefits that might occur.
- An offer to answer any questions about the procedures.
- A statement that participants are free to withdraw from the program at any time without reprisals.
- An offer to help individuals find alternative services should they wish to withdraw from the study.

Three useful references on the topic of ethics in research are, Ethical Principles in the Conduct of Research with Human Participants (APA, 1973), and the ADAMHA Guide for the Protection of Human Subjects (ADAMHA, 1975), and the reports of the President's Commission for the Study of Ethical Problems in Medicine and Biomedical Research (1981).

Methods of Ensuring Confidentiality. Several general methods can be used to help protect the privacy of individuals. These include establishing policies stipulating that only evaluators will have access to information on individuals; sending the results of screenings and risk assessments to private physicians or employees' homes; and contracting with outside organizations to provide employers only with aggregate information. The success of these methods depends on the degree of confidence that employees have in them. However, providing only aggregate data precludes further analysis at the individual level.

To provide continued access to individual data and still protect privacy, several other methods have been developed. Often a master list of names along with code numbers is set up. The master list is kept in a secure location and all data forms are identified only by the code number. This method has been used successfully in the past but has some limitations. Such records could be subpoenaed as a part of a legal proceeding. As there is no complete guarantee of anonymity, employees may distort responses to questions about alcohol and drug use and other behaviors that may be strongly incriminating or socially unpopular.

Another way to protect anonymity and possibly reduce response bias on sensitive questions is the use of random response techniques. This approach protects the anonymity of the question rather than the respondent. In one of the simplest models, two questions are presented--the sensitive question and an innocuous question for which the probability of response is already known. Respondents are asked to chose a question by flipping a coin, and then to respond without letting the interviewer know which question is being answered. Given prior knowledge of the probabilities of question selection and responses to the innocuous question, the proportions of group responses to the sensitive question can be estimated reasonably accurately. A limitation of the random response techniques is that they require large sample sizes since the obtained variance is a function of the proportion of the sample responding to the sensitive question rather than the entire sample (French, 1979; Fox and Tracy, 1980).

The success of worksite based hypertension programs depends critically on steps taken to ensure confidentiality and job security. Such measures can range from excluding management from the premises during screening and treatment to preventing the release of patient records without the written consent of the employee. This guarantees that only program sponsors and health care personnel know which employees are hypertensive and are familiar with their progress in therapy (Alderman, 1984).

Masi and Teems (1983) reported on the development of an evaluation system designed to assess the effectiveness of the Employee Counseling Service (ECS) at the U.S. Department of Health and Human Services. They noted that the issue of confidentiality was the most critical in the development of the evaluation system. Every data collection form and procedure had to be closely examined for its compliance with all privacy and confidentiality regulations.

There often is a delicate balance between protecting individual rights and privacy and protecting the best interests of companies. This issue will become more controversial as greater emphasis is placed on screening, risk reduction, and cost containment.

Debate continues over whether employers should have direct access to the medical records and health information on individual employees. Benefits to employers include better selection and placement of employees, an improved ability to spot unusual occurrences of illness and environmental risks, and more efficient targeting of resources at high risk problems. Critics accept that the above practices may benefit many employees, but they also argue that access to such information can lead to unfair hiring and discriminatory promotion and termination practices.

It is essential that issues of confidentiality be considered at each step in the development of an evaluation plan. Although strict adherence to the principle of confidentiality may place many obstacles in the way of data collection, this can be weighed carefully against the total loss of credibility of the program if data is accessed unethically.

APPENDIX

Evaluation Resources

Several U.S. Government publications and privately subsidized publications on evaluation are available free or at relatively low costs. Several of these are listed and briefly described. The resource guide developed by Zapka et al. (1982) is a good place to start as it provides a relatively large number of annotated references on evaluation.

Locating Resources for Evaluation:

Zapka, J., Schwartz, R., and Giloth, B. (1982). Locating Resources for Evaluation. Chicago: American Hospital Association.

This is a useful resource guide, written primarily for persons working in the health education and health promotion fields. It provides annotated references on evaluation methods for health education and health promotion programs. References are provided in four topic areas: 1) Evaluation design and implementation, 2) data sources, 3) evaluation instruments, and 4) evaluation management. In addition to listing references, brief discussions of each of the four areas are provided.

The publication is currently available at no cost from:

Center for Health Promotion
American Hospital Association
840 North Lake Shore Drive
Chicago, IL 60611

Baseline is a publication provided by the Health Services Research Center of the University of North Carolina at Chapel Hill with support from the W.K. Kellogg Foundation of Battle Creek, Michigan. The Kellogg Foundation's national demonstration program in health promotion/disease prevention gives special emphasis to careful program evaluation.

To date, ten issues of Baseline have been published. Most of the issues are four to five pages in length and provide clear and succinct discussions of basic issues in health promotion program evaluation. The editors for this series are G.H. DeFriese, and W.L. Beery. The titles for each publication are:

1 Background Information (1982).
2 Cost-benefit and cost-effectiveness analysis for health promotion programs (1982).
3 Choosing an evaluation strategy (1983).
4 Goal-oriented evaluation as a program management tool (1983).

5 Formative evaluation of health promotion programs (1983).
6 On the subject of sampling. (1984).
7 Aids to evaluation: Computers and consultants (1984).
8 Health risk assessment in health program evaluation (1984).
9 Measurement Issues: Reliability and validity (1985).
10 Qualitative Methods in Program Evaluation (1986).

Subscription to Baseline is free and single copies of many of the back issues can still be obtained at no charge. The editors are planning to publish two more issues after which all issues will be updated and edited as a manual or book. Correspondence regarding this publication should be addressed to:

Editors, Baseline
Health Services Research Center
The University of North Carolina
Chase Hall 132-A
Chapel Hill, North Carolina 27514
Telephone (919) 966-5011

French, J.F. and Kaufman, N.J. (Eds.). (1983). Handbook for Prevention Evaluation: Prevention Evaluation Guidelines. DHHS Publication No. (ADM) 83-1145. U.S. Government Printing Office, Washington, DC 20402

This is a well written manual which is practical, provides details and covers most of the basic issues in program evaluation. It was developed specifically as a guide for evaluators of primary prevention programs for mental health and substance abuse problems. Thus, many of the concerns and examples are highly relevant to issues in stress management programs.

The Handbook has been available at no charge from the National Clearinghouse for Drug Abuse Information. The address is:

National Clearinghouse for Drug Abuse Information
P.O. Box 416
Kensington, MD 20795

If it is no longer available from the Clearinghouse, it can be purchased from the Superintendent of Documents. The identification number is 0-410-948. The address is:

Superintendent of Documents
U.S. Government Printing Office
Washington, DC 20402

An Evaluation Handbook for Health Education Programs in Stress Management

IOX Assessment Associates of Los Angeles, California, through a contract with the Centers for Disease Control in Atlanta, has produced a series of seven Evaluation Handbooks for Health Education Programs. One handbook deals specifically with the topic of Stress.

This handbook has a brief introductory section on basic considerations in health education program evaluation. It also provides examples of a large number of both standardized and non-standardized, self-report type questionnaires in the Appendix. Data collection instruments appropriate both for children and adults are provided. The wide range of measurement instruments introduced and described in this handbook should be useful to all educators and others who place a major emphasis on cognitive and affective changes in their stress management programs.

Limitations. The manual does not provide examples of behavioral and physiological/biochemical measures. Another limitation of the manual is that information on the reliability and validity of many of the questionnaires had not yet been obtained when it was published.

All of the Evaluation Handbooks for Health Education Programs are available from the National Technical Information Service. The order number for the Handbook on Stress Management is PB 84-171735 A18, and the price is $31.00.

Correspondence regarding this publication should be addressed to:

National Technical Information Service
U.S. Department of Commerce
Springfield, VA 22161

REFERENCES

Alderman, M.H. (1984). Worksite treatment of hypertension. In J.D. Matarazzo, S.M. Weiss, J.A. Herd, N.E. Miller, and S.M. Weiss (Eds.). Behavioral Health: A Handbook of Health Enhancement and Disease Prevention. New York: John Wiley and Sons.

Ahlbom, A., Karasek, R., and Theorell, T. (1977). Psychosocial occupational demands and risk for cardiovascular death. (Swedish) Lakantidningen, 77, 4243-4245.

Alcohol, Drug Abuse, and Mental Health Administration. (1975). The ADAMHA Guide for the Protection of Human Subjects (Rev. ed.). Washington, DC: U.S. Government Printing Office.

American Psychological Association. (1973). Ethical Principles in the Conduct of Research with Human Participants. Washington, DC, American Psychological Association.

Attkisson, C.C., Hargreaves, W.A., and Horowitz, M.J. (Eds.) (1978). Evaluation of Human Service Programs. New York: Academic Press.

Baker, D.B. (1985). The study of stress at work. Annual Review of Public Health, 6, 367-381.

Bernstein, I.N. (1976). Validity Issues in Evaluative Research. Beverly Hills, CA: Sage Publications.

Bertera, R.L. and Cuthie, J.C. (1984). Blood pressure self-monitoring in the workplace. Journal of Occupational Medicine, 26, 183-188.

Breslow, L. et al. (1985). Risk Factor Update Project: Final Report. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control, Center for Health Promotion and Education.

Campbell, D.T., and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 44, 81-105.

Chen, M.S. (1984). Proving the effects of health promotion in industry: An academician's perspective. Health Education Quarterly, 10, 235-245.

Cincirpini, P.M., Hook, J.D., Mendes de Leon, C.F., and Pritchard, W.S. (1984). A Review of Cardiovascular, Electromyographic, Electrodermal, and Respiratory Measures of Psychological Stress. (NIOSH Contract #84-257). Cincinnati, OH: National Institute for Occupational Safety and Health.

Cook, T.D., and Campbell, D.T. (1979). Quasi-experimentation: Design and Analysis Issues for Field Settings. Skokie, IL: Rand McNally.

Dalkey, N.C. and Helmer, O. (1969). The Delphi Method: An Experimental Study of Group opinion. Santa Monica, CA: Rand Corporation.

Delbecq, A.L., and Van de Ven, A.H. (1971). A group process model for problem identification and program planning. Journal of Applied Behavioral Science, 7, 466-492.

Derogatis, L.R. (1975). The Symptom Checklist 90-R (SCL-90-R). Baltimore: Clinical Psychometrics Research.

Edwards, W., Guttentag, M., and Snapper, K. (1975). A decision theoretic approach to evaluation research. In E. Struening and M. Guttentag (Eds.). Handbook of Evaluation Research, Vol. 1. (pp. 139-182). Beverly Hills, CA: Sage Publications.

Elliott, G.R., and Eisdorfer, C. (1982). Stress and Health. New York: Springer.

Evaluation Research Society. (1980). Standards for Program Evaluation. Washington, DC: Evaluation Research Society.

French, J.F., and Kaufman, N.J. (1983). Handbook for Prevention Evaluation: Prevention Evaluation Guidelines. (DHHS Publication No. ADM 83-1145). Washington, DC: U.S. Government Printing Office.

French, J.F. (1979). Randomized response: A method for increasing the privacy of individual responses to surveys. Current Trends in Drug Research, I. Rockville, MD: National Institute on Drug Abuse.

Fox, J.A., and Tracy, P.E. (1980). The randomized response approach and its applicability to criminal justice research and evaluation. Evaluation Review.

Gilberts, R. (1983). The evaluation of self-esteem. Family and Community Health, 6(2), 29-49.

Green, L.W. (1974). Towards cost-benefit evaluations of health education. Health Education Monographs, 1, (Suppl.), 34-36.

Green, L.W., Kreuter, M.D., Deeds, S., and Partridge, K.B. (1980). Health Education Planning: A Diagnostic Approach. Palo Alto, CA: Mayfield Publishing Co.

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

Green, L.W., Wilson, R.W., and Bauer, K.G. (1983). Data requirements to measure progress on the objectives for the national in health promotion and disease prevention. American Journal of Public Health, 73, 18-24.

Holt, R.R. (1982). Occupational stress. In L. Goldberger, and S. Bresnitz (Eds.) Handbook of Stress. New York: The Free Press.

Jemelka, R., and Borich, G. (1979). Traditional and emerging definitions of educational evaluation. Evaluation Quarterly, 3(2), 263-276.

Jenkins, C.D., DeFrank, R.S., and Spears, M.A. (1984). Evaluation of Psychometric methodologies used to assess occupational stress and strain. (NIOSH Contract #84-2756). Cincinnati, OH: National Institute of Occupational Safety and Health.

Kanner, A.D., Coyne, J.C., Schaefer, C., and Lazarus, R.S. (1981). Comparison of two models of stress measurement: Daily hassles and uplifts versus major life events. Journal of Behavioral Medicine, 4, 1-39.

Karasek, R. (1979). Job demands, job decision latitude, and mental strain: Implications for job redesign. Administrative Science Quarterly, 24, 285-308.

Karasek, R. (1981). Job socialization and job strain. The implications of two related psychosocial mechanisms for job design. In B. Gardell and G. Johansson (Eds.) Working Life: A Social Science Contribution to Work Reform. London: John Wiley and Sons.

Kasl, S.V. (1980). Epidemiological contributions to the study of work stress. In C.L. Cooper and R. Payne (Eds.) Stress at Work. New York: John Wiley and Sons.

Kasl, S.V. (1984). Stress and health. Annual Review of Public Health, 5, 319-341.

Kiefhaber, A., and Goldbeck, W. (1984). Worksite wellness. In Prospects for a Healthier America. (pp. 9-19). Washington, DC: U.S.DHHS, Office of Disease Prevention and Health Promotion.

Kiresuk, T. (1973). Goal attainment scaling at a county mental health service. Evaluation, 1(1), 12-18.

Lerner, M., and Shore, L. (1982). Occupational stress and labor organizing: The work of the Institute for Labor and Mental Health. Socialist Review, 12, 121-139.

McLeroy, K.R., Green, L.W., Mullen, P.D., and Foshee, V. (1984). Assessing the effects of health promotion in worksites: A review of stress program evaluations. Health Education Quarterly, 11, 379-401.

Mager, R.F. (1962). Preparing Educational Objectives. Palo Alto, CA: Fearon.

Manuso, J. (1983). Occupational Clinical Psychology. New York: Praeger Press.

Manuso, J. (1984). Management of individual stressors. In M.O'Donnell, and T. Ainsworth (Eds.). Health Promotion in the Workplace. New York: John Wiley and Sons.

Martin, E.V. (1983). Describing Usefulness of Labor-Management Collaborative Approach to Designing Job-Stress Reduction Programs in the Graphic Arts/Communications Industry. (NIOSH Order No. 82-1966). Cincinnati: National Institute for Occupational Safety and Health.

Masi, D.A., and Teems, L. (1983). Employee Counseling Services Evaluation System: Design, issues and conclusions. Evaluation and Program Planning, 6, 1-6.

Mirvis, P.H., and Seashore, S.E. (1979). Being ethical in organizational research. American Psychologist, 34, 766-780.

Murphy, L.R. (1984). Occupational stress management: A review and appraisal. Journal of Occupational Psychology, 57, 1-15.

Neale, M.S., Singer, J.A., Schwartz, G.E., and Schwartz, J. (1983). Conflicting Perspectives on Stress Reduction in Occupational Settings: A Systems Approach to their Resolution. Cincinnati: National Institute for Occupational Safety and Health.

Nutt, P.C. (1981). Evaluation Concepts and Methods: Shaping Policy for the Health Administrator. New York: SP Medical and Scientific Books.

Parkinson, R.S., Green, L.W., McGill, A., Erikson, M., and Ware, B., et al. (1982). Managing Health Promotion in the Workplace: Guidelines for Implementation and Evaluation. Palo Alto: Mayfield Publishing Company.

Pickering, T.G., Harshfield, G.A., Kleinert, H.D., and Laragh, J. (1982). Ambulatory monitoring in the evaluation of blood pressure in patients with borderline hypertension and the role of the defense reflex. Clinical and Experimental Hypertension, 4, 675-693.

President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research. (1981). Protecting Human Subjects: The Adequacy and Uniformity of Federal Rules and their Implementation. (SN 040-000-00452-1). Washington, DC: U.S. Government Printing Office.

Rose, R.M., Jenkins, C.D., and Hurst, W.M. (1978). Air Traffic Controller Health Change Study. Galveston, TX: University of Texas Press.

Rossi, P.H., and Freeman, H.E. (1982). Evaluation: A Systematic Approach (2nd Edition). Beverly Hills: Sage Publications, Inc.

Scriven, M. (1967). The methodology of evaluation. In R.E. Stake (Ed.) Perspectives of Curriculum Evaluation, AERA Monograph Series on Curriculum, Evaluation, No. 1. Chicago: Rand McNally.

Sechrest, L., and Cohen, R.Y. (1980). Evaluating outcomes in health care. In G.C. Stone, F. Cohen, and N.E. Adler (Eds.) Health Psychology-A Handbook: Theories, Applications, and Challenges to a Psychological Approach to the Health Care System. San Francisco: Jossey-Bass.

Sheldon, E.B., and Parke, R. (1975). Social indicators. Science, 188, 693-699.

Shepard, D.S., and Thompson, M.S. (1979). First principles of cost-effectiveness analysis in health. Public Health Reports, 94(6), 535-543.

Shortell, S.M., and Richardson, W.C. (1978). Health Program Evaluation. St. Louis: C.V. Mosby.

Siegel, L.M., Attkisson, C.C., and Cohn, A.H. (1977). Mental health needs assessment: Strategies and techniques. In W.A. Hargreaves, C.C. Attkisson, and J.E. Sorenson (Eds.) Resource Materials for Community Mental Health Program Evaluation 2nd Ed. (DHEW Publication Washington, DC: U.S. Government Printing Office.

Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company.

Spielberger, C.D., Gorsuch, R.L., and Lushene, R. (1968). State-trait Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.

Sokolow, M., Perloff, D., and Cowan, R. (1980). Contribution of ambulatory blood pressure to the assessment of patients with mild to moderate elevation of office pressure. Cardiovascular Reviews and Reports, 1, 295-303.

Stainbrook, G.L. and Green, L.W. (1983). Role of psychosocial stress in cardiovascular disease. Houston Heart Bulletin, 3, 1-8.

Tesh, S.N. (1983, November). The Politics of Stress: The Case of Air Traffic Control. Paper presented at the American Public Health Association 111th Annual Meeting, Dallas, Tx.

U.S. Department of Health Education and Welfare. (1978a). Occupational Stress: Proceedings of the Conference on Occupational Stress. (DHEW, NIOSH Publication No. 78-156). Washington, DC: U.S. Government Printing Office.

U.S. Department of Health Education and Welfare. (1978b). Reducing Occupational Stress: Proceedings of a Conference. (DHEW, NIOSH Publication No. 78-140). Washington, DC: U.S. Government Printing Office.

U.S. Department of Health Education and Welfare. (1979). Healthy People: The Surgeon Generals Report on Health Promotion and Disease Prevention. (DHEW, PHS, Publication No. 79-55071). Washington, DC: U.S. Government Printing Office.

U.S. Department of Health and Human Services. (1980a). New Developments in Occupational Stress: Proceedings of a Conference. (DHEW, NIOSH Publication No. 81-102). Washington, DC: U.S. Government Printing Office.

U.S. Department of Health and Human Services. (1980b). Promoting Health/ Preventing Disease: Objectives for the Nation. (1981 0-349-256). Washington, DC: U.S. Government Printing Office.

Warheit, G.J., Bell, R.A., and Schwab, J.J. (1977). Needs Assessment Approaches: Concepts and Methods. Washington, DC: National Institute Drug Abuse.

Weinstein, M.C., and Stason, W.B. (1977). Foundations of cost-effectiveness analysis of health and medical practices. New England Journal of Medicine, 296, 716-721.

Weiss, C.H. (1972). Evaluation Research. Englewood Cliffs, CA: Prentice Hall Inc.

Windle, C., Rosen, B.M., Goldsmith, H.F., and Shambaugh, J.P. (1975). A demographic system for comparative assessment of "needs" for mental health services. Evaluation, 2(2), 73-76.

Windsor, R.A., Baranowski, T., Clark, N., and Cutter, G. (1984). Evaluation of Health Promotion and Education Programs. Palo Alto, CA: Mayfield Publishing Company.