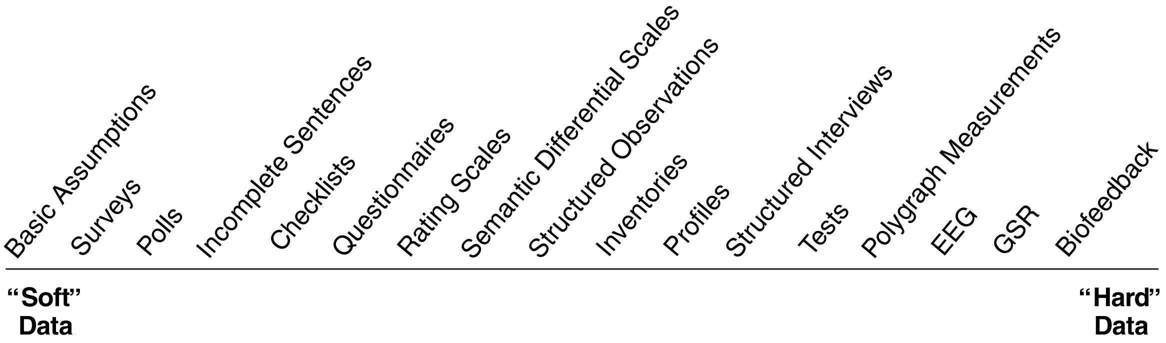
A BACKGROUND FOR USING INSTRUMENTATION IN HUMAN RESOURCE DEVELOPMENT

WHAT AN INSTRUMENT IS AND WHAT IT IS NOT

Despite the inherent difficulty of measuring psychological variables and processes, many devices are available to assess, appraise, evaluate, describe, classify, and summarize the various aspects of human behavior. In human resource development (HRD), these devices are called "instruments"; almost all of them are paper-and-pencil feedback devices that are used to generate data and to personalize theory input within a group setting. The term that is used to describe any particular instrument depends primarily on the format and purpose of the instrument. Typical terms used to describe instruments include: inventory, test, survey, poll, questionnaire, incomplete-sentence form, opinionnaire, checklist, rating scale, profile, semantic-differential scale, reaction form, and evaluation. This makes a distinction between a "test," in which the *correct* answers are determined by the test constructor, and a survey, questionnaire, inventory, or rating scale that does *not* have a specific, "correct" answer for any given item.

For HRD purposes, then, an instrument is used to obtain data and to provide information to the respondents/clients about themselves and/or their group or to exemplify points of a particular theory or other conceptual input. For example, an instrument can collect team members' perceptions prior to a team-building effort or employees and managers perceptions prior to an organization-development effort; an instrument also can be used to provide information to individual respondents on their particular "styles" or modes of behaving in a number of areas and for a number of purposes such as personal growth, management development, stress or conflict management, communication, and so on. Instrumentation frequently is employed to reinforce the learnings from structured experiences and other group activities, and the data that are obtained from the instrument serve to make the focus more personal, more real, and more credible to the participants.

Methods of psychological assessment can be located on a continuum from "soft" to "hard" data. "Soft" refers to highly subjective data that are difficult to measure; "hard" refers to more objective, quantifiable data. Soft data are not necessarily less valid than hard data, but their validity is more difficult to demonstrate. The instruments most commonly used in HRD fall in the middle of the spectrum. They include inventories, questionnaires, surveys, polls, incomplete sentence forms, opinionnaires, checklists, rating scales, profiles, semantic differential scales, and structured interviews. Tests—measurement devices that have items with right and wrong answers—and physiological measurements, such as the polygraph or biofeedback, are almost never used in training and development, although the polygraph has been used in some types of personnel work.



Subjective-Objective Continuum

A RATIONALE FOR USING INSTRUMENTATION

One of our real dilemmas in the human resource development (HRD) field is that we lack a descriptive vocabulary for dealing with other people in a nonpejorative way. Most people can describe their physical world with great precision; our vocabulary for describing colors, sizes, shapes, scents, flavors, and the like is well developed. However, we quickly run out of descriptors when we attempt to describe other people, especially if the behavior of those others has had an adverse effect on us. One of the principal benefits of using instrumentation in HRD is that instruments typically provide participants with a new, rich vocabulary for describing others. With such a new vocabulary, one can begin to describe another person's behavior as "stemming from a strong need for inclusion" or "representing a weak economic value commitment" rather than in more subjective and emotionally laden terms that interfere with, rather than enhance, communication (especially communication with the person being described). Thus, instrumented assessment has many advantages in an HRD context. An instrument is objective and it employs a common language so that participants can give standardized responses that are quantifiable and economical to score and summarize. Participants in today's training settings want actionable feedback, not vague and theoretical models.

In addition to helping participants to identify behavior, the comparison of scores from an instrument provides group members with a convenient and relatively neutral way of exchanging interpersonal feedback. Instruments usually are based on theory and thereby have didactic potential. They promote involvement, leading to increased participation and personalized learning. The involvement with their own scores helps participants to better understand the theory on which the instrument is based—a typical reason for using an instrument in training.

If members of a group think of themselves as studying their group through an objective instrument, the activity seems more scientific, less subjective, and more acceptable. In a group without a leader, an instrument can be used to put the responsibility for processing the group's activities on the members themselves. Instruments also can be used in work teams and meetings to evaluate their effectiveness.

In general, more growth can occur for a group participant if he or she is provided with a method for focusing specifically on his or her own behavior. In addition, the feedback received from other group members can contribute to growth on a different and equally important level. In both cases, the learning is highly specific and can contribute to decisions concerning behavioral change. In addition, being able to relate to the particular outcomes of an instrument may serve to reinforce new behavioral patterns and enhance self-concepts when the participant returns to the environment outside the training setting and when the impact of fellow participants becomes diffused with the ongoing demands of old relationships.

Externally derived data can provide the basis for potentially significant growth experiences and change, especially if the data are understandable, reflect actual behavior, and are discrepant from some ideal or desired self-image. Paper-and-pencil instruments are the most common and convenient approach to gathering such data. They are useful not only to examine the behavior of individuals but also to collect data in groups and organizations.

One reason for the growth of instrumentation in the HRD field is the increasing complexity of modern organizations, in which the ability to access information and gain cooperation are critical factors in achieving success. Organizations are adopting more complex structures and models as they strive to implement performance systems and measures. There is an increasing number of "knowledge workers" who are highly dependent on subjective data to make decisions and evaluate performance. Managerial skills have been recognized as being more important than "position power" in getting things done; performance evaluation has become more complicated and subjective as fewer jobs can be evaluated in terms of "pieces produced." There also is an increase in service functions, which require efficient, concise feedback with action implications. At the same time, individuals are demanding more from their organizations in terms of career planning, meaningful work, and involvement in the decision-making process. Social and business trends all require that individuals and organizations have access to greater amounts of data in a structured and systematic way. Human resource development, in responding to these new requirements, has realized the need for more sophisticated technologies for providing feedback (Peters, 1985).

In addition to providing personal feedback for training purposes, instruments now are used frequently to obtain information that will serve as the basis for organizational change and planning efforts. The computer is being utilized more in assessing data, and it allows the processing of more data from multiple sources to provide more sophisticated and objective feedback for assessment and development planning. Software packages now allow users to customize instrumentation to their specific needs and applications, and computerized instrumentation makes it possible to implement large-scale assessments over long periods of time (Peters, 1985).

Instruments such as inventories, questionnaires, and surveys also are the best way to collect information from large numbers of people (i.e., when the available time and resources would not allow for each individual to be interviewed or observed). It takes less time to collect data from an instrument than it would from other processes such as interviews or verbal surveys. Because the items or questions are standardized and the results are all tabulated the same way, instruments are less susceptible to bias; it also takes less time to analyze the data.

In order to use instruments effectively, HRD practitioners must be aware of the ramifications of the technology: the instrument administrator or facilitator must be personally familiar with the specific instrument being administered; the instrument must be chosen solely on the basis of the expressed needs and goals of the particular client group; and the facilitator must be competent to interpret the data that emerge in order that the feedback can be used in functional ways.

Providing Specific Feedback to Individuals

Instrumented feedback can be more useful than the verbal feedback typically received in groups. Too often participants provide feedback that is absolute; that is, it does not locate the person in reference to degrees of a trait or to a norm. It is not uncommon to hear statements such as “You talk too much,” “You are too quiet,” or “You are too aggressive.” The obvious response is “Compared to what?” Well-constructed scales also can provide feedback on behavioral extremes that may be equally dysfunctional, providing the respondent with a continuum of behavior from which to choose.

A primary value of instrumentation, then, is as a source of personal feedback for individuals in a training group. This involves the individual completion, scoring, and interpretation of scales as the primary step and as the basis for more extensive uses of instrumentation. Participants can be asked to predict one another’s scores so that individuals become more aware of their facades and of the impact they have on others. If appropriate, participants can complete entire instruments for one another for a more in-depth examination of interpersonal perceptions. An example of this is the set of LEAD-Self and LEAD-Other instruments (Hersey & Blanchard, 1973), which are designed to be completed by a manager and his or her subordinates or peers.

Studying Here-and-Now Processes in Groups

A related use of instruments is to help participants to study here-and-now processes within the group and to assist the group in diagnosing its own internal functioning. Instrumented data can focus on what is happening in the life of the group and may specify what changes are desirable. In this way, the group can more quickly arrive at optimal functioning so that more learning can take place. This is useful in teaching awareness of group process and diagnostic skills to intact groups and also as preliminary work for team building and organizational-change efforts. Again, the unique advantage of specificity with instrumented feedback greatly enhances the probability that the group will be able to monitor and manage its own processes effectively.

Teaching Theories of Behavior and Interpersonal Functioning

Many instruments can be used to teach the theory, concepts, and terminology that are intrinsic to the description and interpretation of a particular set of behaviors, model, etc. In fact, some brief instruments are intended primarily to introduce concepts rather than to be used as a source of feedback.

Participants learn more when they actively are involved in the learning process. When participants have invested time and energy in an activity such as completing an inventory related to the model being explored, they also have invested in learning the theory, and the entire process becomes more meaningful in terms of the group experience. Participants also can be encouraged to study the items of the instrument in detail, because the items constitute a behavioral definition of the trait being measured. This, and a personal review of their own scores, can result not only in deeper understanding but also in their considering specific behavioral change.

Manipulating Group Composition

The facilitator may wish to use instrumented outcomes to manipulate group composition for brief, experimental demonstrations of the effects of group composition on task accomplishment. Long-term groups that offer promise of demonstrating group-task competencies also can be built. Extremes of both homogeneity and heterogeneity can be avoided through the use of instrumented data.

Researching the Outcomes of Training and Other Interventions

The measurement of human phenomena can be a needed and realistic expectation of learning and change experiences, contributing toward growth, evolution, and behavioral change. Instruments also can provide the means of assessing growth and change on both individual and group levels. Even scales with relatively low reliability can be used effectively to study group behavioral patterns and attitudes in a pre-, post-, and post-post design. This measurement of outcome can provide some of the most crucial feedback for individuals in a training setting and may help to validate the experience for them and for the group in general. In addition, instrumented research will provide feedback to the facilitator on the effectiveness of his or her own style and intervention skills and will aid in designing interventions for other groups.

Summary

Instrumented survey-feedback tools (generally inventories or measurement scales) can be used in a number of ways by trainers and consultants. Data from inventories can be interpreted normatively or intrapersonally, but it is important that they be coordinated carefully with the goals of the training design. Some uses of instrumentation include the following:

- *Providing instrumented feedback to group members*. Participants complete, score, and interpret their own scales. They can be asked to predict one another’s scores. They can fill out scales for one another as feedback.
- *Studying hear-and-now processes in groups*. It sometimes is helpful to use an instrument to assist the group in diagnosing its own internal functioning. The data can be focused on what is happening and what changes are desirable.
- *Teaching theories of interpersonal functioning*. Some brief instruments are intended primarily to introduce concepts. Participants are involved with theory by investing in an activity such as completing an inventory related to the model being explored.
- *Manipulating group composition*. For brief, experimental demonstrations of the effects of group composition, various mixes of group members can be established. Long-term groups can be built that offer the promise of beneficial outcomes. Extremes of both homogeneity and heterogeneity can be avoided.
- *Researching outcomes of training and other interventions*. Even scales with relatively low reliability can be effective in the study of group phenomena when used with pretest or follow-up procedures.

TRAITS

It is important in using instrumentation to remember that one is dealing with outcomes based on visible elements of human behavior, or traits. Prior to using an instrument, the facilitator should complete his or her understanding of the nature of the traits on which the particular instrument is based. The facilitator must be knowledgeable enough to ease the anxiety of the participants concerning what the instrument will “reveal” about them—to help them to anticipate the learning experience in a nonthreatening way. A presupposition of learning theory is that if the participants are highly anxious, they will be largely incapable of hearing, seeing, and learning what is presented.

Behavioral scientists, like natural scientists, build taxonomies or ways of naming observable phenomena. The reason for naming is to provide a common ground for communication regarding what may be experienced by individuals. The naming does not create the phenomena; it merely attempts to label, in some meaningful way, what already exists. Traits, then, are sets of categories invented by behavioral scientists to permit the orderly description of behavior. This definition can be illustrated by the following anecdote.

> Three baseball umpires were involved in a heated discussion of what they considered to be strikes and balls.
>
> First umpire: “Well, it’s easy, fellas. I call ’em as they are: If it’s a strike, I call it a strike; if it’s a ball, I call it a ball.”

Second umpire: “Wait a minute! I see it different! I call ’em the way I see ’em. If I see a strike, I call a strike, and if I see a ball, I call a ball!”

Third umpire: “You’re both wrong! They ain’t nothin’ ’till I call ’em! It’s just a baseball that got thrown ’till I holler ‘Strike!’ or ‘Ball!’”

This anecdote gets to the heart of the issue. People do not “have” traits in themselves. They do not, for example, have a trait of inclusion or a trait of affection or a trait of control, *per se.* These are labels imposed on people’s behavior to add some order, understanding, and predictability to the behavior. One of the difficulties in using instrumentation is that many people tend to infer that the trait being measured is an integral part of their psychological makeup or behavioral pattern. It is important to stress that these simply are imposed categories; they do not exist in themselves any more than a strike or a ball exists before it is called by an umpire.

A related concept that may help to debunk the image of infallibility that participants often assume about instruments is the story about the Air Force way of dealing with things: to measure it with a micrometer, mark it with a crayon, and then cut it with an axe. The final outcome of instrumentation is equally inexact, even though the intent is to be as precise as humanly possible. The key word here is “humanly.” At best, the outcomes merely suggest types of behavior; yet, this suggestion can be of great value if it is seen for what it actually is: an indication.

It is not necessary to adopt a trait-factor theory of personality in order to employ instrumentation effectively. It is possible to process instrumented feedback in terms of widely varying theoretical positions, from analytical to existential. The following two statements illustrate two extreme positions.

> “If a thing exists, it exists in some amount. If you haven’t measured it, you don’t know what you are talking about.”

> “No number or combination of numbers ever can adequately describe a dynamic, emerging person. The important characteristics of humans are immeasurable.”

In the context of one’s own theoretical frame of reference, it is possible to incorporate “objective” data to good effect so long as one bears in mind that the process simply abstracts from the mass of information that is available about an individual.

One of the most powerful learnings derived in human resource development is that people are far more alike than they are different. Participants discover that many of the differences among people are noncritical. It is a mistake to concentrate so much on individual differences that we ignore the commonalities. Instrumentation can demonstrate the large overlap across persons of a wide array of human traits. Outcomes can be interpreted in a perspective that acknowledges that, at a humanistic level, all differences do not necessarily make a difference. One can use instruments to discriminate among people, to spread them out for study or instruction. We propose that it is equally advantageous to use instrumentation to demonstrate graphically the common core of humanness that can bind us together.

THE DISADVANTAGES AND ADVANTAGES OF USING INSTRUMENTS

It is important to recognize both the advantages and disadvantages of using instruments in human resource development. These can be dealt with most effectively by discovering ways to minimize the problems and maximize the advantages.

Minimizing the Disadvantages

The “Labeling” Effect

One of the key disadvantages of using instruments is that the participants may fear that someone will obtain an indelible fingerprinting of them, that they will be exposed, that somebody will get into their minds and “psych them out.” This fear may be accompanied by resentment and a loss of learning potential. An accompanying problem is that some participants may accept their scores as unquestionably accurate descriptions of themselves: they *are* “assertive,” “aggressive,” or “withdrawn.” Some participants may attach pathological or quasi-pathological definitions to their traits and turn them into self-fulfilling prophecies. This problem of labeling also can occur when participants are dealing with one another. They may still refer to Joe as “at the ninety-eighth percentile on control” after the group session is over despite the fact that Joe may have spent a great deal of time and energy during the workshop in experimenting with new behaviors and may have modified his control pattern considerably. This problem of not allowing people out of their old stereotypes is particularly counterproductive in intact groups. Joe may discover three years later that some individuals are still relating to him as if he were “at the ninety-eighth percentile on control.”

It is extremely important that the facilitator make an effort to reduce this tendency to overgeneralize the accuracy and stability of the instrument. To avoid having participants interpret the instrument in this manner, the facilitator should discuss the margin of error and other factors that contribute to less-than-absolute results. An instrument can be described as analogous to a thermometer; the reading would be expected to vary from time to time.

Participants also should be encouraged to explore the instrument thoroughly so that they can see how it was designed and how their scores were derived. They need to acknowledge the fact that all they have done is to give their best answers, at a particular point in time, to the situations or questions in the instrument, and that they added numbers denoting those answers to come up with a score. If they have trouble understanding the scores that resulted, they should be encouraged to go back to each item to see how they responded to it and how they scored it and perhaps to compare their responses with those of others, item by item.

It also is useful to show the participants how instrumentation is related to everyday choices, fraught with inconsistency and subject to influences of all kinds, such as one's psychological set at the moment, one's physical state, and so on. Efforts such as these can encourage a realistic way of looking at "test" outcomes. Participants should be helped to understand that instrumented feedback, like other forms of feedback, can indicate only what *may* be true of the individual at a given point in time.

Attempts To Take Flight

Another problem that may arise for the facilitator is that instruments sometimes promote flight from personal and interpersonal issues for the participants. An instrument may generate a rash of nitpicking (inappropriate) responses. Participants sometimes question items, reliability, validity, or the value of the instrument rather than choosing to see it as a tool for potential learning. A lot of time can be wasted in arguing about the instrument itself—a result of the fact that participants have received, or fear that they are about to receive, information that disturbs them. They may be afraid that others will interpret their data in a negative fashion, so they attack some aspect of the instrument in order to change the focus or to minimize the impact of the data. A nonhostile "flight" may involve a tendency to engage in a "psychological" discussion about the traits the instrument is measuring.

The facilitator can alleviate these flights by dealing thoroughly with the mysticism and reality of instrumentation. He or she also can intervene to refocus the group discussion on the data obtained from the instrument and away from the instrument itself so that participants become involved once more in the learning process.

It is a great help to establish the expectation prior to the group experience that an instrument will be part of the design and to discuss why instrumented feedback will help to meet the group's needs and goals.

The use of instruments can be a means of dissipating some of the ambiguous but potentially growth-producing tension that is generated by encountering and reacting to others face-to-face. It also may tend to pull individuals away from the interpersonal processes of the group if data from the instrument have created an emotional overload. Participants may become so preoccupied with integrating the data that their behavior is dysfunctional to the rest of the group. Some of this difficulty can be eased if the facilitator takes time to process the data from the instrument sufficiently so that the participants can "handle" it and integrate it without detracting from the business of the group.

A balance must be maintained, consistent with the goals of the group, between instrumentation and the interpersonal learning needs of the participants. Processing the data can and should include more than the facilitator's interpretation of scores. Participants should have the opportunity to talk through their scores and to compare them with the scores of others in the group and with appropriate norm groups, if such information is available. The facilitator should emphasize and legitimize the different life perspectives and orientations among people and should encourage participants to explore these differences. The facilitator may use his or her own scores to illustrate how personal orientations govern responses to the instrument and to share with the participants the personal impact of the feedback that he or she received from the instrument scores.

Dependence on the Facilitator

A third disadvantage of the use of instrumentation is the tendency for the activity to foster dependence on the facilitator. The administration of the instrument and the subsequent feedback process can put the facilitator in the role of expert rather than of process consultant. Participants who initially find it anxiety producing to be in a group situation in which the “leader” does not assume a traditional leadership role may find it hard to let the facilitator out of the directive, authority role. It is particularly important for the facilitator to shift the responsibility for learning during the feedback-processing stage back to the participants themselves as early as possible, making the transition from a highly structured activity to minimally structured group processes.

It also is important to emphasize that the responsibility for learning rests with the participants. The meaning of the feedback for each individual is within that person, and he or she integrates the data with a personal understanding of himself or herself. If a participant seems to have received more feedback than he or she is ready to handle, the facilitator can encourage the person to take time to assimilate the information and to work through it in order to put it into perspective.

The “Test” Stigma

The idea of being “tested” may provoke subtle anger or undue anxiety on the part of participants. Instruments can trigger unpleasant memories of school grading practices. The facilitator should avoid using the word “test” and should stress that the instrument is nonevaluative. Virtually all the instruments used in human resource development do not have “right” or “wrong” answers and are merely tools for obtaining information; this fact should be made clear to the participants.

Manipulation of Scores

Many instruments are subject to distortion; participants may lie or answer in ways that they believe are socially desirable. Most instruments have some degree of transparency, and the more sophisticated the group is in terms of test taking, the more potential there is for participants to distort their scores. However, if the participants have a commitment to their own learning, the tendency to distort will be lessened greatly. If the participants have volunteered for the experience, their level of commitment generally will be high. If the group is an intact, nonvoluntary one, the facilitator should attempt to inspire commitment to the learning goals of the experience. This commitment must exist if the experience is to be productive on any level.

Enhancing the Advantages

Although there are problems in using instruments, there are means of coping with or avoiding these problems. In addition, there are many advantages. Instruments are highly involving because participants are working with information about themselves; however, because they supply most of that information by responding to the instrument, the activity is not as threatening as some other vehicles for feedback.

Early Presentation of Theory

An instrument gives the respondents the opportunity to understand the theory involved in the dynamics of the group situation—understanding that can increase their involvement. Through judicious use of an appropriate instrument during the first group session, the facilitator can offer the participants a theory about personality styles, group development, interpersonal relations, or leadership that can be built on throughout the rest of the group experience.

If separate theory sessions are part of the training design, participants will find them stimulating and meaningful if they explain the rationale for the instrumented feedback that the participants have received. Theory sessions without such a previous activity might be experienced by the participants as a sudden change from an active, involved situation to a passive, listening situation.

Early Understanding of Constructs and Terminology

A related advantage of using an instrument is that it gives the participants some constructs and terminology early in the group experience that they can use in looking at their behavior and in categorizing and describing what goes on between individuals or within an individual. Participants seem to form commitments to information, constructs, and theories when their instrumented feedback describes them in terms of those constructs and theories. The participants' learning is crystallized when their "selves" are tied to useful information about interpersonal relations or groups, that is, when it has personal impact.

Early Feedback and Time To Practice New Behavior

Through instrumentation, a participant can be given feedback about his or her personal behavior early in the group experience. It often happens that an individual does not receive feedback from other participants about his or her style or way of relating to others until the last day or the last two or three hours of the workshop. It often takes that long for the other participants to develop the skills necessary to give effective feedback and before an atmosphere of trust is developed in the group so that members can feel comfortable in exchanging personal feedback. In this case, the participant receives information that he or she needs to know but he or she has no time to work on new behavior. Instruments administered early in the group experience help to compensate for the lack of feedback from others. Participants can generate personal agendas for behavior modification based on characteristics revealed by the instrument, with the remainder of the workshop to work on them. Individuals can “contract” with the group members to experiment with new behaviors. The personal input into the instrument also increases the chances that participants will form personal commitments to change and grow; this is not as likely to occur as a result of feedback from others, which is easier to discount or forget.

Personal Input To Foster Acceptance of Feedback

The feedback from instruments is relatively “low threat.” Because the individual has filled out the instrument form himself or herself, he or she is more likely to trust the data. At least the individual does not have the dilemma of trying to determine whether the information is primarily a function of his or her behavior or of the mind-set of the person who is providing the feedback or of some chemistry between the two of them. An instrument form obviously can hold no personal malevolence toward an individual, and participants can readily see that the information they obtain actually came from their own responses to the form.

Comparison of and Involvement with Data

Another advantage is that instruments not only provide feedback about the individual, they also allow the person to compare himself or herself with others. We all are aware that we may be more or less dominating than others, that we may enjoy being with people more or less than others, that we may have a greater or lesser need for others to like us, and so on. However, it often is an eye-opening experience to find out that one is stronger in a particular characteristic than 99 percent of the people in a certain norm group. This information can cause a person to examine carefully whether this characteristic is becoming dysfunctional, i.e., getting in the way of his or her performance at home or on the job.

Instrumentation can promote involvement with data of all kinds. Participants often come to a training session never having heard the word “feedback” or thinking that it was purely computer terminology. Participants can be taken from the concept of feedback as it applies to data processing to its relationship to instruments and then to its application in interpersonal relationships. After initial experience with feedback from an instrument, participants can practice giving and receiving feedback in a relatively nonthreatening experience such as predicting one another’s scores. This process allows the participants to learn to give and receive constructive feedback early in the life of the group and gives the group experience a greater chance to impact the behavior of the participants.

Latent Issues Surfaced

Instruments surface covert issues that can be dealt with in the group setting. This is true whether the issues are within an individual, between individuals, within the group, or within an organization. An instrument that uncovers these issues validates the public airing of these concerns and makes them legitimate topics to be dealt with, discussed, corrected, and improved.

Control over Focus

Instruments aid the facilitator in that they focus the energies and time of the participants on the most appropriate material and also allow him or her to control, to some extent, what is dealt with in the group session. In this way the facilitator is able to ensure that crucial, existing issues are worked on—that the group does not tackle less important issues in order to avoid grappling with the more uncomfortable ones.

Assessment of Change

A final, but important, advantage is that instruments allow longitudinal assessment of change in a group, an organization, or an individual. This assessment can be useful in organization development work for demonstrating that the group interventions are compatible with the goals of the client, the consultant, and the organization. It also can be valuable for group research and personal feedback.

Summary

The following lists summarize the disadvantages and advantages of using instrumentation with small groups and ways to deal with each:

Disadvantages

Engenders fear of exposure.

Encourages “labeling.”

Promotes flight from confrontation.

Generates time-consuming nitpicking.

Relieves potentially growthful tension.

Fosters dependence on the facilitator.

Makes the facilitator an “expert.”

Can result in feedback overload.

Triggers anger and anxiety about “tests.”

Makes distortion of feedback possible through manipulation of scores.

Advantages

Enables early, easy learning of theory.

Promotes personal involvement and commitment.

Develops early understanding of constructs and terminology.

Supplies personal feedback sooner than other participants are able to.

Facilitates contracting for new behavior.

Fosters open reception of feedback through low threat.

Allows comparisons of individuals with norm groups.

Promotes involvement with data and feedback process.

Surfaces latent issues.

Allows facilitator to focus group and control content.

Facilitates longitudinal assessment of change.

Avoiding the Disadvantages of Instruments

1. Legitimize the use of instrumentation with the participants.
   - Establish clear expectations concerning instruments and their value to the group experience prior to the beginning of the session.
   - Be ready to intervene to refocus the group discussion if participants use the instrument as a flight mechanism.
   - Minimize anxieties so that more learning can occur.
2. Make a concerted effort to remove the mysticism surrounding instrumentation.
   - Discuss the margin of error and other factors that contribute to less-than-absolute results.
   - Allow and encourage participants to explore the instrument thoroughly so that they see how it was designed and how their scores were derived.
   - Clarify the theoretical basis of the instrument.
3. Ensure that sufficient time is made available for participants to process the data derived from the instrument.
   - Provide an opportunity for participants to talk through their scores and to compare their scores with those of others.
   - Emphasize and legitimize different life perspectives and orientations.
4. Assure the participants that they have control over their own data.
   - Define carefully the ways in which scores are to be shared or not shared.
   - Emphasize that scores will not be reported to the participants' supervisors.

ETHICAL CONSIDERATIONS

There is no commonly accepted ethical code that binds users of training materials, but a number of practices in relation to the use of instruments have ethical implications. Materials should not be reproduced from copyrighted publications without permission. Even if the use can be categorized as "classroom use" and the number of copies reproduced is small enough to qualify as "fair usage" under copyright law, the materials still should bear the author's name and the title, publisher, and copyright notification of the original source. In such cases it is a courtesy to authors to advise them of the intended use of the instrument; they may be able to supply current information that will be of use to the facilitator. It is unfortunate that scales so often are "borrowed" from complex instruments and that elements from one instrument are "adapted" for inclusion in new ones. The problem here is not only a question of honesty; the risk is that the elements that are used will not have the same validity out of the original context. Oversimplification or misinterpretation can invalidate the purpose and results of an instrument.

The question of validity and reliability is another consideration. Scales sometimes are named to indicate more validity than has been demonstrated; norm groups are inadequately described; and validity and reliability evidence often is nonexistent. Although many instruments used in training have not been used previously for research purposes and have only face validity, it is important that the instrument not be so transparent that respondents can anticipate the purpose of the scoring and answer accordingly.

Much of what participants experience in any training or development group depends on the skill and personal style of the facilitator(s). Before attempting to use instrumentation, facilitators must understand how to administer an instrument, present its theoretical background, help the participants to predict their scores, score their instruments, interpret the data, and process the results. If the facilitator is not skilled in using instrumentation, participants may be harmed by the process. (The discussion that follows presents a detailed, seven-stage process for using instruments and also discusses the issue of facilitator style.)

In addition, participants should not be co-opted into revealing themselves through instrumentation; it is intended as a learning and self-development experience. Therefore, scores on scales that denote pathology should not be published within the group.

HOW TO PRESENT INSTRUMENTATION

THE SEVEN PHASES IN PRESENTING AN INSTRUMENT

There is a sharp distinction between just "giving" an instrument to a group and presenting it properly, i.e., getting the most value out of it in terms of the goals of the experience and the needs of the participants. There are seven basic phases in the presentation of an instrument:

1. Administration
2. Theory Input
3. Prediction
4. Scoring
5. Interpretation
6. Posting
7. Processing

To illustrate these seven phases, we will describe the manner in which we present the *FIRO-B*, an instrument often used in interpersonal training. The *FIRO-B* originally was developed by Will Schutz in 1957.

1. Administration

This stage usually will take ten to fifteen minutes. First, a nonthreatening atmosphere should be established. One way to do this is to use the word "instrument" rather than "test" and to indicate that there are no right and wrong answers. Then the use of the instrument should be legitimized by discussing the purposes of the instrument (i.e., the reasons for taking it and how it fits into the goals of the session). In administering the *FIRO-B*, it is important to avoid giving any clues about the nature of the traits that are being measured.

After the instrument has been distributed, give clear, objective instructions about how to respond to the instrument. Give directions sequentially before participants begin to take the instrument. Encourage the participants to answer honestly in order to promote greater self-learning. In administering the *FIRO-B*, caution the participants on two points: (a) although they will experience a sense of repetitiveness in the items of the instrument, each item is to be considered and responded to independently, and (b) some of the answers do not quite fit (participants can be encouraged to be creative in following the intent of the instrument).

The question of whether or not participants should identify themselves on an instrument form depends, in part, on the instrument itself, its purpose, and the needs of the facilitator and participants. If the participants feel anxiety about self-disclosure (as they might in a work setting), more truthful responses might be obtained if they are allowed to remain anonymous. Even participants in a training workshop may adjust their answers because they fear evaluation by other group members. In such cases, respondents may be asked to code their forms (using part of their social security numbers or telephone numbers, for example) so that the forms can be identified later and retrieved by the respondents. Of course, the facilitator should attempt to establish the norms of openness and honesty in filling out the instrument forms and, if appropriate, should encourage the group members to "own" their responses.

A difficulty, particularly in larger groups, is that individuals in the group will finish the instrument at different times. In a nonauthoritarian way, the administrator of the instrument can establish the expectation that as people finish the instrument they will wait quietly for the others to finish. When everyone has completed the instrument, the group moves to the second phase.

2. Theory Input

This phase usually takes ten to twelve minutes. It is the time to clarify the theoretical basis of the instrument. Sometimes this can be done best through the use of analogy and visual aids. The theory behind *FIRO-B* is that all human interaction can be divided into three categories: issues surrounding inclusion, issues surrounding control, and issues surrounding affection.[1] Schutz's theory of group development suggests that a group proceeds through inclusion issues into control issues and finally into affection issues and then recycles. To illustrate these categories, the facilitator can ask the participants to consider a group of people riding in a boat. The inclusion issue, with the boat, is whether or not individuals have come along for the ride. The issue of control with the boat is who is running the motor or operating the rudder. The affection issue concerns how closely people are seated together in the boat.

The facilitator can follow this discussion of categories with an explanation of the *FIRO-B* six-cell diagram.

	INCLUSION	**CONTROL**	**AFFECTION**
Expressed Behavior			
Wanted Behavior			

The dimensions illustrated by the diagram are inclusion, control, and affection in terms of what the individual *expresses to* others and those same three issues in terms of what the individual *wants from* others. It is important to clarify that the expressed behavior is one's *own* behavior and that the wanted behavior is what one wishes *from others.* When the participants understand this concept and the meaning of each of the six cells, it is time for the prediction phase.

[1] Schutz revised his dimensions in 1982 with the publication of *The Schutz Measures*. In these instruments, the dimensions of behavior are inclusion, control, and openness.

3. Prediction

This phase takes about two minutes. All participants are asked to predict whether they will score high, medium, or low in each of the six cells and to write their predictions in a corner of each cell in the diagram on their copy of the instrument. When the predictions have been made, it is time to begin the fourth phase, scoring.

4. Scoring

This discussion will concern itself with instruments that can be scored *in situ*, not those that must be sent out for computerized analyses. For the *FIRO-B*, the scoring process generally takes six to eight minutes. There are a number of ways in which to score instruments. Some require templates; some are self-scoring; and for some the scores can be called out or written on newsprint or handed out on duplicated sheets of paper. If the scoring is fairly understandable, simple explanations or handouts may suffice. It is important to gauge the level of sophistication of the particular group in selecting the most appropriate way to score an instrument. Many instruments available to HRD practitioners are self-scoring; these provide the participants with immediate feedback and a sense of how the scores are derived. Virtually all such instruments come complete with scoring and interpretation or profile sheets.

In some cases it is more efficient for the facilitator or another staff member to do the scoring. This is appropriate if the scoring is difficult. Another reason for this approach is that sometimes people can create such a task out of the scoring process that they lose or diminish the actual results of the instrument, i.e., the scoring detracts from the data being generated. The obvious negative factor in this approach is that individuals do not receive immediate feedback. This can be mitigated by administering the instrument before a meal break and having the results available immediately after the meal. The important thing is that the scoring must not detract from the data being generated by the instrument.

Once the facilitator has worked through scoring the first scale on the *FIRO-B* with the participants, they generally can derive their own scores for the other scales at their own pace. After the scoring has been completed and the participants have posted their scores in the cells along with their predictions, the interpretation phase begins.

5. Interpretation

The way in which the *FIRO-B* is interpreted can vary greatly depending on the participant group and the style of the facilitator. Typically, it takes from five to seven minutes per participant. We like to handle interpretation in two stages: the first stage is an interpretation of the facilitator’s own scores or those of another staff member; the second stage is a dyadic interpretation between pairs of participants. In a typical design, the scores of another staff member first are interpreted by means of a six-step method, so that participants can begin to see how interpretations are made. Then that staff member interprets the facilitator’s scores in front of the group to afford participants an opportunity to see some variance in the style of interpretation. This modeling of interpretation is a very important defusing element. If the staff members are willing to show their scores to others, the individual participants find it easier and less threatening to share their scores with other members of the group.

The following are the six steps in the first phase of interpretation employed for the *FIRO-B.*

1. Actual scores are compared cell by cell with the predictions of high, medium, or low. The conversion for this design is as follows:

Prediction	**Actual Score**
High	7-9 points
Medium	3-6 points
Low	0-2 points

2. Actual scores then are compared to norm averages. Norms are numerical summaries of the behavioral responses of a group of people to standard stimuli (an instrument). For training purposes, they should be presented as descriptive statistics: average scores or percentiles of various groups of people. These descriptions give the respondents a framework with which to compare their scores with those of various groups of people. A respondent can learn from seeing how his or her own highs and lows compare with those of others and how he or she ranks in comparison with others. It is important to emphasize to the participants that norms are not standards and should not be misinterpreted as such. Normative data should be used as nonevaluatively as possible so as to avoid any implication that there are "right" ways of responding or that there is a profile that the participants "should" have.

The major concern related to norms is that they be based on a relevant reference group. Usually this means that norms of locally based groups are most meaningful to participants, but tables included in instrument manuals may be useful if the groups represent the participants in some way. For example, the *FIRO-B* development group is not described in the manual, but there are averages available for members of twelve occupational groups ranging from traveling salesmen to teachers.

Average Scores of the *FIRO—B* Norm Group

Average Scores of the FIRO-B Norm Group

	INCLUSION	CONTROL	AFFECTION	
Expressed Behavior	5.4	3.9	4.1	13.4
Wanted Behavior	6.5	4.6	4.8	15.9
	11.9	8.5	8.9	29.3

In this phase, we consider scores that are discrepant from the norm by two or more points to be significant for purposes of discussion.

3. *Column* scores are examined to see the significance of inclusion, control, and affection scores by their relative importance to one another. For example, if the highest score is for control, control issues are the most important to that individual; if the second highest score is for inclusion, it is the second most significant concern for the individual, and so on. A second part of this step is to look at the column scores in relation to the norm scores and to compare the individual scores with the totals for each of the columns. If there are more than three points of discrepancy, they are worthy of discussion.

4. Next, the scores are examined by *row*—the expressed and wanted aspects. The first comparison is the relative importance of expressed behavior versus wanted behavior in terms of which is a more characteristic or logical pattern for the individual. The second comparison is the actual score in relation to the norm score totals for the two dimensions. This process allows individuals to see their scores in relation to the scores of others. A discrepancy of more than four points is worthy of discussion.

5. This step deals with what we call the Social Interaction Index. It is derived from adding either the columns or the rows to arrive at the sum of all six cells. This trait then is viewed in relation to twenty-nine points, which is the norm for the sum of the scores for the *FIRO-B.* If an individual's score is five points higher or lower, it is significant for discussion.

6. An interpretation is made of the "fit" between the profiles of two participants. For example, the expressed control of one person is compared with the wanted control of the second person, and the compatibility of their behavior is discussed.

In the second phase, dyadic interpretation, the facilitator asks individuals to form pairs. Once the dyads have been established, they are directed to exchange scoring sheets so that they are interpreting each other's scores. The dyads can be allowed five to seven minutes for A to interpret B's scores while B remains silent as a stoic, noninformational receiver. When that interpretation is completed, A and B exchange roles and B interprets A's scores while A remains silent. When this five-to-seven minute interpretation is completed, the dyads can move into a five-to-ten minute discussion of the instrument in which the two members share the personal impact of their scores. An optional, total-group discussion then can focus on generalizations.

6. Posting

The sixth major phase in presenting an instrument is posting. It usually takes from five to eight minutes. Posting scores on newsprint sheets or on chalkboard has the potential to dissipate some of the concerns that people have about negative values and lack of social desirability for any particular score. At the same time, it can generate additional, useful data for the group to process, including group and subgroup profiles. Generating scores and then posting them for discussion can be particularly effective in dealing with subgroups within a large workshop.

In some cases, it may be important to emphasize that participant's scores will not be reported to their managers. As with any personal information in training, any participant has the right to choose not to post his or her scores, although, in practice, this rarely happens if the norms of openness, sharing, and experimentation have been established.

7. Processing

The final and perhaps most crucial phase of instrumentation is processing. Group processing of the data generated by an instrument has the potential to simultaneously defuse negative affect and promote integration of the concepts. It generally will take from fifteen to twenty minutes to process and begin to integrate the data from the *FIRO-B* with a typical subgroup. Sufficient time must be allowed for this critical step, and the amount of time required will vary, depending on the nature and complexity of the data and the sophistication or receptivity of the participants.

The processing phase allows participants to compare their scores and to develop an understanding of how their scores "fit" with their self-images and others' images of them. During this stage, the facilitator should take care to legitimize differing behaviors, orientations, and perspectives. At the end, he or she also can solicit feedback on the way in which the instrument was administered.

The optimal size of the group for processing data generated by an instrument varies from six to twelve participants. In a group composed of participants who are strangers to one another, the smaller size increases the potential for individual speaking time. In an intact group, the goal is to include as many people as possible in order to maximize the common exposure to the information being shared.

Questioning

Some questions that are effective in processing are:

- Which scale scores seem to fit your self-concept most accurately?
- Which scale scores seem to fit your self-concept least accurately?
- Based on the common history of the members of the group, which scale scores seem to be most/least like those that other group members would have predicted for the individual whose scores are being discussed?
- What value does each individual place on a high or low score for a particular trait? For example, the *Survey of Interpersonal Values* instrument produces the following six scales: support, conformity, recognition, independence, benevolence, and leadership. For each individual, is a high score on conformity socially desirable or undesirable? This question promotes a forum of value clarification that provides an opportunity for disclosure (what I hold valuable) and awareness (what other members of the group see as valuable).

Potent Scores

Occasionally, instruments produce feedback that is simultaneously accurate and discrepant with an individual's self-concept. For example, a score of low benevolence on the *Survey of Interpersonal Values* may be disturbing to a minister; a low score in time competence on the *Personal Orientation Inventory* may be disorienting for an executive who sees himself or herself as being extremely competent; and a high score in expressed control on the *FIRO-B* may be disconcerting for a teacher who sees himself or herself as egalitarian. In leading the processing phase, the facilitator must be tuned in to scores that are disorienting, i.e., not easily integrated. This situation often can be handled most effectively by a one-to-one discussion with the person after the group-processing session.

It is the facilitator's responsibility to assist in the integration of all data generated by instrumented feedback. This responsibility may, in some cases, mean that one promotes dealing with disconcerting feedback (as opposed to permitting participants to discount it). It often takes days for individuals to accept dissonant feedback. Conversely, instruments can produce inaccurate data that should not be accepted indiscriminately.

An important caveat in helping participants to interpret their scores is that the trainer, consultant, or facilitator must recognize and state that the scores obtained by individuals on any instrument are the result of their answers to a series of questions at one point in time, and that such scores should not be treated with undue reverence. Such responses typically change over time, for a variety of reasons. The individual's interpretation of the question the next time may affect his or her answer, a variety of experiences may change the person's self-perception, and so on. HRD professionals are encouraged to use and to present instruments simply as one additional means of obtaining data about individuals, with all the risks and potential payoffs that any other data source would yield.

FACILITATOR STYLE

It is important for the facilitator to develop a psychological atmosphere that is conducive to participants' receiving instrumented feedback readily. The climate should be nonclinical, open, and experimental. The facilitator's style should be light rather than heavy and should not convey the impression that this is a deadly serious business that is going to yield some delicate data. In "selling" the instrument to the participants, the facilitator should attempt to induce a psychological set toward frankness in responding to the instrument items.

The Gestalt concept of *presence* can be useful in thinking about one's style in relation to instrumentation. The facilitator who has presence exhibits confidence, demonstrates that he or she is "on top of" the situation, and appears to be organized and alert. It is somewhat akin to charisma. People tend to follow such a person's instructions without challenge. If the facilitator lacks presence, the participants may tend to ask many questions, nitpick about items, and cast aspersions on the validity of the procedure. Being present in this context presupposes that the facilitator is well prepared to use the instrument (knows the procedure that he or she is about to direct, is comfortable in explaining the theory related to the scales, and is flexible in managing the learning situation).

In introducing the instrument, the facilitator should be sensitive to the possible emotional impact that the idea of "being tested" may have on participants. He or she should defuse the experience by relating the instrument to the learning goals of the training event and by pointing out that the intent is to be instructive rather than diagnostic. As has been said before, the facilitator can model openness by interpreting his or her own scores and by soliciting feedback about them.

The following list can enhance the effectiveness of the facilitator in using instrumentation.

Do's and Don'ts in Using Instrumentation

Do's	Don'ts
Do complete the instrument yourself first.	Don't use the word "test."
Do tell participants how the instrument aids the goals of the training.	Don't give instructions while participants are reading.
Do encourage participants to be open in describing themselves on the instrument.	Don't give too many instruments at one time.
Do allow plenty of time for processing.	Don't put undue pressure on participants to publish scores that may make them appear to be "sick."
Do watch for participants who may be experiencing difficulty in integrating their scores with their concepts of themselves.	Don't diagnose participants' weaknesses for them.
Do solicit feedback on your style: particular things you did that helped and impeded learning.	Don't label participants.

INSTRUMENTS AS PART OF OVERALL TRAINING DESIGNS

Instruments are highly useful components of training designs, especially when used in conjunction with structured experiences, which permit participants to share their data, receive feedback and consensual validation, and deepen their understanding of the theory behind the instrument.

Implicit in this discussion is the idea that training means developing new behaviors. A complete learning model incorporates *doing* as well as being or understanding. Useful training is designed to be transferable from the training setting to the real world.

Initially, in order to validate an instrument, the facilitator usually will divide the participants into subgroups to focus on common behaviors in each subgroup. To carry the use of the instrument further for skill development, the focus must move to the individual. For this purpose, heterogeneous subgroups are more desirable, because the unique behaviors of individuals are easier to see when they do not become submerged in the set of behaviors common to all subgroup members. Furthermore, people typically find it easier to perceive a behavior that is clearly different from their own.

In using an instrument to “tie down” observable behavior, the scales, even though behaviorally validated, may be inadequate because not enough concrete detail is provided by the scale label. Therefore, it is helpful to focus on individual items, which are more clearly related to specific behaviors.

Once the instrument has been accepted and concrete behaviors of individuals identified, it is up to the facilitator to provide participants with the opportunity to define specific behavioral change goals and then to provide a structure for behavioral skill practice that can lead to goal attainment. “Practice” can be provided in the form of discussions, rehearsals, role plays, or some combination of these. The length of time and depth of detail involved may vary considerably.

LABORATORY/WORKSHOP DESIGNS

In designing laboratory experiences in which instruments will be used, it is wise to consider several things. One major concern in using more than a single instrument in any training design is that the traits measured be supplementary. The two instruments should not measure exactly the same characteristics; this adds little or no new information. However, the traits measured should not be so diverse that participants have a difficult time putting together congruent pictures of themselves. For example, two very similar personality inventories would not be used in the same design because the data obtained from the second would be redundant. On the other hand, using two diverse instruments would provide a disconcertingly diverse focus to the feedback.

A second concern is that the facilitator anticipate the type of interaction affect that will be produced by the instrumented feedback. He or she should avoid overloading the participants with data that will generate a heavy, emotional atmosphere that could move the interaction to a nonproductive level. It is equally important to avoid instruments that could produce feedback that is too sophisticated for the participants to handle well and to avoid producing feedback that is inappropriate to the goals of the learning experience or that may cause a shutdown of interaction. For example, it would not be productive to administer a complex personality inventory to a group of managers during a team-building workshop.

The key to managing the integration of data productively when incorporating instrumentation in a training design is the careful selection of instruments and the consciousness of processing issues as they relate to the interaction in the group. (The discussion that follows tells how to evaluate and select instruments.) If the decision is made to use two complementary instruments, the design must include time to process each of the instruments and time to process the relationship between the two sets of scores produced.

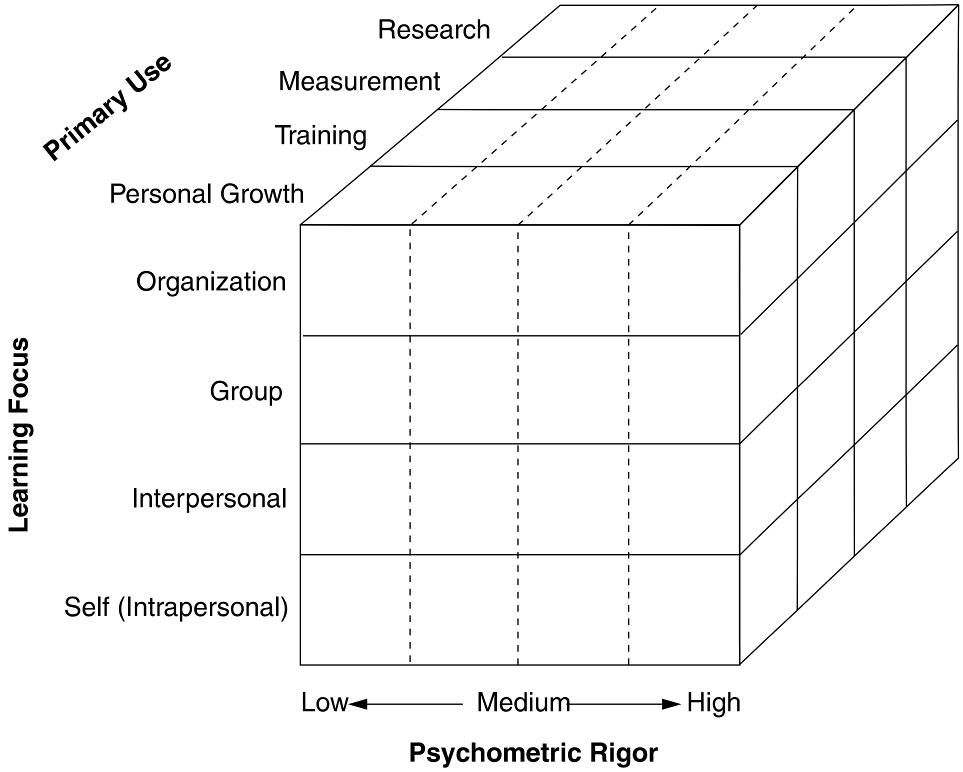
Many instruments have complementary pairings with structured experiences. Pfeiffer & Company’s *Annual* series, for example, contains many structured experiences that either are built around instruments or utilize them to introduce or reinforce the concepts to be learned. The well-integrated use of instruments and structured experiences tends to reinforce learnings and crystallize concepts.

EVALUATING AND SELECTING INSTRUMENTS

Knowing how to present an instrument is not enough to ensure that its use will be successful. The other half of the coin is knowing how to evaluate and select the instrument(s) that will be used. The number of published instruments available to HRD professionals has become so great that choosing the right one for a particular use is difficult. This discussion contains a model and guidelines for selecting instruments for various purposes.

A MODEL OF INSTRUMENTATION

A model of instrumentation could be created using any number of dimensions; the one that follows is a simple model covering the significant dimensions of (a) psychometric or technical rigor, (b) primary use (training, research, etc.), and (c) learning focus (individual, group, etc.).



The Basic Dimensions of Instrumentation

TECHNICAL CONSIDERATIONS [PSYCHOMETRIC RIGOR]

It is important to have a knowledge of technical topics related to instrumentation before studying instruments for possible use in training settings. The list that follows includes most of the technical concerns that the facilitator needs to take into account when selecting an instrument.

Technical Considerations in Using Instruments

Validity	What does the instrument measure? Will the data be useful?
Reliability	How accurate or stable are the scores derived from the measurement?
Objectivity	Is the scoring dependent on the judgments of the scorer or is there a standard key? Can the instrument be scored by untrained people such as participants?
Theoretical Base	Is the instrument based on a workable model?
Behavioral Orientation	Are the scores derived from the respondents' descriptions of their behavior?
Observability	Can the scores be related to the observable behavior of respondents?
Special Training	How much professional preparation is required to use the instrument?
Language	Is the instrument written at an appropriate reading level? Does it use a special vocabulary or jargon?
Sophistication	What background is required to use the instrument?
Complexity	What kinds of feedback can be derived from the items? How complicated is the interpretation?
Supplementation	Will the instrument yield data that add to what the participants already know?
Adaptability	Can the items be adapted/amended to fit a particular situation?
Transparency	How obvious is the rationale underlying the items?
Fakeability	How easy is it for respondents to manipulate their scores?
Norms	Are relevant norms available?
Availability	Can the instrument be obtained easily?
Copyright Restrictions	Can the materials be photoreproduced or edited without special permission?
Time Required	How much time is needed to prepare, administer, score, and interpret the Instrument?
Expense	What is the cost of the materials, scoring, analyses, and background documents? Are these reusable materials?
Special Materials	Does the instrument require that any special apparatus be obtained and set up in advance?
Noxiousness	Would the items—or the scale itself—offend intended respondents?
Scoring Complexity	Can the instrument be self-scored? Are electronic/clerical options available?
Data Reduction	How many scores are derived? Can these be summarized for ease in interpretation?
Handouts	Are easily read, interpretive materials available for the facilitator to distribute to respondents?
Familiarity	How likely is it that participants will have responded to this instrument before?

The most critical technical considerations are validity and reliability; these are discussed in detail below.

Validity

The basic questions of validity center around what traits are being measured, what the scores mean, and how useful the data are. Six types of considerations are suggested, and these are outlined in the following paragraphs.

Content Validity. Content validity, or face validity, is the minimum validity acceptable. It refers to the first impression the user has of the instrument—whether the instrument appears to be measuring the things it is supposed to measure. A careful examination of the instrument should show a reasonable, logical, clear connection between the instrument and what is measured. For example, a selection test should include simulated job samples; an appraisal form should include job elements deemed important to performance by a formal job analysis. The user is cautioned not to have much confidence in or make important decisions based on instruments that do not appear to have content validity. Furthermore, respondents and clients are very likely to question the instrument. Although it is possible for an instrument to measure something it does not appear to be measuring, such occurrences are rare. Of course, it also is possible for an instrument to fail to measure something that it appears to be measuring.

Predictive Validity. Normally, one would expect more than just content or face validity, and there are other ways to check on whether an instrument measures what it claims to measure. One of these is predictive validity: some concrete evidence that instrument scores are related to certain measurable characteristics or behaviors of the persons tested, as predicted by the logic behind the instrument. If the instrument is a measure of the ability to learn to fly, people who score high on it should learn faster and fly with more skill. If they do not live up to this prediction, the validity of the instrument may be questioned. There are a number of instruments that are used to select people for potential occupations. However, there are many instruments that are not appropriate for selection or prediction. Scales that measure introversion-extroversion, dominance needs, styles of relating to others, and personal temperament, for example, are not intended to predict performance. For such instruments, there are other criteria for determining validity (these criteria can be used to assess predictive instruments as well). One such criterion is concurrent validity.

Concurrent Validity. Concurrent validity is established by showing that there is a relationship with a present criterion, i.e., the measure is shown to relate statistically to an independently determined, concrete measure obtained simultaneously. For example, those who presently are engaged in an occupation and are doing well should score higher on an aptitude measure than those who are doing poorly. On a scale measuring political conservatism, student members of the Young Republicans should score higher than student members of the Young Democrats.

Convergent Validity. If an instrument is measuring what it is supposed to measure, it should relate positively to other measures of the same thing (i.e., they all should be “converging” on the same trait). Because they all are sampling from the same behavioral domain but from slightly different angles, they should have substantial overlap. Similarly, an instrument is said to have convergent validity if, in numerous cases, it is statistically shown that a ratee is rated similarly by several raters on the same dimension (i.e., the raters agree or “converge” on the rating).

Discriminant Validity. The other side of the coin from convergent validity is whether an instrument is *un*related to scales that measure traits it is *not* supposed to be measuring. An instrument also is said to have discriminant validity if it is shown statistically that a ratee’s different dimensions are rated differently by a rater (i.e., the various dimensions “discriminate” to yield a proper, separate, and different rating for each dimension).

Construct Validity. Many instruments that measure personal or group characteristics are related to, issue from, or are the basis for theories. The scale measures a concept or construct, such as achievement motivation, that takes on full meaning through the theory. The theory attempts to explain what childhood experiences lead to high achievement motivation, what the preferred leisure activities are for the person with a high need for achievement, the types of situations in which such a person will do well, the types in which he or she will do poorly, and so on. If the scale actually measures the kind of achievement motivation explained by the theory, high scorers will report those predicted childhood experiences more than low scorers will; high scorers will perform better in the prescribed situations than low scorers will, and so on. If the predicted pattern of relationships is found, both the instrument and the construct will have been validated simultaneously. If the predicted pattern is not found, the user must determine whether the fault lies with the instrument, the theory, or the testing of the instrument and theory. Usually *some* hypotheses are confirmed, so that clues are available concerning what may be right and what may be wrong with both the theory and the instrument.

An instrument is said to have construct validity if statistical and logical tests (the multitrait-multirater approach) show it to have convergent and discriminant validity.

Validity is not inherent in instruments. The user must validate each instrument for the specific uses for which he or she will employ it. This means studying the validity evidence available in the instrument manual or other supporting documents and carrying out one’s own evaluation of the usefulness of the instrument. In one sense, validity resides in the user rather than in the items; using the same instrument in a variety of ways with different people can build experience and supply data that can add validity to an instrument. In another sense, validity is situation specific; it resides not so much in the instrument as in the particular use of it. For example: In *training* the validity of the scale is measured by whether it will help participants to learn more effective behavior; in *organizational assessment* the consideration is whether the instrument taps those process dimensions that are correlated with production. In personnel selection the question is one of predictive—or discriminative—validity, i.e., whether the instrument is significantly related to a meaningful success criterion; and in research the major concern is the theoretical constructs being measured, i.e., whether the scale measures the concepts derived from theory sufficiently well to permit meaningful tests of hypotheses derived from the model used.

Reliability

A basic question that must be asked of any measuring device is "Is it consistent, i.e., will the instrument yield the same results for the same people on separate occasions, given the same conditions?" This is *test-retest reliability*. As an example, imagine a ruler made of taffy. Each time a person measures something with the ruler, it has become shorter or longer through handling. The person cannot decide the effects of different kinds of fertilizer on the height of certain plants because she cannot measure the height in a reliable or consistent way.

Another way of checking reliability is to split the instrument into two equal parts to see if the two parts yield highly similar results (as they should). This is called *split-half reliability*. If an instrument has several separate scales, each scale must be split in half by randomly sorting items into two groups.

Low reliability means that scores are due to something other than the trait one wanted to measure—there is a lack of consistency in the measuring. These factors can be called "error" or "error of measurement." An unreliable instrument *cannot* be valid.

Reliability indices give scores that range between 0.00 and 1.00. A reliability of 0.00 means that nothing consistent is being measured and that scores are being determined by error, chance, or random fluctuations in conditions. A score of 1.00 means that error, chance, or other extraneous conditions have no effect on the measurement procedure.

A reliability index of .85 or higher generally is considered to be effective for all purposes. An index of between .60 and .85 indicates reliability that is effective for measuring and talking about groups of people and in doing research but may be too low for placing high confidence in an individual's score. Reliability scores below .60 indicate that the variability from extraneous factors is so great that the instrument (a) probably should not be used to diagnose an individual and (b) should be used with caution to talk about groups of individuals. Even though such an instrument may be useful in research, the error in it may hide or mask significant differences or changes that exist.

Higher reliability is required of an instrument used for personal feedback because as reliability goes down, error of measurement goes up. Error of measurement is an index of how close a person's hypothetical "real" score is to the one that he or she received from the instrument. A large error of measurement means that the person's scores would vary widely if you gave him or her the instrument a number of times; a small error of measurement means that the scores would be very similar over a number of administrations for the same person.

Measures that have relatively low precision can be utilized to study group phenomena. Averages of groups have more stability than do individual scores. An instrument used to measure employee morale over time need not be so accurate as an inventory intended for career-development counseling. A comparatively crude index of the level of interpersonal trust in a group or team can be employed if the interest is in studying how the average changes as the group develops.

Reliability refers to whatever is left after the error of measurement has affected the scores. The term has been used to describe two rather different elements: homogeneity of the statements and stability of the scores.

Homogeneity of Statements

Many indices of reliability measure the extent to which statements in the scale measure the same characteristics. If two statements measure the same thing, as scores go up on one, they go up on the other, and vice versa. They vary together, or “covary.” Homogeneity is a characteristic of a good measure; if an instrument purports to measure a certain trait, all the parts should be measuring facets of that trait. Although scores of the various statements on the scale should vary together, they should not vary together perfectly, because then there would be duplicate measures of exactly the same thing, and that would add no new information.

At the other extreme, one would not want statements that are unrelated, because it is impossible to say what a score measures, and responses might be attributable to guessing or chance. The ideal of homogeneity is a scale with statements correlated on the average between .15 and .50. Such statements can be viewed as measuring various aspects or facets of the same thing.

Stability of Scores

The aspect of reliability measured by the test-retest method is stability. The idea is that one should obtain the same score on repeated measures. If scores are attributable to guessing, chance, momentary reaction to ambiguous wording, or other extraneous factors, they will vary from one administration of the instrument to another. The correlation coefficient is used to calculate test-retest reliability.

Reliability and validity are closely related considerations. An instrument cannot be valid without some precision, but a highly stable measure may not necessarily be useful. Reliability is necessary but not sufficient. The major concern is the utility of the outcomes from the use of the instrument.

Transparency and Fakeability

At least two other aspects of psychometric rigor are worth examining. First, one may ask to what degree the instrument is transparent. That is, how obvious is it that a leadership-style instrument is intended to measure leadership? If it is very clear, it will be easier for people who take the instrument to understand the results and they will be more likely to trust those results. This can be a benefit in many situations, but it is counterbalanced by the problem of fakeability. That is, the more transparent the instrument, the easier it usually is to fake the results.

Fakeability becomes a problem when it is clear that certain results are more desirable than others. For example, when managers fill out a particular leadership questionnaire, they typically report that they believe in behaving considerately toward subordinates and that they place low to moderate emphasis on giving directives and orders. When their subordinates fill out a modified version of the same instrument, the results usually are quite different. Because contemporary U.S. managerial culture favors a person-centered, less directive style *in theory*, U.S. managers usually describe themselves in this way even though that is not how they actually behave. The more transparent the instrument, the more prone people who take it are, consciously or unconsciously, to fake certain responses.

The technical considerations presented here about instrumentation are intended to ensure that the data to be used in HRD work are not misleading. The overriding concerns are validity and reliability, and information about these characteristics often is missing or inadequate. It is incumbent on the facilitator to explore the technical characteristics of any instrument that he or she is considering using. The facilitator must be prepared to answer any technical questions straightforwardly and must determine the validity of the instrument for his or her particular purposes.

PRIMARY USE

Many different uses exist for instruments. Four primary uses—personal growth, training, measurement, and research—are shown on the model, although these categories are not mutually exclusive.

Personal Growth. Instruments intended to be used for personal growth should promote self-analysis by providing the user with valid information about himself or herself. This information may be exclusively internal (attitudes) or external (behavior). Many instruments also provide clear guidance on what the user can do about the results, either through a program for self-development or through manuals and materials for the facilitator.

Training. Training instruments are similar in purpose to those designed for use in exploring personal growth, but are oriented far more toward behavior than toward attitudes or personality. Training instruments are directed toward supporting specific behavioral changes.

Measurement. Some instruments are professional tools for use in assessing organizations, groups, relationships, or individuals. For example, some measure vocational interests and aptitudes; some diagnose interpersonal relationships in various contexts (families, groups, etc.) for the benefit of the professional counselor.

Research. Some instruments are intended primarily for research use. Although valid measurement is quite important, the purpose is not to use the data in a pragmatic sense but to build, support, or test some theory. Many instruments used for diagnostic measurement formerly were research tools. Most research instruments never are published or widely disseminated, although there is now an index of unpublished research instruments so that scholars and scientists have access to one another's tools. However, many research instruments are designed especially for one specific project and are essentially useless for any other purpose.

These four purposes for using instrumentation are not perfectly exclusive. Many instruments can be adapted for more than one use. A training instrument may well be used for personal growth with only minor modifications.

LEARNING FOCUS

Most instruments have one of four primary learning foci: organization, group, interpersonal, or self (intrapersonal). Of course, most can be applied at more than one level. It often is possible to use aggregate scores by pair, group, or organization. Before aggregating individual scores, it must be considered whether the aggregate score has meaning. For example, it might be interesting to see how many individuals in a group or organization fall into each category on a particular survey, but it might make no sense to find an average group score.

PSYCHOLOGICAL DEPTH AND IMPACT

Another major dimension along which an instrument can be rated is the psychological depth or intensity of psychological impact it may have. For example, there are several instruments that measure the extent to which an individual feels controlled by or able to control his or her environment (external versus internal control). Much research has shown this measure of personality to have important behavioral implications (Rao, 1985; Rotter, 1966), including feelings of depression on the part of those individuals with strong external orientations. With such persons, the use of an instrument measuring internal/external control (with resultant interpretations and implications) could have severe and undesirable effects unless the facilitator is skilled in paying attention to the needs of individual participants and in dealing with the ramifications of interpretation. At the other extreme, an instrument that deals with behaviors that are clear and easily changeable is likely to have little psychological impact.

An instrument that has deeper psychological impact requires skills of a high order on the part of the facilitator in order to avoid potential psychological harm to participants. Harrison (1970) says that the HRD practitioner should intervene at the *least* depth needed for the diagnosed problem. This is consistent with standard medical practice: intervention to cause the minimum trauma consistent with the diagnosis and needed cure. Thus, if two instruments are available to measure the same traits, the best choice usually would be the one that provides guidance for behavioral change without probing too deeply into the participant's psychological make-up.

OTHER CONSIDERATIONS

The purpose for using an instrument should be examined carefully and stated clearly, and the learning focus should be clear. Next, an instrument must be assessed in terms of four secondary criteria: administrator concerns, respondent concerns, pragmatic concerns, and client concerns.

Administrator Concerns

The trainer, assessor, or consultant who administers an instrument has some special concerns. First, the administrator must consider the *special training* he or she needs to use the instrument effectively. Some instruments require extensive training in proper scoring and interpretation of results. Others require considerable conceptual background to be able to present the theoretical basis. For example, the use of *the Survey of Organizations* (SOO) requires that the consultant have a good grasp of Likert's (1971) theory of organizations. Many psychological assessment instruments are sold only to individuals who are "certified" psychologists.

A second concern for administrators is *objectivity of scoring*, the degree to which scoring is based on trained judgments versus a standard "key" or index of rules that assigns specific scores to specific responses. Objectively scored instruments avoid the need for special scoring skills and may make it easier for respondents to accept the accuracy of the score.

A third concern is *scoring complexity*. Even an objectively scored instrument may have many complicated scales. A well-constructed scoring key can reduce this concern, as can a simple, self-scoring format or a procedure by which marking one's response automatically enters it on a hidden scoring form by means of NCR or carbon paper. When scoring complexity is high, another alternative may be an electronic scoring service with rapid turnaround and feedback, as is now provided for many organizational surveys.

A final administrator concern is the need for *special materials*. For example, some instruments require audiovisual materials; some require special marking pens, etc. Even instruments that must be administered in a large-group setting can present problems if only a few people can be spared from their jobs at any one time.

Respondent Concerns

Before using most instruments, the facilitator should evaluate how appropriate they are for a particular respondent group. At least three specific concerns are relevant. First, the administrator must consider whether the instrument—or certain items on it—could be *noxious* to users, that is, whether the instrument could provoke negative reactions. For example, Peters, Terborg, and Taynor's "Women as Managers Scale" (WAMS)

(Terborg, 1979) would be inappropriate for use with managers from a country in which women are informally but absolutely excluded from managerial roles in organizations.

A second concern is respondent *familiarity* with the instrument because some of the group members may have taken it in the past and recall their scores or the items. Many instruments that normally are not particularly fakeable can be faked easily by respondents who are familiar with the items, scoring, and rationale. Even when respondents do not intend to fake their scores, their familiarity with the instrument may bias those scores.

The third concern is the *language level* that respondents must have in order to respond adequately to the instrument. Some instruments use language that assumes a higher level of education than most respondents may possess. Others may use jargon or technical terminology with which some potential user groups may not be familiar. For example, to demonstrate the cultural and language bias inherent in many "IQ" tests, Morgan (1981) developed the Central West Virginia Cultural-Awareness Quiz, designed so that those who do not share the jargon of Appalachia do poorly. This clearly illustrates the issue of language level.

Pragmatic Concerns

Also of concern in using instruments are several practical issues. Obviously, the perceived attractiveness of instrument, the clarity of its format, and the ease in responding to or scoring it contribute to the cooperation of group members in utilizing the instrument. An important consideration is *expense*. The administrator should not overlook possible scoring-service or scoring-supplies expenses, the cost of background materials needed to understand the interpretation, special licensing fees, and whether the materials are reusable. Related to this is the issue of *copyright restrictions,* not only whether the instrument can be reproduced but also whether it can be modified for a special purpose or a different user population. *Adaptation* may not be practical, even if it is permitted. It might be easier to choose a different, more appropriate instrument or even to create a new one. *Accessibility* is another issue. Some publishers may have a reputation for slow delivery, or some instruments may be circulated on a small scale and may be difficult to locate. Finally, the *time required* to prepare to use the instrument, to administer it, to score it, and to interpret it must be considered. When a great deal of time is required, an instrument may become impractical for certain uses or simply may be too costly in terms of total time and costs incurred.

Client Concerns

The concerns of the person who actually will use the results of an instrument must be considered. Individuals, groups, or organizations involved in training; individuals in counseling; and individuals, groups, or organizations undergoing formal assessment need to know the meaning of their results. If scores can be related to observable behavior, it will greatly facilitate the use of the instrument for personal growth and training purposes. If the instrument is behaviorally oriented to begin with, users will be better able to see a relationship between the instrument and their own behavior.

If the instrument generates many, separate, complex scores, interpretation can become quite difficult, even for a professional. If the instrument is complex, a greater burden is placed on the administrator (trainer or consultant). Simple, straightforward, *interpretive materials* can relieve some of that burden. Data sometimes must be *reduced*, and it can help to draw a chart or graph of the results. The use of *normative data* also is extremely helpful. Finally, the existence of a clear and compelling *theoretical base* can be critical if the results are to make an impression on the users and affect their behavior.

In examining the applications and uses of instruments, we have identified a broad range of dimensions that need to be considered. The following chart reflects the relative amount of concern each dimension warrants for the purposes of personal growth, training, organizational assessment, personnel selection, measurement, and research applications.

WHAT TO LOOK FOR IN AN INSTRUMENT

	Instrumentation Application					
Area of Concern	**Personal Growth**	**Training**	**Organizational Assessment**	**Personnel Selection**	**Measurement**	**Research**
Administrator Concerns						
Special Training How much professional preparation is required to use the instrument?	High	High	High	High	High	High
Objectivity of Scoring Is the scoring dependent on the judgments of the scorer, or is there a standard key?	Medium	High	High	High	High	Medium
Scoring Complexity Can the instrument be self-scored? Are electronic/clerical options available?	High	High	Low	Medium	Low	Low
Special Materials Does the instrument require that any special apparatus be set up in advance?	High	High	Medium	Medium	Medium	Medium
Respondent Concerns						
Noxiousness Would the items—or the instrument itself—offend intended respondents?	High	High	High	Medium	High	High
Familiarity How likely is it that respondents will have responded to the instrument before?	Medium	Low to Medium	Low	Medium	Low	High

	Instrumentation Application					
Area of Concern	**Personal Growth**	**Training**	**Organizational Assessment**	**Personnel Selection**	**Measurement**	**Research**
Language Is the instrument written at an appropriate reading level? Does it use a special vocabulary or jargon?	High	High	High	High	High	High
Pragmatic Concerns						
Expense What is the cost of the materials, scoring, analyses, and background documents? Are these reusable materials?	Medium	Medium	High	Medium	High	Medium
Copyright Restrictions Can it be photocopied or edited without special permission?	Medium	High	Medium	Medium	Medium	Medium
Adaptability Can the items be adapted/amended to fit a particular situation?	Medium	Medium	High	Low	High	Low
Accessibility Are the materials readily available?	Medium	Medium	Medium	Medium	Medium	Medium
Time Required How much time is needed to prepare, administer, score, and interpret the instrument?	High	High	High	Low	High	Medium
Client Concerns						
Observability Can the scores be related to observable behavior of respondents?	High	High	Medium	Low	Medium	Low

Instrumentation Application

Area of Concern	Personal Growth	Training	Organizational Assessment	Personnel Selection	Measurement	Research
Behavioral Orientation Are the scores derived from the respondents' descriptions of their behavior?	Medium	High	High	Low	High	Low
Interpretive Materials Are easily read interpretive materials available to be distributed to respondents?	High	High to medium	Medium	Low	Medium	Low
Data Reduction How many scores are derived? Can these be summarized for ease in interpretation?	High	High	High	Medium	High	Low
Normative Data Are relevent norms available?	Low	Medium to Low	Low	High	Medium	High to Medium
Theoretical Base Is the instrument based on a workable model?	High	High	High	Low	High	High

It may not be possible to rate any one instrument on every factor discussed here, but it should be possible to obtain an average rating using two or three factors in each category and, in this way, to make a reasonable, overall judgment regarding any instrument. The check sheet that follows can be filled out and filed with the instrument for quick reference.

Instrument Evaluation Check Sheet

Title:____

Author:____

Source:____

Primary Use: ☐ Personal Growth ☐ Training ☐ Organizational Assessment
☐ Personal Selection ☐ Measurement ☐ Research

Learning Focus: ☐ Self ☐ Interpersonal ☐ Group ☐ Organization

Psychometric Rigor

Reliability	low	moderate	high
Validity	low	moderate	high
Transparency	very transparent	somewhat transparent	not at all transparent
Fakeability	easily faked	somewhat easy to fake	very difficult to fake

Administrator Concerns

Special Training	requires much special training	requires some special training	requires no special training
Objective of Scoring	completely subjective judgment	partly subjective judgment	completely objective
Scoring Complexity	extremely complex	moderately complex	not at all complex
Special Materials	requires extensive special materials	some special materials required	no special materials needed

Respondent Concerns

Noxiousness	clearly offensive to users	somewhat offensive to users	not at all offensive to users
Familiarity	considerable familiarity	some familiarity	no familiarity
Language	not appropriate for users	partly appropriate for users	fully appropriate for users

Pragmatic Concerns

Expenses	very costly	moderately costly	not at all costly
Copyright Restrictions	cannot be reproduced	can be reproduced with restrictions	can be freely copied
Adaptability	not practical	difficult but possible	easy to adapt
Accessibility	not at all accessible	accessible with some difficulty	readily accessible
Time Required	long	moderate	short

Client Concerns

Observability	scores not related to observable behavior	indirectly related to observable behavior	scores clearly related to observable behavior
Behavioral Orientation	scores not derived from behavior	some scores derived from behavior	scores derived directly from behavior
Interpretive Materials	none available	available, not very useful	available, excellent
Data Reduction	multiple scales hard to relate to one another	some data reduction possible	scales easy to summarize

Normative Data	no norms	limited normative data available	extensive norms provided
Theoretical Base	no theory base	some theory references researchable	comprehensive, clear theory base provided

Once the facilitator has applied the criteria described in this discussion to instruments that he or she is considering, the final question is: Is this instrument appropriate for the purpose for which it is intended and the needs of the respondents? As is discussed elsewhere in this volume, no instrument should be used to measure anything other than what it purports to measure. An instrument cannot be substituted because it is “close enough” to what is needed without altering the response data and the interpretation of that data, perhaps with serious consequences. Therefore, if the instrument under consideration does not satisfy the selection criteria, an alternative is to develop a new instrument that will meet the needs of the respondents and satisfy the purpose for which it is intended. Considerations and steps in the development of instruments are presented next.

DEVELOPING INSTRUMENTS

When an instrument cannot be located that meets the special needs of the group with which the facilitator will be working, an instrument can be developed. The time required for such development will be hours or perhaps a day. We will not describe the formal procedures for developing instruments of established reliability and validity for widespread psychometric application.

A DEVELOPMENTAL SEQUENCE

A number of steps are taken in generating numerical data that can be useful in HRD interventions. The following is a functional sequence in the development and administration of a new instrument.

Definition The trait to be measured is defined. Presumably the characteristic is related directly to the goals of the training.

Specification Behavioral instances of the trait are specified and are written as items for a questionnaire, or they are incorporated into the instrument in some manner, depending on the style of instrument the developer has chosen.

Scaling A response format is devised that will yield numbers for interpretation. (Various scaling options will be discussed later.)

Keying A scoring procedure is selected, e.g., weighting response numbers. The developer examines the key as it applies to the instrument to ensure that it will be functional.

Duplication The scale is published in a pencil-and-paper response format or through some other device to make it available to the participants.

Administration The administrator facilitates the understanding of the instructions for completing the instrument.

Scoring The key is applied.

Interpretation Participants are instructed on what high and low scores presumably mean. Limitations on validity and reliability are specified.

Norming Frequency distributions and averages are developed for the participants and are posted. The administrator may ask participants to post their own scores to indicate the group “photograph” and to facilitate discussion.

Critiquing The instrument can be refined by studying its utility in the training setting. Participants can be encouraged to share what they have learned.

DEFINITION

Before developing an instrument to measure something, it may be a good idea to determine whether the dimension or trait in question could be measured better by some other means or whether it is worth the trouble. For example, assume that one were interested in measuring the participation of members of a group. If participation is defined as the number of words spoken by a group member during a session, one could measure that dimension perfectly by attaching a microphone to every member and calculating the number of words spoken by each member. Granted that there are some research designs in which such information might be desired, the limitations of such a measure for use in small-group discussions is obvious. The number of words spoken provides almost no information for group members to use; they are more apt to be interested in the quality, nature, or impact of the words spoken. Most important issues in social interaction are heavily laden with such value judgments. When group members rate one another on "participation," it may be difficult to ascertain how much they are rating quantity of talking and how much they are rating other factors such as influence and leadership. It is not sufficient to assume that because a dimension such as participation is being rated reliably (i.e., group members produce the same score), the members are in agreement on exactly what they are rating. When all members of a group find agreement easy in rating some dimension, it may be a clue that the information to be obtained and the potential for important learnings may be limited (Hanson, 1981).

Although an instrument may have face validity, the question of whether or not the instrument is measuring what it is supposed to measure should be subjected to experimental studies specifically designed to provide construct validity. A minimum effort to establish such validity is to review the dimensions with the group of respondents prior to their completion of the instrument to make certain that everyone is in agreement about the definitions of the items or dimensions. The question is not whether the instrument developers agree on the items, but whether or not the *respondents understand them* (Hanson, 1981).

SCALING

After the content of the instrument has been specified, the facilitator devises a scaling procedure. Several uncomplicated and useful strategies are discussed here.

Summative Scale

The summative scale, or Likert scale, was developed by Rensis Likert (Likert, 1932). It is constructed by first devising statements that a person might use to describe himself or herself, e.g., "I am nervous right now," or "I usually let other people decide what we will do together." Some reasonable categories then are developed for the respondent to use to indicate his or her feelings. The most usual response categories are (1) yes, unsure, and no; (2) true, unsure, and false; and (3) strongly agree, disagree, and strongly disagree. The following example illustrates an instrument item in the third category.

Circle the response that most clearly reflects your feelings.

SA = Strongly agree

A = Agree

SLA = Slightly agree

U = Uncertain (undecided/unsure)

SLD = Slightly disagree

D = Disagree

SD = Strongly disagree

1. I am a very persistent and steady worker. SA A SLA U SLD D SD

Researchers have moved away from the seven-point Likert scale because of the possibility of ambiguity (e.g, the meaning of “agree” seems very similar to the meaning of “slightly disagree”; if you have only some slight disagreement with a position, it means that you basically agree with it). Others have moved away from the use of the “uncertain” or “unsure” response because it allows the respondent to avoid commitment. We propose a set of four responses: “Disagree; Unsure, probably disagree; Unsure, probably agree; Agree.” This acknowledges that a respondent may be unsure about his or her reactions and, at the same time, it measures the intensity of his or her feelings. Even though the uncertainty is acknowledged, the respondent is asked to indicate which way he or she tends to lean. On the other hand, it is possible to so indicate if one’s convictions are sure. For example:

Circle the response that most clearly reflects your feelings.

A = Agree

UA = Unsure, probably agree

UD = Unsure, probably disagree

D = Disagree

1. I find it easy to influence people. A UA UD D

A score for a respondent is obtained by assigning a number to each of the response categories and then adding up the numbers for each response indicated. The number can be assigned before the instrument is administered to the person so that the respondent places a number in the response space rather than a letter; e.g, 1 for Disagree; 2 for Unsure, probably disagree; 3 for Unsure, probably agree; and 4 for Agree. The person’s score would be the sum of the scores he or she wrote down in responding to the statements. If a statement is worded in opposition to the characteristic being measured, the scoring key is reversed so that a response of 1 is counted as 4, 2 is counted as 3, and so on. As an example, in devising an instrument to measure need for interaction with others, let us assume that we have four statements, with a person’s response recorded in front of each of them.

3 1. I enjoy being with a lot of people.

3 2. I often organize parties or get-togethers.

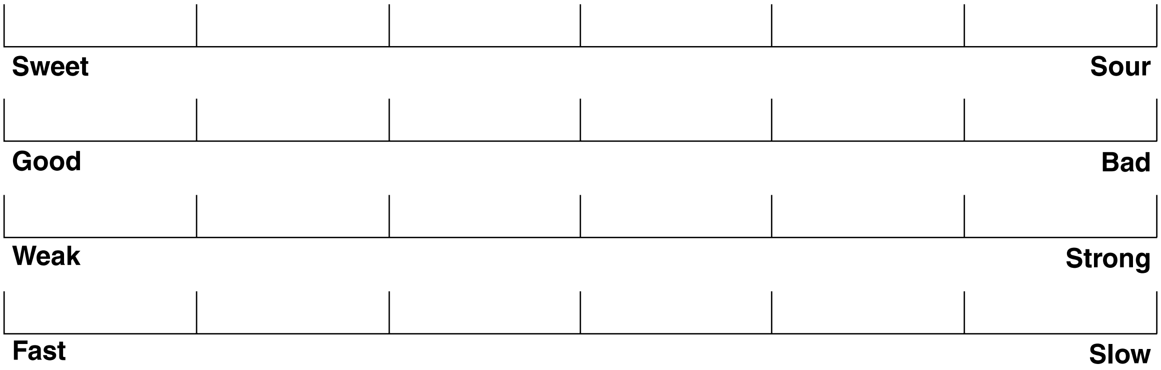
1 3. I prefer reading to talking with people.

4 4. The good life is the life of friendship, wine, and song.

Statement number 3 is reverse-worded, so the response of 1 will be counted as 4. Thus, the person’s score on the four-item scale is 3 + 3 + 4 + 4 = 14 out of a possible range of from 4 to 16. This respondent appears to be socially extroverted. This person might have responded with an “Agree” (4) to statement 1 if it had not had the words “a lot” in it. It is desirable to have some statements that are more extreme than others to help to separate those individuals who are high on a characteristic from those who are *very* high on it.

Rating Scales (Semantic Differential)

The semantic-differential approach is the most well-known use of the rating scale. In the use made by Osgood, Suci, and Tannenbaum (1957), objects (groups, organizations, practices, people, countries) are rated on a series of bipolar scales such as the following:



Sometimes the spaces are numbered from 1 to 7 and sometimes they are numbered from -3 to +3. The first system gets rid of minus signs; the second system indicates clearly whether an attitude is positive or negative. Osgood and his colleagues analyzed their data to try to reduce the number of measures to a few broad ones that seemed to indicate how people react to social objects. They found three basic dimensions that accounted for many of the ways in which people understood or reacted to things. These are: evaluation (whether the person likes it); potency (whether the person thinks it is powerful); and activity (whether the person thinks it is moving). Scales that measure the evaluation dimension are:

1. Sweet-sour
2. Beautiful-ugly
3. Tasteful-distasteful
4. Kind-cruel
5. Pleasant-unpleasant
6. Bitter-sweet
7. Happy-sad
8. Sacred-profane

9. Nice-awful
10. Good-bad
11. Clean-dirty
12. Valuable-worthless
13. Fragrant-foul
14. Honest-dishonest
15. Fair-unfair

Scales that measure the potency dimension are:

1. Large-small
2. Strong-weak
3. Heavy-light

Scales that measure the activity dimension are:

1. Active-passive
2. Fast-slow
3. Hot-cold

More evaluative adjectives are given in the first category because it is the strongest factor in people’s reactions and is used most often (it seems to be an effective measure of pure attitude or pure affect).

The basic technique can be expanded in two ways. First of all, the number of rating scales can be varied. The instrument developer can add bipolar labels that fit his or her area of interest, e.g., open-closed, risky-cautious, demanding-yielding, colorful-drab, deep-shallow. Second, the social objects that can be rated are limited only by the time, energy, and creativity of the developer. Examples of objects to be rated include: this organization, this organization ten years from now, me, my ideal self, me as others see me, our team, our relationship, this apple, the person in this group whom I admire the most, Fred, Ed, etc. An example follows.

Place a check in the space that indicates your reaction to the object.

Object: The organization

Active						**Passive**
Beautiful						**Ugly**
Bitter						**Sweet**
Valuable						**Worthless**

Object: My boss

Active						Passive
Beautiful						Ugly
Bitter						Sweet
Valuable						Worthless

As with the summative scales, numbers are assigned to ratings, e.g., the spaces from left to right are 7, 6, 5, 4, 3, 2, and 1, with the exception of the Bitter-Sweet scale, which has reversed response numbers of 1, 2, 3, 4, 5, 6, and 7. The user adds the responses to all the bipolar scales that measure the same characteristic. The total is that person's score or rating of that specific object on that specific characteristic.

Some people prefer to use an even number of points on a scale, based on the assumption that the respondents will be forced to select a rating either above or below the nonexistent middle point (the choice point for people who want to play it safe or who do not want to make the effort to discriminate). On the other hand, not having a midpoint may force a respondent to ignore a reality that he or she actually perceives. Having too few points on a scale does not allow much discrimination. Having too many points may give the false impression that the item is so refined that the number of choices is justified. In making such decisions, the designer must use his or her own judgment in terms of the purpose of the scale, the clarity and specificity of the behaviors or things to be rated, and the face validity of the items (Hanson, 1981).

The number of points on a scale to be labeled will depend on the facilitator's and/or respondents' needs for greater or lesser reliability. The more points that are labeled, the easier it is for respondents to rate them with reliability. This greater reliability would be important primarily if the scales were to be used for research purposes. In labeling a scale, it also is important not to use words that change the concept of the original item. For quick assessments (e.g., to help groups to process their meetings), the designer may choose to label only the two extremes of the scale.

Forced-Choice Scales

One of the problems with most scaling techniques is that the respondent can work at making himself or herself look good rather than at describing what he or she really is like. The forced-choice technique presents two alternatives and asks the respondent to choose which one he or she prefers. The respondent may hate them both or may think that both are excellent; regardless, he or she must choose one and reject one. The following is an example.

Choose the activity you prefer by placing a check in front of it. You may check only one in each pair. Be sure to respond to each pair.

1. _____ Walking through the woods on a cool, sunny day.

   _____ Attending a football game on a cool, sunny day.

2. _____ Lying on the beach talking to a member of the opposite sex.

   _____ Reading a good novel at home.

The advantages of the forced-choice approach are: (a) the respondent is forced to choose between alternatives—a behavior that is part of life, and (b) the choices usually are equated in their social desirability so that neither choice makes the respondent who chooses it appear healthier, more motivated, more moral, or more intelligent than the other choice.

One of the advantages of the forced-choice method, the difficulty for the respondent in skewing answers in a socially desirable way, also leads to a disadvantage. With forced-choice techniques, all respondents obtain the same scores, i.e., if you sum their scores on scales A, B, and C, they always will equal the same number. Thus, a person whose total involvement in the three areas measured is extremely high will obtain the same total score as a person whose total involvement in the three areas is near zero. A second disadvantage of this method is that it may irritate respondents by requiring them to choose between two equally attractive alternatives or, worse, between two equally unattractive alternatives, neither of which they would consider normally.

Three modifications can be made in the simple forced-choice approach. One is to present three or four alternatives and have the respondent select the one he or she likes *least*. This approach provides more flexibility than just two alternatives. A second method is to provide two alternatives, but instead of asking the respondent to choose one, allow him or her to allocate five points between the two choices. Thus, choice A may receive a rating of 5 and choice B a rating of 0, or choice A may receive a rating of 3 and choice B a rating of 2, and so on. This approach avoids forcing an artificial choice. All responses are not counted because some alternatives are used to indicate neither (or none) of the scales being studied. If a person responds to an alternative that is irrelevant to the characteristics being measured, it reduces his or her scores on the scale being assessed and lowers the total score.

Sociometric Ratings

One application of the rating-scale technique is the rating of members in a group (department, family, task group, etc.). Assuming that people are using a seven-point rating scale on the same dimension, ratings can be summed in two ways, as shown in the matrix that follows. One can see at a glance whether person A rated person B in a way that is discrepant from the way in which B was rated by the others in the group; whether person A rated person B in a way that is discrepant from A's usual rating scale; who received the highest total rating (Roger); who received the lowest (Elizabeth); who tended to rate everybody high (Marsha); and who rated everybody low (Roger). It also is possible to look at people’s self-ratings (diagonal cells) to see if some are unusually hard or easy on themselves.

Matrix of Ratings Given and Received

	Ratee					**Ratings Given to Others**	
	Bill	Roger	Marsha	Nancy	Elizabeth	Total	Mean
Bill	4	6	2	5	1	18	3.6
Roger	2	4	1	3	1	11	2.2
Marsha	6	7	5	5	4	27	5.4
Nancy	5	7	2	4	1	19	3.8
Elizabeth	3	5	1	2	1	12	2.4
Ratings Received from Others							
Total	20	29	11	19	8	87	17.4
Mean	4.0	5.8	2.2	3.8	1.6	17.4	3.5

In general, the sociometric technique is one that can be done on the spur of the moment, does not need any forms, allows open discussion in the group (this should be agreed on beforehand by group members), and provides a number of informative and discussable indices.

Meanings of Numbers

Scores derived from instruments do not have meaning except as they are related to both the content of the items and the method of deriving numbers. Interpreting an index of anxiety based on counting how many items a respondent marked as true is different from determining the meaning of a score in a forced-choice inventory. The basic number systems relevant to instrumentation are as follows:

Nominal Numbers are used to “name” things and are assigned arbitrarily. No “greater than” is implied. An example is the numbering of football jerseys.

Ordinal Numbers represent ranks. There is a notion of “greater than” but not “how much greater.” An example is the ranking of graduates in a class.

Interval Numbers denote by how many units one score is greater than another. However, there is no true zero point. An example is a Fahrenheit thermometer.

Ratio Numbers indicate positions on a scale with a true zero point, such as weight. Comparisons such as “twice as much” are possible. An example is a bathroom scale.

Most instruments used in HRD work are assumed to generate ordinal and interval data. When a group facilitator collects ordinal data, however, he or she is restricted in the types of statements that can be made about the data. For example, if participants are asked to rank-order the other members of the group on some characteristic, they must be reminded not to interpret such *rankings* (ordinal) as though they were *ratings* (interval).

DEVELOPING GROUP NORMS

In developing one's own instrument or in working with an instrument for which the norm groups are significantly different from the members of one's participant group, it is necessary to know some simple statistical procedures for developing norms. The simplest norm is a hand count of scores, with a posted tally. If participants seem reluctant to reveal their scores, they may be asked to write them anonymously on small pieces of paper. A chart such as the following then can be constructed.

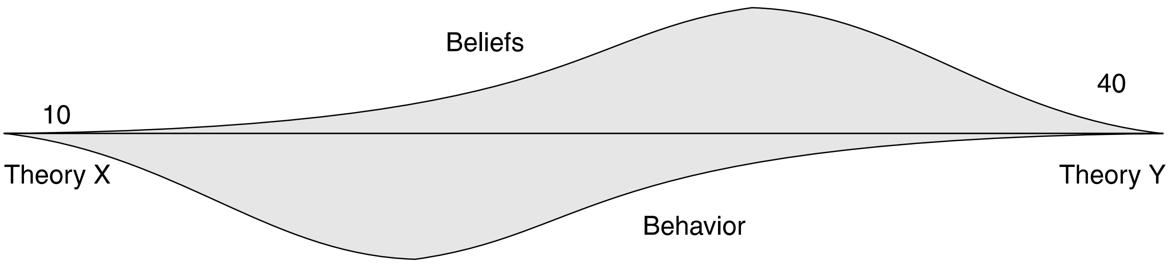
Name of Characteristic

(Rounded) Score	Frequency
7	______
6	______
5	______
4	______
3	______
2	______
1	______

Almost equally simple are averages (means, medians, and modes) and ranges that can be obtained from such a tally. For example:

	Dimensions	1	2	3	4
How I see myself in this group	Mean	______	______	______	______
	Range	______	______	______	______
How I see this group	Mean	______	______	______	______
	Range	______	______	______	______

The tally can be reduced graphically, making a group "photograph." Two scores can be depicted simultaneously on the same numerical scale. The following pattern, for example, often emerges on the instrument "Supervisory Attitudes: The X-Y Scale" *(Pfeiffer & Company Library 19*, 187-192):



X-Y Scale Norms

The tally can be converted into percentile ranks by a simple procedure.

1. Add two new columns to the tally: one for cumulative frequency (from the bottom up) and one for approximate percentile;
2. Divide the number of participants by one hundred; and
3. Multiply each entry in the frequency column by this factor.

The result is a set of percentiles that are interpreted this way: "___________ percent of the people scored this given score or *less.*" The procedure is illustrated here with hypothetical data.

Group Leader Self-Disclosure Scale Norm

Personal History

Score	Frequency	Cumulative Frequency	Approximate Percentile
7.0	0	61	99
6.5-6.9	1	61	99
6.0-6.4	1	60	98
5.5-5.9	2	59	97
5.0-5.4	6	57	93
4.5-4.9	8	51	84
4.0-4.4	10	43	71
3.5-3.9	9	33	54
3.0-3.4	9	24	39
2.5-2.9	7	15	25
2.0-2.4	5	8	13
1.5-1.9	2	3	5
1.0-1.4	1	1	2
0.5-0.9	0	0	
0.0-0.4	0	0	

If the group is small, a simple tally may be preferable. Norms for comparable groups can be accumulated over time to provide more stable statistics and to facilitate comparisons that are, on the average, more meaningful.

VALIDATING INSTRUMENTS THROUGH USE

One way to obtain clear behavioral evidence of instrument validity is for the facilitator to form relatively homogeneous groups whose members' behaviors (individually or as a whole) are likely to be consistent with the instrument scores. The procedure for this type of validation is relatively simple. First the instrument is administered, and then homogeneous groups are formed. For example, with an instrument that categorizes scores on separate scales (e.g., "self," "interaction," and "task" orientations), after administration and scoring but *prior* to any explanation, group members are divided into score categories, and subgroups are formed of members who scored high on the same scale.

The second aspect of the validation procedure requires a task provided by the facilitator. Ideally, the task should be highly suited to the precise types of behavior predicted by or inferred from the instrument; however, it is adequate if the task is designed to permit a range of various behaviors. In other words, if the facilitator were to give all subgroups the same task and allow only a very short period of time in which to accomplish it, all subgroups would exhibit task-oriented behavior. With a task that allowed time for a wider range of behavior, subgroups typically would exhibit the behaviors expected. Thus, subgroups composed of highly task-oriented individuals would focus on the task given them; subgroups composed of highly self-oriented individuals would accomplish little (often with discord, perhaps over leadership roles); and subgroups composed of highly interpersonal-oriented individuals would have an enjoyable time socializing and might or might not accomplish the task.

Although these dynamics usually are quite easy to identify, the use of observers or, even better, videotape to record brief vignettes of interaction in each subgroup can help to capture clearly and strongly the behaviors of interest. Subgroup members often are surprised to see what they remember as happening "to some degree" recorded so emphatically on videotape. Thus, these procedures can be quite powerful for demonstrating the reality of the behaviors that the instrument measures through what otherwise may seem to be rather abstract scales. Participants' belief in the validity of a measure is a necessary precondition for the use of the instrument in training.

It is important to remember that once the behaviors that the instrument measures are demonstrated, the focus must shift to the individual and his or her plans for altering or modifying some behavior. This is the "training" phase of instrument use. It goes beyond merely understanding what an instrument measures; it involves the application of learning to real life. As is discussed previously, the goal of using instruments in training designs is to develop self-awareness. The individual participant must be able to understand which of his or her specific behaviors might be changed. Although new behaviors may be obvious to some participants, for others it may be difficult to develop plans for behavioral change. The facilitator must provide a training structure that allows participants to develop and then in some way to practice alternative behaviors. A wide range of actions and options for doing this exists. Please refer to the discussion entitled “Instruments As Part of Overall Training Designs” for more on this topic.

GENERATING CONTENT WITHIN THE GROUP ITSELF

One of the most effective and unique aspects of instrument use is generating instruments on the spot with the participants when the facilitator identifies a need that cannot be met by any of the instruments he or she may have available. This strategy also is useful with participants who are apt to engage in the defensive tactic of “nitpicking,” challenging the validity of items and/or scales.

Items that can be used to form an instrument can be developed quickly with the assistance of participants. This approach has the advantage of developing a sense of ownership, commitment, and relevance. If the recipients of instrumented feedback are involved in specifying what is to be measured, they are more likely to learn from the process and less likely to reject the data.

Spontaneously derived instruments are helpful in exploring concepts, generating here-and-now data for discussion and/or processing, developing new models for exploring human interaction, and facilitating a sense of ownership about learning. It is possible for a facilitator to develop instruments that measure exactly the characteristics in which he or she is interested, and facilitators should feel free to use their creativity to get at those characteristics in whatever way seems most appropriate.

Of course, few scales evolved in this way would stand up to statistical scrutiny. Homemade instruments tend to be unreliable, yielding unstable scores. The caveat is that one must beware of over-interpreting the results. The results should be viewed by the participants as suggestive, not conclusive.

One can create an instrument that can generate intensive discussion by having the participants brainstorm a list of the qualities or characteristics relevant to what is being studied (e.g, in a leadership workshop, they would list the qualities of a good leader). Then each participant would rank the items on the list from first to last in order of importance. A large group may be broken into small groups for discussion. Participants would compare their rankings, obtain a group average, discuss the differences and similarities in their biases, and attempt to reach consensus. A crude measure of influence can be obtained by summing the differences between the group-consensus ranking and the original points of view of individuals. Presumably, those with low scores would be the most influential.

Adjective Checklists

Adjective checklists can be developed by some procedure such as brainstorming, free association, or critical incidents. For example, the facilitator can ask that each participant write down two or three adjectives that describe his or her best friend and two or three that describe the one person with whom he or she has had the *least* satisfactory relationship. The adjectives generated then are alphabetized and used for various purposes such as feedback, intergroup perception checking, and evaluating.

An example can be drawn from a team-building session. The facilitator wants a way to focus an exchange of interpersonal perceptions about influence within the work group. Each member is instructed (a) to recall an individual in his or her past (not present) who has influenced the member both significantly and *positively* and (b) to write down two or three adjectives that describe how he or she experienced that person. Then the process is repeated with each participant listing a person who influenced him or her both significantly and *negatively*. These two lists of adjectives are called out by the participants, alphabetized, and posted. This "instrument" then is used to guide the discussion of individuals in the team. There are no questions about reliability, validity, objectivity, or relevance.

For an instrument such as an adjective checklist, it is not necessary to go through a mechanized duplicating process. Such a list can be written on a chalkboard or on newsprint for the participants to see, or they can make their own copies on blank sheets of paper. Instructions can be given orally.

Attitude Scales

Attitude scales can be derived from declarative sentences about issues facing members of the group or the organization. The facilitator can select an appropriate topic (attitude object) such as job enrichment, MBO, participative management, or consensus as a decision-making strategy. Each participant writes a statement about the topic. These statements are posted and numbered, without discussion. (The facilitator must be careful not to edit the statements except to make them unidimensional. For example, the statement "Job enrichment is best done by imposing changes designed by management because workers do not have the proper perspective" is really two statements and should be posted as such.) The response scale is then introduced. Usually this is some variation of the familiar Likert scale, such as:

SA = Strongly Agree
A = Agree
U = Uncertain or Undecided
D = Disagree
SD = Strongly Disagree

Participants write down the number of each statement and their response to it. These responses are tallied by a show of hands, and the facilitator guides a discussion of the results.

In one supervisory skills session, the facilitator noticed a controversy among participants regarding the subject of women as managers. Because the group had resisted an earlier paper-and-pencil inventory that the facilitator had introduced, a spontaneously developed instrument seemed to be the right way to focus attention on both the subject and self-assessment. Participants were instructed to write down a completion of the following sentence: "As managers, women . . . ." Members of the training group were reluctant to read their sentences aloud in order for the facilitator to post them, so small discussion groups were formed to explore the reasons why people were hesitant to reveal their attitudes. Reports indicated that this was a controversial topic in the company and that it never was addressed publicly. The facilitator then had participants write their sentences on cards and pass them in anonymously. The statements were randomized and posted. Members then used the five-point Likert scale to record their agreement or disagreement with each item. Individuals asked for a group tally on any item of interest, and a lively discussion followed.

Semantic Differential Scales

Participants can develop a list of bipolar adjectives, such as hot-cold, light-heavy, high-low, etc. These are posted as the ends of continua, with six or seven points in between, e.g.:

empty/________/________/________/________/________/________/________full

fast/________/________/________/________/________/________/________slow

Then a topic to be rated is announced. Each participants rates the topic according to his or her associations with it. These ratings are tallied and written on the posted instrument as the agenda for discussion. The group sees its profile of association for the topic rated. According to the research of Osgood, Suci, and Tannenbaum (1957), who developed this type of instrument, three types of bipolar adjectives should be included: those that are evaluative (sweet-sour, happy-sad, etc.), those that measure potency (large-small, strong-weak, etc.), and those that refer to activity (active-passive, fast-slow, etc.).

Several years ago, public-school principals were discussing racial integration in their schools in a management-development seminar on conflict. The interchange was energetic, but the facilitator sensed that the participants were using familiar terms in ways that implied a variance of meaning. What "the neighborhood concept" meant to one person was not what it meant to another. The discussion was not creating understanding, and emotions were running high. A set of semantic differential scales was constructed to clear up the connotative meanings of various, emotionally loaded subtopics, such as forced busing, magnet schools, and ethnic studies. This intervention helped to foster more effective self-expression and listening.

Behaviorally Anchored Rating Scales (BARS)

Participants individually write down important dimensions to be studied. These lists then are melded on newsprint, and the group members select the most salient ones by some appropriate method (e.g., voting for three, ranking, or rating). Subgroups are formed to correspond to the selected dimensions, and these groups construct behaviorally anchored rating scales (BARS). (For a structured experience and a professional development article on using BARS, see *Pfeiffer & Company Library* 18, 113-119 and 20, 167-178). These scales are presented to and edited by the group. Then the BARS are used to perform ratings to generate data for discussion.

An example of this technique comes from team building. In sensing interviews, members of the group had voiced dissatisfaction with their weekly staff meetings. In the beginning of the offsite session, the group constructed BARS on six different aspects of effective staff meetings (e.g., cooperation, open communication, attention to process). One of the scales was participation.

Participation 1 • • • 2 • • • 3 • • •4 • • •5 • • • 6 • • • 7 • • • 8 • • • 9

The meeting is dominated by one or two members	People contribute when they have strong feelings	Everyone's input is solicited before decisions are made

These scales were then used by individuals to rate their usual meetings and that particular meeting. A progress chart was constructed on which averages were posted. The BARS were used at the end of each three-hour segment of the team-building session and became the bases for team self-assessment at ensuing regular staff meetings.

Model-Based Questionnaires

The group is given the task of creating a model of the process to be studied. Subgroups are formed around the major aspects of the model and they are instructed to develop one or two questionnaire items to measure the selected dimensions of the model. Each member of a subgroup makes a copy of the subgroup's item(s). As soon as all subgroups have completed this task, an "assessment circus" is held, in which everyone fans out from subgroups and "administers" items to one another. One participant approaches another, shows the item(s), records the response(s) on a *separate* sheet of paper, and then solicits a critique of the item(s). When all participants have been surveyed, the subgroups are reassembled to tally the data and to discuss the critiques of their items. Representatives are chosen to make brief reports to the total group. Then the facilitator leads a discussion of the aggregate data.

In a train-the-trainer workshop, the staff wanted to involve participants in exploring the norms of the "learning community" in order to increase commitment to shared responsibility. The group brainstormed community norms and selected a set of five that seemed particularly relevant to the workshop at that time. Members volunteered to work on a norm area of interest, and subgroups each wrote one or two items. After all the data were collected and collated, subgroup representatives had a confrontation meeting with the workshop staff to present the findings. (The other participants were observers.) The staff members had their own meeting at that time to explore methods to deal with dysfunctional norms and to reinforce facilitative ones. Subgroups reassembled to rewrite their items, based on the critiques; these items formed the basis for a later diagnosis of community functioning.

Conclusion

Spontaneously developed instruments can be quite useful in training programs in which the validity and reliability of an instrument are not as important as its application to the concerns of the group. For example, a trait can be defined, and then participants can brainstorm behavioral instances of high and low values of the characteristic. Traits can be selected from an adjective checklist to form the basis of a rating scale. Group-process phenomena can be identified and can become the content of reaction instruments; such dimensions of group development as trust and openness can be further differentiated into behavioral referents. Developing instruments spontaneously not only serves as an effective training and consulting intervention, but it also serves as the basis for scales that can be refined later into publishable form.

Although the basic precepts presented here hold true in the process of developing any instrument, our focus primarily has been on instruments for training and development groups. Instrumentation also is a very useful tool in organizational diagnosis. Because it is likely that an instrument may not already exist that would meet the specific needs of (or accurately describe) a particular organization, the discussion that follows presents guidelines for designing and conducting organizational surveys.

DESIGNING AND CONDUCTING ORGANIZATIONAL SURVEYS

There are many standardized instruments available for use in the applied behavioral sciences, including HRD work. Perhaps the most commonly used instrument is the familiar attitude and opinion survey. Such instruments are used in group and organizational settings to measure behavioral dynamics, morale, organizational climate, leadership, and a host of other variables that describe or relate to human behavior. Some are focused on a single variable while others are comprehensive instruments for use in organizational assessment and development. The *Survey of Organizations* (SOO) (1980), for example, often is used in conjunction with survey-guided development (Bowers & Franklin, 1977; Franklin, Wissler, & Spencer, 1977; Hauser, Pecorella, & Wissler, 1977). When the issues of concern to a manager or group facilitator are narrowly, rather than broadly, defined, a full-scale organizational assessment is not needed. What is then desirable is a survey instrument designed specifically to meet the user's needs.

Everyone is familiar with the results of some survey, but how those results are obtained usually is a mystery known only to the experts—whoever they are. Actually, the development of a survey instrument primarily is a matter of care, common sense, and skill developed through practice (and based on the ability to write coherently). No set of instructions can provide basic aptitudes or the equivalent of practice, but the following guidelines describe how to take care, what common sense is, and where to invest efforts that amount to practice.

STEP 1: DEFINE THE OBJECTIVES

What is the purpose or intent of using the survey? What precisely is the survey trying to find out? Why are these data needed? Can the data be obtained in some way other than a survey? If not, the objectives should be defined as precisely as possible—in writing—and should be limited to those that are really important, i.e., the reasons for doing the survey. If there are more than four or five basic issues, the survey probably will be too long and the respondents—the people who are asked to fill out the questionnaire—will not respond.

STEP 2: IDENTIFY THE POPULATION TO BE STUDIED

The population is everyone from whom the surveyor would need to have a response in order to completely and correctly answer the basic questions. It is important to be precise in defining the population to be used. Although this does not mean that a list should be made of everyone in the specific population, this is the time to consider how the survey instrument physically will be brought to the people in the respondent population. For example, will the members of the group be assembled in one place or will the questionnaire be mailed to them at their home or business addresses?

STEP 3: SELECT THE SURVEY SAMPLE

Ideally, one would like to conduct a census, a survey that includes everyone in the population of interest. Obviously this is not realistic if we are talking about, for example, “all managers.” The researcher must settle for a sample of the population, preferably a sample that will provide the same results as if all managers actually had responded. Less obviously, the consultant helping a particular company may also be faced with an unrealistic task if, for example, there are 783 managers in the entire company. With limited resources, an individual consultant may find it impossible to obtain and analyze this much data. Typically, one conducts a census-type survey only when the population is relatively small or when the need for total participation is extremely great.

Sampling techniques have been developed and refined extensively in the last several decades, but the two most significant ones are randomness and stratification. Normally, every person in the population should have an equal chance of being picked to be in the sample. This can be accomplished by random selection. Because any number of factors easily can interfere with random selection, this step requires extreme care. Suppose, for example, that the consultant to a small company decided to sample 20 percent of the 783 managers by listing all of them by their social security numbers and selecting 156 of them. If the consultant were to pick the first 156 and the list were in numerical order (low to high), the sample would not be random but biased. Because Social Security numbers are not assigned randomly, managers from certain parts of the country or of certain ages would be excluded from the sample. If the consultant were to take every fifth number, the sample would be random; but the best method would be to *list the numbers randomly* and then select every fifth one. This procedure still might not yield a representative sample because there are different numbers of managers at each level of management and fewer high up in the hierarchy. Thus, high-level managers would be less likely to be represented in the sample than low-level managers.

To correct this problem, one must stratify the sample—in this case by managerial level. This can be done by grouping the social security numbers by management level and picking 20 percent of the people in each level. The final sample then would be random and also would represent the population accurately. One can, of course, stratify the population on any basis that logically will reduce bias or make the sample more representative.

STEP 4: CONSTRUCT THE INSTRUMENT

In order to develop a concise instrument, one must have some skill in writing and also must be able to endure the tedium of rewriting over and over again until the questions (or items) are as nearly perfect as possible. A typical organizational survey has at least four basic parts: the cover letter, the items, the scales, and the codes. Each of these must be prepared extremely carefully in order to achieve optimum results. Obviously, much practice is a primary way to attain skill in this area.

The Cover Letter

The cover letter should be written clearly and simply, without the use of jargon and technical words. It should speak to the respondent on at least three issues: (a) why the survey is being conducted; (b) what the benefit of the survey might be, especially with respect to the respondent; and (c) the guaranteed anonymity and security of responses. Respondents also should be thanked for their participation. The cover letter should not be long; two to three paragraphs usually is adequate. The more the letter looks like a "real" letter, the better. The use of letterhead stationery is highly desirable, and, whenever possible, each letter should be signed individually. (This usually is possible only when fewer than one hundred questionnaires are being used.) The more personalized attention the respondent perceives, the higher the response rate will be.

The Items

The items are, of course, the heart of the survey instrument, but writing the questionnaire items surely is the most tedious aspect of survey design. It also is the most important. One begins with the objectives defined in step 1 and attempts to translate them into specific questions. Often, one can receive much guidance from the efforts of others, because the topic to be examined probably has been studied before. One often can borrow items appropriate to fairly specific needs from well-developed research questionnaires and use them with appropriate modifications. In many cases, one must obtain written permission from the author of the source instrument. Obviously, one must be careful to maintain professional ethics in the use of questionnaires and questionnaire items authored by others.

There are a few rules about writing questionnaire items. First, each item must ask only one question and must be unambiguous and specific. It is easy to write double- and triple-barreled items such as "To what extent does your supervisor give subordinates responsible or interesting work?" There are two separate items to be considered here: "responsible" work and "interesting" work. Another example is "To what extent are group members participating and involved?" In this example, "participating" would refer to the amount of talking, and "involved" would be an interpretation of participation. It is extremely important to mention only one dimension per item (or deal with only one topic per question).

Another of the most frequent errors in constructing items is to mix the scale (the rating or frequency) with the item, e.g., “Are participants highly involved?” rather than “How involved are participants?” (Hanson, 1981). Specificity leads to clarity, and clarity of the item is extremely important in order that all respondents can make choices from the same frame of reference.

Thirdly, questions should be worded so as to avoid social-desirability bias. That is, most people would agree or strongly agree with the statement “I support the values on which our society is based” and would disagree with the statement “Everyone should try to get as much as he or she can when selling a commodity.” The socially desirable responses to these and questions such as “To what degree do you do a good job at work?” are fairly obvious; there is little point in asking them. If the topic is important, items must be developed so that the “right” response is not obvious. For example, the “get as much as you can” item could be stated in this way: “To what extent do you agree that profits should be maximized?” Another question might be: “Compared to others doing similar work, rate the quality of your own output.” A scale of “better than anyone else,” “better than most,” “as good as most,” “not as good as most,” and “not as good as anyone else” might be used. The bias of social desirability also can be avoided by avoiding “loaded” words. When objective terminology is used, people usually will respond quite honestly, even when the response is not so favorable to them personally.

The fourth rule is: avoid threatening the respondent. If one were trying to measure the feelings of auto workers about job security, an agree-disagree item such as “less productive workers should be laid off first” probably would threaten many people. A person who is threatened frequently will refuse to complete the survey. The threat in an item may not always be obvious; threat often is a matter of circumstance, such as the financial stability of an organization or the economy in general. When an implicit threat is inevitable, reassurance of anonymity often helps.

Good item construction depends on common-sense writing skills, i.e., avoid leading questions; try to phrase items objectively; use common rather than obscure terms; and strive for brevity and clarity. Again, only practice can provide the skills needed to write effective questionnaire items.

The Scales

Scaling need not be overly technical. As has been discussed before, the most commonly used scale is the Likert scale, with five or seven multiple-choice alternatives such as “to a very great extent . . . to a moderate extent.” Other dimensions that commonly are used include “agree-disagree,” “how much,” “how often” (frequently-infrequently, never-always, once a day-once a year), “to what degree,” and “how important.” These simple scales generally make certain assumptions that render them equivalent to much more technically sophisticated scales.

There is no one best set of scale labels. “To what extent” and “agree-disagree” probably are the most used. When a questionnaire has more than twenty items, it generally is less boring for the respondent if more than one scale is used. When actual frequency of behavior is being measured, the "how often" or "never-always" sets are most relevant. When personal values or the rewards one wants from work are being examined, the "how important" scale might make the most sense. Frequently, it is possible to phrase one's questions so as to use whichever scale labels one prefers.

It frequently is helpful to word items so that the answers can be graded on a continuum rather than discretely. For example, a scale that measures *degrees* of managerial control (high control, some control, little control, no control) results in a continuum; whereas on an instrument that identifies *sources* of managerial control (fear, threats, punishment, etc.), the items must be graded discretely, i.e., individually. A continuum generally is indicated by the use of adjectives or adverbs (high-low, often-not often, moderate-very much), and discrete items generally are nouns or verbs (reward, punishment, does, does not). It is important not to use discrete items as if they were a continuum (Hanson, 1981).

Research shows that the "right" number of points on a scale usually is between five and nine. This is the comfortable range of discrimination for most people. Therefore, it usually is safest to use a seven-point scale, although a five-point scale does make the handling of data easier. It is important to avoid restricting the range of responses by using only two or three categories. The rest of this would be meaningless data obtained at considerable cost.

The Codes

Finally, we come to the matter of coding responses. To prepare to analyze the data, one must construct a code book. This usually is a copy of the questionnaire, marked up to indicate how each item is to be scored and how to deal with problem responses. For example, a response marked between 3 and 4 would be treated according to a general rule, explicitly stated in the code book. Other typical concerns are failure to respond to an item and reversed items (those for which low scores are "very good" and high scores are "bad"). The latter may be coded in reverse so as to be consistent when the data are tabulated. One must think carefully about how to score the returned questionnaires.

STEP 5: PRETEST THE INSTRUMENT

No matter how well developed the survey instrument is, there still will be at least minor problems that must be identified and corrected. This is the function of the pretest. A small number of instruments are prepared as mockups; these may be typed instead of printed. Volunteers are recruited to respond to the items on the questionnaire as though they were members of the sample population. In fact, some of these pretesters should be members of the population from which respondents are to be drawn. Immediately after completing the instrument, each of these volunteers is interviewed, in order to identify flaws or errors on the form. Even if the sample of volunteers is quite small, a pretest is crucial. It is almost certain that some (although, unfortunately, not always all) errors will be identified.

STEP 6: PREPARE THE FINAL DRAFT

After the errors identified through pretesting are corrected; the problems are resolved; the typographical errors are corrected; and an attractive, clean, final copy is prepared for reproduction; then the instrument form must be checked carefully. From here on, errors probably will be too costly to remedy except by discarding some data, which also is a costly cure.

STEP 7: ADMINISTER THE INSTRUMENT

Ideally, one would administer the instrument to everyone in the sample at the same time, perhaps in one large group meeting. Usually this is impossible. Even the next-best situation, having several group-administration sessions, is not always feasible. It may be necessary to distribute questionnaires to individual respondents who complete and return them either by hand, intra-company mail, or national mail. Although this can increase privacy and anonymity for respondents, it also usually leads to decreased return rates. Therefore, when such a voluntary return procedure is unavoidable, the surveyor must do everything possible to boost the return rate. There are at least four ways in which this can be done.

First, *the importance of the study* must be emphasized (its intrinsic relevance, its potential usefulness—to management, for example—and the possible benefits to the respondent). These factors should be emphasized in three ways: (a) in the cover letter; (b) in informal conversation when the questionnaire is being handed to the potential respondent; and (c) by explaining that some authority (e.g., management) supports the survey (as shown by management's willingness to have it administered during work time). Telling people that they should not fill out a questionnaire during work time is, by the way, the same as saying that management thinks it is worthless and will ignore it. Most potential respondents then will ignore it too.

The second aspect to be emphasized is *the confidentiality and privacy of individual responses*. Again, this should be stated in the cover letter and verbally. It often is wise to explain that the person affected (e.g., management) will see only aggregate data such as averages and percentages, not any person's or work group's data. If possible, questionnaires to be returned by mail should be sent to an address other than the organization's or interested party's. If people return questionnaires at the work place, a sealed box should be provided so that respondents can drop their completed questionnaires into it. It should be specified verbally and on the questionnaire form that no names or identifying marks are to be put on the questionnaires or response forms.

The third aspect of administering the instrument is to invest as much time and effort as possible in *personal contact with the potential respondents*, explaining the objectives of the survey verbally and asking for and answering questions. Respondents should be promised a summary of the results. Survey forms should be given directly to potential respondents, and they should be thanked for their participation. In short, everything possible should be done to maximize personal contact. Respondents should be asked when their responses can be expected, and an attempt should be made to obtain verbal commitments to a specific time frame. These investments of time and energy will be repaid amply in terms of return rate.

Finally, the *questionnaire instructions must be as clear as they possibly can be.* The respondents should be told whether they are to make choices, indicate rankings, circle answers, etc., and how this is to be done. They should be told if they are to turn the page. They should be reminded not to skip or omit any items. They should be told how to respond each time a new type of item, scale, or topic is used.

A deadline should be set for receipt of the responses, and this date should be included on the questionnaire and mentioned verbally to the respondents. The date given should be at least one week earlier than the actual deadline. Shortly before the stated deadline, all respondents should be reminded of the date by means of a letter or memorandum. An offer can be made to provide another copy of the questionnaire if the first was misplaced. Normally, one would not attempt more than two such reminders.

The actions and techniques described above are time consuming, but the more carefully they are attended to, the better the response rate will be.

STEP 8: CODE THE RESPONSES

This is tedious work. Accounting ledgers make good tally sheets for raw data, and there also are more costly forms designed for this purpose. Special forms are particularly useful when the data are to be transferred to computer disk or tape for computer analysis (which certainly is the easiest way to analyze data). As dull as it is, this step is very important because minor errors can have serious impact. A few of the survey questionnaires always should be checked at random to see if there are errors in coding.

STEP 9: TABULATE THE RESULTS

The aim here is to present the data so that people can understand and make interpretations from the information generated. Sample tabulations of responses for each item, using percentages (not just raw numbers), generally will suffice. This can be indicated on a “doctored” copy of the questionnaire, with percentages filled in where the check marks would go. Many results may be ignored later, but it is important to begin by tabulating everything. An example of a tabulation of respondents’ job satisfaction follows.

Sample Tabulation of Job Satisfaction

Job Satisfaction % (N)					
Completely	Very	Mostly	Slightly	Not at all	Total
61(82)	18(25)	10(13)	7(10)	4(5)	100(135)

The next step is cross-tabulation for items that have some important relation to one another. For example, to determine whether older workers are less satisfied than younger workers, one would cross-tabulate age by satisfaction, as is shown in the second example.

Sample Cross-Tabulation of Worker Satisfaction

Age	Worker Satisfaction % (N)					
	Completely	Very	Mostly	Slightly	Not at all	Total
Under 25	75(15)	20(4)	5(1)	0(0)	0(0)	100 (20)
25-30	70(24)	15(5)	10(4)	5(2)	0(0)	100 (35)
31-40	60(24)	15(6)	10(4)	10(4)	5(2)	100 (40)
41-50	50(15)	25(8)	10(3)	10(3)	5(1)	100 (30)
Over 50	40(4)	20(2)	10(1)	10(1)	20(2)	100 (10)
Overall	61(82)	18(25)	10(13)	7(10)	4(5)	100 (135)

The items in parentheses are the numbers of respondents of given ages who gave specific satisfaction-level responses. For example, of the forty people between thirty-one and forty years of age, 60 percent (or twenty-four) were completely satisfied. Of the ten people over fifty in this sample, 40 percent (or four people) were completely satisfied. One may not know if a trend such as this is statistically significant, but might decide later that it is worth testing.

Obviously, there must be a reason for setting up cross-tabulations. If the data are to be analyzed by computer, it is quick and inexpensive to cross-tabulate everything by everything, but then one must wade through mounds of printout. So even when the calculations are easy, it is worth spending some time to decide what, if any, variables should be cross-tabulated.

Tabulating results is another important step (and can be tedious unless one uses a computer). It leads directly to data interpretation, the final output of all survey work.

STEP 10: PREPARE THE REPORT

Before preparing a final report, one must pull together all thoughts about the survey in a brief overview or summary paper. The aim is to organize these ideas and the data, not to communicate results. Based on this summary paper, the needs of the organization, and the circumstances of the survey consultation, one then can proceed with a formal, final report.

The summary paper should begin with about a page of description, highlighting what the data show and referring to the tables. It is a good idea to review the data and the tables several times and then to review them once more, this time looking for important omissions. Then a second, more detailed, summary is written. This second summary might be just a minor revision of the first, but by approaching the task this way one allows as much of a chance as possible to pull it all together without missing important findings or interpretations.

The final report will be based on the second summary, but should be tailored to the circumstances. Some consultants believe that this means that if management is quite uninterested in doing anything with the results of the survey and would be threatened by their negative tone, it is wise to prepare a brief, bland report that plays down the negative aspects. Our opinion is that there should be a clear contract at the beginning, with both consultant and client agreeing on the purpose and use of the survey. Thus, management agrees at the beginning that it wants the data. If this does not seem to be the case, it would be a sham (and probably a mistake) to conduct the survey in the first place.

The exact form of the final report will depend on how it will be used and on other circumstances. If the data will be used to work on problems, with small groups involved at all levels, the report should avoid inferences and conclusions; should contain data grouped by unit, department, or division; and should be clear to a nonexpert. If top managers will work on the data to derive action plans, then more summary, charts, and recommendations usually are desirable. The guidelines for preparing the final report are to consider who will use it and to consider the purpose for which it will be used. The surveyor should prepare a report that (a) will not, in itself, do harm to the people or organization studied, (b) is targeted to the users, and (c) is an appropriately usable form.

CONCLUSION

This has, of necessity, been an abbreviated discussion of the ten steps in planning and conducting paper-and-pencil surveys of groups and organizations. To cover all the practical details involved in this endeavor would take a longer book. Moreover, the skills needed to effectively design and conduct organizational surveys cannot be learned by reading but can be developed only by practice, correction, and more practice.

The ten-step model is as follows:

Step 1: Define the objectives.
Step 2: Identify the population to be studied.
Step 3: Select the survey sample.
Step 4: Construct the instrument.
Step 5: Pretest the instrument.
Step 6: Prepare the final draft.
Step 7: Administer the instrument.
Step 8: Code the responses.
Step 9: Tabulate the results.
Step 10: Prepare the report.

RESEARCH USES OF INSTRUMENTATION IN HUMAN RESOURCE DEVELOPMENT

The parts of this volume have progressed from the general to the specific, from considerations germane to the use of all instruments to specific applications of the technology. Even more specific than the use of instrumentation in organizations is the use of instrumentation for research purposes.

The use of instruments in research appears to be deceptively easy, i.e., all one has to do is have people fill out an instrument and see if their scores relate to some group or individual criterion. The deceptiveness lies in the number of problems that can arise in using instruments in research. However, one can anticipate many of the factors that can cause failure in finding significant differences or that make results open to several different explanations. The major problem areas are reliability, validity, pretesting and posttesting, transparency and social desirability, management, and human concerns.

THE RELIABILITY PROBLEM

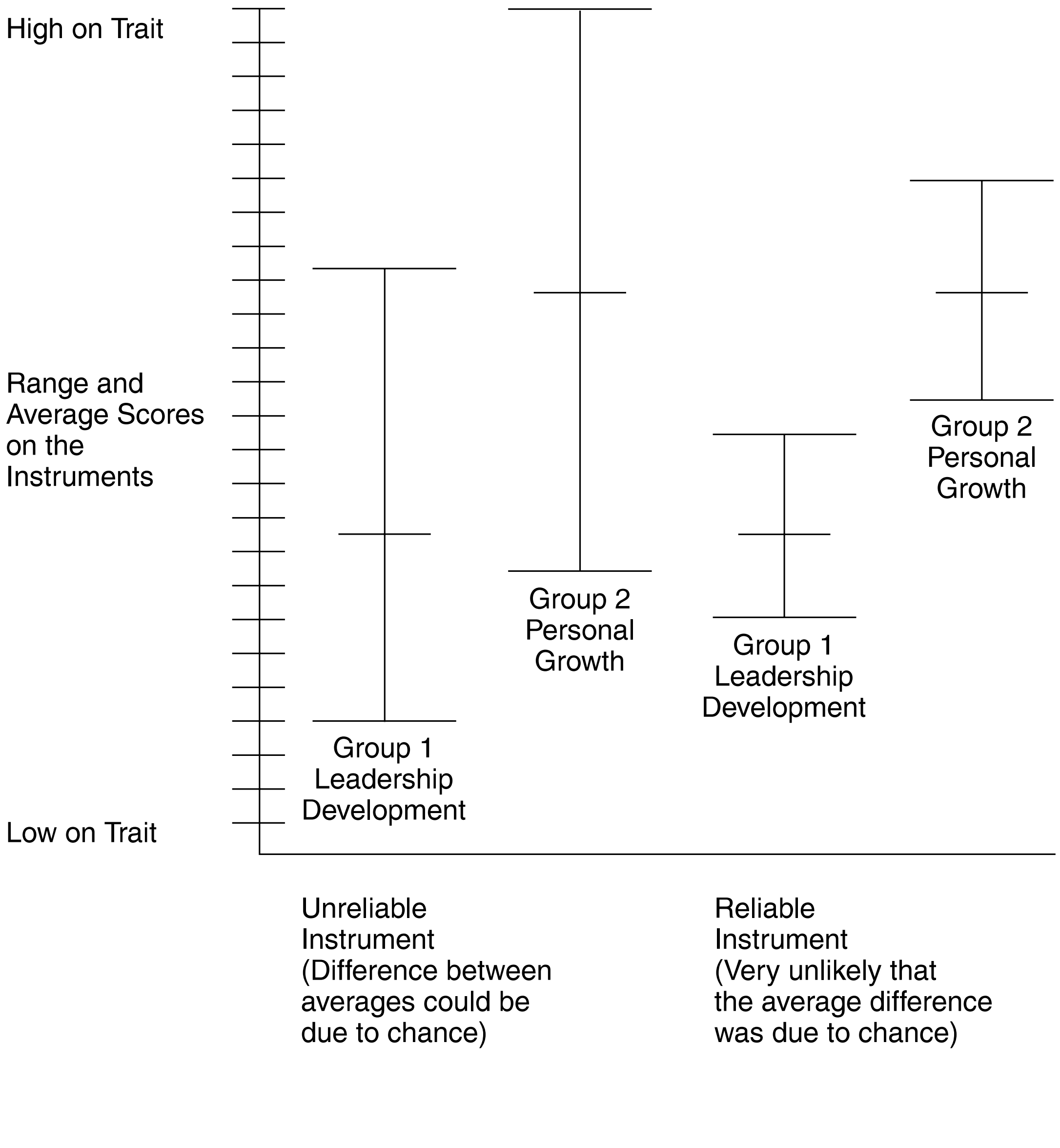
The term “reliability” can strike terror in the hearts of users of instruments. Many people are not sure what the word means and, therefore, are not sure whether or not their instrument has a sufficient amount of it.

Test reliability means that the instrument is measuring something other than chance responses, guesses, or some characteristic that comes and goes. One way in which this is demonstrated is by showing that people respond in the same way to an instrument the second time as they did the first time (usually after an interval of about two weeks). This is called test-retest reliability. A second way of demonstrating reliability is to examine the extent to which responses to two halves of an instrument or two different forms of an instrument are correlated, i.e., measuring the same thing. High correlations indicate that all the individual questions relate to the same general attitude and that some actual characteristics are being measured.

The importance of reliability is that it determines the error of measurement. The higher the error of measurement, the less sure one can be that a respondent’s score is close to the score that the person would obtain if he or she took a parallel form of the instrument or took the same instrument again after a period of time.

How much reliability is enough? This question is more important for a trainer than it is for a researcher, because a trainer may be trying to draw conclusions and pose alternatives based on scores. If an instrument is unreliable and the trainer does not know it, it is possible that he or she may offer an erroneous alternative. A researcher, on the other hand, has a built-in check against conclusions based on unreliable instruments:

statistics. Statistics allow comparison of the amount of change in a person's responses with the amount of random change in scores or comparison of the difference between two or more groups of people with the variation in scores within the group. Unreliable instruments show an excess of random variation in scores. The greater the error in the instrument, the less likely it is that the differences between average group scores will appear statistically significant (so large a difference between groups that it is unlikely to have occurred by chance). If one is using an unreliable instrument, one is not likely to find a difference between groups or conditions that is greater than the large error of measurement, and one is not likely to uncover significant findings from using the instrument. The figure that follows illustrates this point.


High on Trait
Range and Average Scores on the Instruments
Low on Trait
Group 1 Leadership Development
Group 2 Personal Growth
Group 1 Leadership Development
Group 2 Personal Growth
Unreliable Instrument (Difference between averages could be due to chance)
Reliable Instrument (Very unlikely that the average difference was due to chance)

Nonsignificant results do not permit one to say that "there is a difference between the groups" or "there was no change." One can only say that one is not able to demonstrate that there was a difference. An unreliable instrument is not likely to lead to trouble, *per se,* but the user will be frustrated by his or her inability to draw any conclusions about what has been studied.

Unreliable instruments lead to conservative, nonrisk errors. This is not to say that such conservative errors are not something to be concerned about. Science proceeds by the gradual accumulation of confidence in or disenchantment with a theory, procedure, or test. If people test a valid theory or procedure with an unreliable instrument and find no support for it, they have done a disservice to the theory, the theorist, and the quest for "truth." They also have wasted their own time and effort.

THE VALIDITY PROBLEM

The validity of an instrument is the degree to which it measures what the publisher and author say it measures. An instrument can be reliable yet invalid; that is, it may be measuring an undetermined entity consistently. On the other hand, a very unreliable instrument cannot be valid because it is not measuring much of anything but random responses and guesses.

One way in which validity usually is demonstrated is by showing that an instrument correlates positively with another instrument that is measuring the same general characteristic (convergent validity). A high positive correlation (+.50 or better) means that high scores on instrument A were obtained by people who received high scores on instrument B, and low scores on instrument A were obtained by people who received low scores on instrument B.

Another method used to determine the validity of an instrument is to show whether people who should score in a predicted way on a certain characteristic do. For example, William F. Buckley, Jr., should score nearer the conservative end of a liberal-conservative scale than should Senator Edward Kennedy.

There are other ways to validate instruments (see the discussions of validity and reliability under "Technical Considerations" earlier in this volume), but the two approaches above should provide an indication of how it is done.

Even if an instrument has substantial reliability and validity coefficients as reported in its manual or in published research, one cannot proceed with the assurance that it is a "good" instrument without first examining it very carefully (cutting through the labels, psychological jargon, and general descriptions) to discover exactly what the instrument measures. This can be done as follows:

1. Examine the questions in the instrument and ask, "Do these questions really reflect what the author says the instrument is measuring?" If not, what do they measure? If so, do they measure other characteristics as well?
2. Examine the correlations with other instruments and ask, "What are all the factors that could cause a person who scores high on this instrument to score high on the other instrument?"

3. Look at the differences between norm groups (e.g. men versus women, college students versus construction workers) and ask, “In how many other ways do these groups differ?” Could this instrument be measuring one (or more) of these factors instead of, or in addition to, what it purports to measure?
4. Complete the instrument and think about your own experience with it before deciding whether it is a valid measure of what it claims to be.

The most serious problem with validity is the use of good instruments to measure what they have not claimed to measure. For example: A researcher wants to investigate the effect of loving attention from members of the opposite sex on individuals’ perceptions of their own sexual attractiveness. He cannot locate a measure of perceived sexual attractiveness, so he uses a previously validated measure of self-esteem. Although the instrument has some questions relevant to sexual attractiveness, most of them measure other aspects of self-esteem. He finds that his experimental situation had no significant effect on the scores he obtained. He did not get results because the instrument was not valid as a measure of perceived sexual attractiveness. In this case, it probably would have been preferable for the researcher to develop his own instrument with questions that related exactly to what he was expecting to measure. If he had found significant results, he might have had some confidence that he had tested the characteristic in which he was interested. Even if he had obtained significant results with the self-esteem scale, he would have been faced with the fact that he had not tested his predictions in a valid way.

Even if an instrument is reliable and valid, if it does not measure the trait that is being researched, it is almost useless. A compromise would be to use two instruments that come as close as possible to measuring what is desired and then adding one’s own instrument to the research. It is a good idea to remember that the majority of theory-testing articles published in social psychology use only one—or, at the most, three—questions to measure the effect obtained in the experiment. The results are obtained from strong, experimental manipulations, and the few questions are aimed exactly at the effect that is expected.

PRETEST AND POSTTEST PROBLEMS

One of the most useful ways to do research from both a common sense and a statistical viewpoint is to use a pretest and posttest design. The logical way to measure change is to evaluate a group of people on a characteristic prior to an experience and then to evaluate them again on that characteristic after the experience that was supposed to change them. If the scores are the same or close enough that error of measurement could account for the difference, then the people did not change. If the group average shifted significantly in the expected direction between the first evaluation and the second, then they did change. Using a pretest-posttest design makes more intuitive sense than using a “posttest-only” approach in which one does not measure the experimental group initially but compares their scores at the end of an experience with the scores of a group that did not go through the experience. No matter what the outcomes are, it is often hard to convince others, or even oneself, that the two groups were the same except for the experience.

The statistical advantage of the pretest-posttest approach is that it allows the researcher to use a form of analysis that ignores differences between people and measures only how much change occurs within people. Thus, a smaller change in the group can be a statistically significant difference. With the posttest-only approach, however, differences between members of the two separate groups would be a prime concern, so a larger change might have to occur in order to be statistically significant.

One argument against a pre-post approach is that participants may remember how they responded the first time. This may make their responses the second time open to conscious comparison by them. A second problem is that completing the instrument before the experience may sensitize them to how they are expected to change, and they may simulate change rather than actually change as a result of the experience. One solution is to "bury" the research question(s) in a larger questionnaire given before the experience. In this way, participants respond to so much material, they are less likely to remember when they complete the questionnaire the second time how they filled out any specific question the first time. They are also unaware of the "important" parts of the questionnaire until the end of the experience. A second solution is to use a line scale to measure each participant's attitude or characterization of himself or herself. The respondent draws a slash along a line that has few "anchoring" terms or descriptions (three at most, both ends and the middle). For example:

Right now I am feeling:

High Anxiety, Heavy Concern	Usual Anxiety or Concern	Good, Very Relaxed

A variation on this approach would be to use a line-response format (e.g.: High Anxiety _______________ Very Relaxed) in place of the usual five-category response format after a number of statements. However, this format poses problems in calibration because the respondent may not mark slashes on the scale consistently (equally far from the anchor for each increment) from item to item. Continued recalibration against the last mark, rather than against the original anchor, is necessary to assure accurate measurement.

Another technique is to divide an instrument into two parts, administering the first part to half the group and the second part to the other half of the group. After the experience, each half of the group takes the part of the instrument that it did not take the first time. Group scores for the pretest and, later, the posttest equal the sum of the scores across the two halves.

Another problem with the pretest-posttest format that may be endemic to training settings is the commitment to providing "learning experiences" and to giving people feedback about themselves. There is social pressure to have people score themselves and interpret their score patterns. If that is done, they are sensitized to the dimensions being assessed and may modify their behaviors in ways they would not have had they not taken the instrument. Even if a person does not change as a result of taking the instrument and obtaining feedback from it, the others in the group may change because of the instrument, thus presenting the individual with a group experience that is different from what it would have been if the instrument had not been administered.

In doing research connected with a training event, there is a conflict between giving participants useful feedback to help them to set personal agendas for the workshop and withholding all instrument data from them so that they will not be affected by knowledge of their scores. One control group is needed to discover if pretesting generates behavior or reactions that differ from the behavior and reactions of people who did not take a pretest. Another possible control would be to administer the pretest to a group and not give the feedback to the participants. The differences on the *posttest* between the experimental group (scored and discussed pretest scores), the first control group (no feedback from pretest), and the second control group (no pretest) will tell whether taking a pretest causes changes in posttest responses and whether taking a pretest and discussing one’s own scores causes even greater changes.

TRANSPARENCY/SOCIAL DESIRABILITY PROBLEMS

Instruments vary in terms of how obvious their intent is. Some of them measure one variable, and the respondents are virtually certain that they know what the variable is by the time they read the fifth statement. Other instruments measure twelve or more different characteristics; these sometimes contain distractors—or checks—such as “I breathe air.” The respondent has so many plausible guesses about what the questionnaire is measuring that (theoretically) he or she stops trying to figure it out and starts responding to each item at face value.

It is important to remember that transparency is not necessarily problematic. It causes trouble primarily when the instrument is being used to *evaluate* people (particularly when they are experiencing “evaluation apprehension”) or when the way in which it is being used lets the respondent decide whether or not he or she wants to “help” the researcher. However, when respondents are filling out an instrument for their own use, transparency becomes more of an annoyance than a real difficulty. It is an annoyance because the respondents are trying to remember how they have reacted to each situation described or how they feel about some statement. Their awareness that they can move their scores up or down some obvious dimension may be a temptation. This becomes a distraction because they may then try *too* hard to avoid “cheating,” thereby depressing their scores artificially.

It has been found that when people are put in a situation that arouses evaluation apprehension, they attempt to present themselves in ways that will make them appear more “socially desirable.” If the researcher suspects that the situation in which the instrument is being used may so strongly motivate some respondents to appear healthy, competent, tolerant, etc., that they distort their answers in a socially desirable direction, the researcher should take steps to make it difficult for them to distort or should "correct" for the distortion in computing results.

There are a number of ways in which this can be done. First, one can bury the real intent of the instrument among a number of distractor questions. However, if the questions in which the researcher is interested have a "look good" answer, the respondents will still distort. In this case, the questions may have to be rewritten. The second approach is to structure the questions so that there is no obvious "healthy" response. For example, a question might have a less guarded response if it is couched in general, rather than personal, terms. A third way of minimizing distortion is to ask respondents to check the statements that would make them "look good" if they agreed with them. Then have them check the statements that would make them "look good" if they *disagreed* with them. After examining the statements that received a lot of checks, the researcher could either rewrite the instrument (perhaps reversing the wording of some statements) so that a high (or low) score could not be obtained by giving "look good" responses or the researcher could develop a social desirability key. This key would include all the "look good" responses. A person would receive a score on the key whenever he or she gave a "look good" response. If a respondent were to obtain a high score (in the upper 10 to 15 percent of respondents) on social desirability, the researcher might wish to exclude that person from the analysis because there would be a good chance that the person was more concerned with looking good than with being honest.

MANAGEMENT PROBLEMS

A major concern in doing research in training settings is the effect of administrative timing on participant responses. Timing of measurement phases becomes important because participants often are at high emotional levels at the beginning and end of a workshop.

Using Pretests

The first day or first few hours of a workshop usually are marked by feelings of anxiety and general excitement. In many ways, the opening hour of the workshop is a poor time to administer a questionnaire. Often it is the facilitator's only opportunity to do pretesting, however, so he or she must live with the problematic elements of this timing. To help alleviate some concern, the facilitator can reassure the participants that they will receive as much feedback about their scores as possible and that their scores, although not anonymous, will be kept confidential.

One way to eliminate the difficulties involved in first-day administration of instruments is to mail the pretests to the participants a week or two before the training experience begins. The difficulty with this approach is that there are likely to be a few people who sign up for the workshop after the questionnaires have been mailed, a few who do not receive the questionnaires, and a few more who leave their questionnaires at home.

A second way to avoid measuring at the height of “opening” tension is to wait to administer the pretest until after a session or two has been conducted. The participants will have had an opportunity to become familiar with the training conditions and to put their irrational fears to rest; however, they also are somewhat different from the way they were before the first one or two sessions. As with many of the issues discussed in this volume, the user is faced with compensating, counter-balancing factors and must choose those alternatives that he or she considers most relevant to the particular situation and those most important to the intent of the research design.

Using Posttests

Just as the beginning of a training experience has its unique set of problems, so does the closing carry its special concerns. At the end of a personal-interaction event, people tend to be feeling happy, sentimental, euphoric, or perhaps even mystical. They usually are not in a mood to fill out a questionnaire. There is a legitimate question about the value of their responses if they are forced to fill out a questionnaire at that time. One solution is to give them the questionnaires and ask them to mail them back within three days. The questionnaires also can be mailed to them with the same instructions. The problem with these solutions is that some people will neglect to return them. Gathering research data through a mailing process has its own inherent problems, most of them centering around a human reluctance to interrupt personal priorities to respond to another person’s priorities when the other person is not present. Success in mailing questionnaires depends on the amount of patience and energy one is willing to expend in order to obtain the research data.

HUMAN PROBLEMS

An underlying premise of HRD work is the assumption that people should be open and honest with one another and that such behavior leads to the most productive outcomes of training. The tendency of researchers to avoid telling participants what they are attempting to accomplish by failing to give them feedback about their scores on pretests and by not processing data during the experience violates this pervasive assumption about openness. Participant hostility and resentment is a common result of researcher standoffishness. Hostile participants can produce invalid results. The researcher must choose between withholding feedback versus running the risk of participants knowing “too much,” between a clean research design versus a contaminated design, and between risking nonvalidity through insignificant results versus invalidity through biased results.

A SAMPLE RESEARCH DESIGN

When a researcher is concerned with studying the comparative outcomes of various training designs, he or she also must be interested in training versus no training. Control groups (those whose members receive no training) are used to compare the effects of experimental conditions with the absence of treatment. An effort must be made to keep all groups equal except for the specific training interventions to be tested.

For example, let us assume that one wants to explore the differential effects of a workshop that has a theory-centered design (a heavy focus on lecture material) and a workshop that is based on the use of structured experiences. It is conceivable that neither workshop will be found to produce lasting, observable changes in the behavior of the participants, and this possibility must be taken into account in the research design.

One might take applications for a training experience and randomly assign each participant to one of three groups:

1. Experimental Group 1: These people will participate in the workshop that focuses on theory.
2. Experimental Group 2: These applicants will attend the workshop built around structured experiences.
3. Control Group 1: These applicants will be excluded from the workshop experience. (As alternatives, they may be asked to wait until later to attend the workshop or they may be brought together for informal discussions.)

Ideally, one would have thirty to forty people in each of the two experimental groups and more in the control group. In addition, one may wish to establish other control groups, such as the following:

4. Control Group 2: Nonvolunteers, drawn from the same population from which the workshop applicants are drawn.
5. Control Group 3: A representative sample of people in general. (Sometimes relevant data on these people can be found in manuals of instruments.)

One might use a variety of instruments to assess outcomes. These should be selected to measure variables directly related to the goals of the workshop. The scales can be given on a pretest, posttest, and follow-up basis to determine which changes are temporary and which are lasting. One also may discover “delayed reactions.” Data can be collected from the participants themselves, from their associates, and from fellow participants and facilitators during the training.

EXPERIMENTAL STUDIES IN TRAINING

When an organization invests in equipment or materials, it generally takes steps to ensure that the investment is prudent—that the product is worth the price. To this end, when the purchase of untried items is contemplated, many companies either conduct their own analyses or contract for appropriate testing services. Curiously enough, however, many organizations buy or inaugurate expensive training programs or devices without proof of their usefulness. Experimental studies can provide that proof.

It is not necessary for all training programs to be subjected to experimental analysis. Prior use by other organizations, together with sound evidence of effectiveness, may be sufficient. However, all innovations, including new types of training equipment, need to have been tested under such conditions that results can be considered applicable to the new training situation. The greater the investment an organization plans to make in teaching its employees new skills, the more prudent it becomes to make an effort to ascertain the usefulness of the training involved.

WHAT CONSTITUTES AN EXPERIMENTAL STUDY

An experiment can be contrasted to the mere collection of opinions from trainees regarding the usefulness of a training experience and their suggestions for improvement. Such subjective reports can be helpful both to the designers of training and to those who implement it, but an experiment requires far more effort than an opinion survey.

Much of what is known about training innovations of the recent past—programmed instruction, the classroom communicator, and the computer, for example—comes from carefully executed experiments. Knowledge of the principles and techniques of experimentation helps in evaluating innovations that are being considered. The training administrator who understands the strengths and limitations of investigative procedures is in a better position to judge the claims of program vendors.

The crucial element in an experiment is control, that is, the elimination of all variables or conditions that might affect the conclusion except those of primary interest. Every training experiment incorporates the following elements:

1. An “experimental” or “independent” variable;
2. A “dependent” variable that presumably reflects changes resulting from the independent variable; and
3. One or more “control” variables that are ordinarily related to the dependent variable.

From *The Complete Book of Training: Theory, Principles, and Techniques* (Chapter 7), by G. Douglas Mayo and Philip H. DuBois (1987), San Diego, CA: Pfeiffer & Company. Used with the permission of the authors.

The main purpose of the experimental design is to eliminate or at least reduce the effects of the control variables on the dependent variable. In other studies with different designs, however, effects of control variables could be of great interest.

Because there may be more than one change brought about by training, an experiment may explore several different effects. For example, an innovation resulting in an improvement in reading skill also may be responsible for improvement in other areas of endeavor and in attitude toward work. In designing a study, one ordinarily identifies a single dependent variable as the target, but changes in other measurable variables may be of interest.

Hypotheses

The first step in designing a study is to develop an hypothesis, which is a statement that there is a relationship between the independent variable, which is more or less managed by the experimenter, and a dependent variable, which is not manipulated by the experimenter and which represents a goal of the training. Examples of independent variables are as follows:

1. The presence or absence of a specified type of training;
2. One type of training versus another type;
3. Different degrees of the same type of training;
4. Intervals of time between training sessions or programs;
5. The use or lack of use of a particular training device;
6. A comparison of the usefulness of two or more such devices; and
7. The usefulness of feedback or other procedures that might be motivating.

For technical reasons, the hypothesis studied directly in an experiment is generally stated negatively as the “null” hypothesis, that is, “the experimental variable has no effect on the dependent variable.” A null hypothesis is exact, and it can be refuted at a stated level of probability. Statisticians appreciate both of these features. However, because the refutation of a null hypothesis gives evidence that the positively stated experimental hypothesis is correct, we prefer to formulate positive hypotheses and collect data that may support them.

The Identification of Dependent Variables

In identifying dependent variables, one might start by determining how the trained employees will enhance the organization and what new and useful skills will have been gained if the training is to be considered successful. These determinations would lead to the identification of one or more dependent variables. In some cases identification of dependent variables is simple: Workers are taught new skills in machine operation, equipment maintenance and repair, order processing, or record keeping. In any of these situations, the resulting increase in proficiency would be apparent. In other cases, as on a production line, little variation in skill can be tolerated; consequently, no proficiency variable exists other than a “go”/“no-go” dichotomy. Another dependent variable might be the time or the amount of training required to reach the standard.

One approach to the identification of dependent variables is to list the direct objectives of the training program, together with possible side benefits. Dependent variables often are the same as the criteria of training effectiveness. Obviously, criteria are specific to training content; what is central to one program may be tangential or not applicable to another.

In the case of an organization that has introduced a new type of managerial training and is interested in determining the degree to which this training attains its objectives, the first task is to list measurable results, which might be the following:

1. Better internal communication;
2. Reduction of absenteeism, accidents, complaints, and disputes;
3. Increase in productivity and longevity of employees on the job; and
4. Improvement in employee morale and departmental *esprit de corps*.

These criteria are chiefly measures of the behavior of subordinates and reflect actions of the managers only indirectly. Furthermore, by the nature of their jobs, different managers have charge of functions that differ widely in composition and mission. One should not expect that the construction of an observational scale reflecting managerial performance will be easy; it is likely to require skill and hard work. The effort, however, could be rewarded in better understanding of both a managerial-training program and the role of supervisors in the success of a company.

Much easier to construct as a criterion is a test of managerial knowledge. Although some instruments of this sort are available from commercial publishers, it is probably better to use a device that parallels the training program and represents achievement in it. It can cover principles, practices, and the solution of managerial problems. Such a device cannot stand on its own because verbal knowledge of management and the ability to apply that knowledge are by no means identical. An investigation is necessary to determine whether increase in managerial knowledge is actually related to increase in managerial skill, as judged by changes in the behavior of subordinates.

Experimental Group and Control Group

In many psychological experiments, two groups are used: the “experimental” group and the “control” group. For both groups, all the “control” variables must be equivalent. As we discuss later, there are several procedures designed to make the two groups equivalent at the beginning of the study.

It is important to remember that group membership in and of itself may constitute the “experimental” variable. In a training study, the experimental group participates in or uses the innovation, while the control group is treated in the conventional fashion or receives no training at all, depending on the hypothesis being investigated. Either way, the independent variable may be thought of as dichotomous, with two arbitrary values: 1 for the presence of the innovation and 0 for its absence.

Actually, there may be two or more experimental groups in the same study, representing different degrees or different kinds of experimental innovation as well as two or more control groups, each representing absence of the experimental innovation.

The Four Requirements of a Research Design

Any research design must provide for the following elements:

1. Variation in the independent variable, normally through the existence and treatment of at least one experimental group, together with nontreatment of at least one equivalent control group;
2. Differential treatment of the experimental and control groups;
3. Elimination or at least reduction of the effects of all control variables by methods to be discussed later, such as random assignment, use of the same group at two points in time, or matching; and
4. Measurement of the dependent variable in the two groups at the conclusion of the study.

The purpose of the entire procedure is to investigate whether or not there is a dependable relationship between the independent and dependent variables. Sometimes, when an effect has been demonstrated, it is possible also to assess its strength and practical importance.

KINDS OF EXPERIMENTS TO EVALUATE TRAINING

The design of an experiment is simply a detailed plan. Normally, it describes the reasons that the study is needed and the setting in which it will be conducted. This information may be followed by a description of possible approaches and why the selected approach was chosen. Often, several sources of data will be included in the plan. It is important that the designer not attempt to substitute a large volume of low-quality data for data of good quality. No amount of poor-quality data will result in an adequate evaluation; it may even prove to be misleading.

The design should include copies of the measuring instruments and information concerning their characteristics, such as validity and reliability. Any sampling procedure should be described; if random sampling is not employed, the alternative should be justified.

A plan for analysis should be included, and examples of the type of information that will be developed should be set forth. For example, the plan may state that if a certain result is shown by the data, a particular conclusion can be drawn and a specified action is indicated. It is advisable to prepare a number of such statements during the design phase. Not only do they apprise all concerned of what may be expected, but they sometimes point to correctable deficiencies in the study. This is a much better time to identify design defects than after data have been collected and analyzed.

Random Assignment As a Control-Group Design

Random assignment of trainees to one or more experimental groups and to a control group normally requires a pool of prospective trainees and control of their assignment by the person or team responsible for evaluation. Obviously, there are a great many instances in which this requirement cannot be met; but when it can be met, this procedure has excellent potential for providing valid information. The term "random" has a special meaning in research and evaluation. Random assignment does not simply consist of designating the first, third, and fifth trainees (and so forth) who enroll in a training program as members of an experimental group and the second, fourth, and sixth (and so forth) to a control group. Although assembling groups in this way may approximate the results of random assignment, it is better to assign each trainee a number and then use a table of random numbers to designate half of the trainees as members of the experimental group and the remaining trainees as members of the control group. When several experimental conditions are included in the design, this procedure requires only minor modification in order to assign trainees to more than one experimental group. Three examples of random assignment are described in the following paragraphs.

Use of Posttest Only

After random assignment has been made to one or more experimental groups and to a control group, the training innovation is implemented without the administration of a pretest. A pretest is not necessary because proper random assignment assumes that the groups are equivalent. This assumption is accurate if the number of trainees is fairly large—fifty or more in each group. The appropriate test of statistical significance is based on probability theory and takes into account chance differences.

Use of Pretest and Posttest

When this design is used, a pretest is given after random assignment of the trainees to the experimental and control groups. The experimental group then participates in the training innovation. As in other designs, the control group may receive the regular training or may receive no training at all, depending on what question the evaluation is expected to answer.

After the training, a posttest is given to both groups. The primary comparison is between the gain made by the experimental group and the gain (if any) made by the control group. If the gain by the experimental group exceeds that of the control group by a statistically significant amount, the difference may be considered to be evidence of the effectiveness of the training innovation. A simple extension of this design occurs when more than one experimental program or condition is included in the design. Basic comparisons and interpretations remain the same.

Use of a Matched Sample

In this design, trainees who have the same score or characteristic on one or more control measures are paired. The closer the relationship of the score or characteristic to the posttest or other criterion measure, the more effective the matched-sample design becomes. After the pairs have been formed, one member of each pair is assigned to the experimental group and the other is assigned to the control group by a procedure that tends to avoid systematic bias, such as the use of a table of random numbers. From this point, either of the first two procedures described above may be used. Statistical procedures appropriate to the matched-sample design take into account the strengthening of the design due to matching; the result is that smaller differences between groups indicate greater statistical significance than otherwise would be the case. Other, less-adequate designs that employ matching but do not employ random assignment can be used when random assignment is impractical.

Limitations of Random Assignment

Despite its advantages, random assignment has certain limitations. As noted by Cook and Campbell (1979), random assignment does not completely ensure that the experimental group and the control group will be equivalent prior to the training. The effectiveness of random assignment requires fairly large numbers. When the number of trainees is small, there is less assurance that the experimental group and the control group will be equal before the training begins and, hence, less assurance that a correct conclusion will be drawn from post-training differences. Tests of statistical significance reduce the probability of such an error by requiring a larger difference between the experimental group and the control group when the number of trainees is small, but it is still possible for the data to suggest an erroneous conclusion.

Although the minimum number of trainees required to conduct an adequate evaluation varies with circumstances, a good rule of thumb is to conduct an evaluation with no fewer than thirty-five trainees in the experimental group and a similar number in the control group. However, one should keep in mind that in cases in which a significant difference is not found, numbers larger than this normally are necessary if one wants to be able to infer that any difference of importance would have been shown by the data.

From another point of view, random assignment or any other control-group design may create inequities in the way individual trainees are treated. Not only is this undesirable from the standpoint of fairness, but individuals assigned to either the control group or the experimental group may behave in an atypical way because of being so assigned, thus obscuring differences that otherwise would result. Examples of this occur when some members of the control group think that they are not receiving the best training available or when some members of the experimental group believe that they are being manipulated or “treated like guinea pigs.” Sometimes it is possible to assign trainees to groups without their knowing that they are participating in an experiment. Obviously, this solves the problem of the effect; but it may raise ethical issues, as discussed in the code of ethics of the American Psychological Association (1982).

It is fair to conclude that random assignment involving a large number of trainees tends to make pertinent characteristics of an experimental group and of the corresponding control group initially equal; but it would be inaccurate to conclude that random assignment completely equates all pertinent characteristics, especially those having to do with human reaction. Even so, when practical considerations permit its use, an evaluation design involving random assignment to a control group and one or more experimental groups will usually prove superior to other designs in program evaluation.

Quasi-Experimental Designs for Program Evaluation

When it is not possible to assign trainees randomly to experimental and control groups, it still may be possible to collect worthwhile information through the use of quasi-experimental designs. Such designs are somewhat similar to the experimental designs previously described, but conclusions from them tend to be more equivocal. Normally, the results obtained from quasi-experimental designs permit a greater number of different interpretations than are possible with carefully designed and executed experiments.

A quasi-experimental design should not be thought of as an acceptable alternative to an experimental design when control of conditions, including random assignment, is possible. Campbell and Stanley (1963) describe a number of quasi-experimental designs and precautions that should be observed in interpreting the resulting data. Cook and Campbell (1979) further develop this topic with special emphasis on field settings. Three examples of quasi-experimental designs are as follows:

1. A design that uses a nonequivalent control group with pretest and posttest;
2. A time-series design; and
3. A regression-discontinuity design.

Nonequivalent Control Group with Pretest and Posttest

In this design, intact classes are assigned randomly to a control group and to one or more experimental conditions. The trainees as individuals are not randomly assigned to either (or any) of these classifications. Only after individuals have become members of groups by usual assignment procedures are intact groups assigned randomly to different conditions. For random assignment to have much effect, there should be, if possible, a number of groups involved; but the basic procedure can be used with only one control group and one experimental group.

The control group and the experimental group are given the same pretest and posttest. The only planned difference is that the experimental group participates in the innovative form of training, while the control group participates in the old form of training or no training at all, depending on the question that the study is expected to answer.

As a rule, the statistical procedure known as the analysis of covariance is used in the analysis of the data. This procedure seeks to compensate for differences between the pretest scores of the experimental group and those of the control group. In effect, the procedure predicts posttest scores from pretest scores and indicates the probability that the difference between the predicted scores and the observed scores for the experimental group and the control group could have occurred by chance. If the probability is low, it may be possible to infer that the difference was the result of the experimental training. However, this statistical equating of groups does not necessarily compensate for initial lack of equivalence in the groups. Lord (1969) has pointed out some of the problems involved.

Time-Series Design

As the name implies, the time-series design involves measuring the performance of several groups both before and after the innovation is introduced. Thus, a series of measures is established in which the change to be evaluated occurs near the middle of the series. One can readily see that this is a stronger design than simply comparing the performance of the last group before the innovation was introduced with the performance of the first group to participate in or use the innovation. The measurement of several groups before and several groups after the introduction of the innovation makes it easier to detect an atypical group than would be the case if only two groups were used.

This design lends itself to both graphical presentation and statistical analysis. If most of the groups measured prior to introduction of the innovation have about the same level of performance, and most of the groups measured after the introduction have a higher level of performance, this result suggests that the improvement may be due to the innovation. On the other hand, if performance is about the same both before and after the introduction of the change, it suggests that the innovation is not superior to the old form of training.

Regression-Discontinuity Design

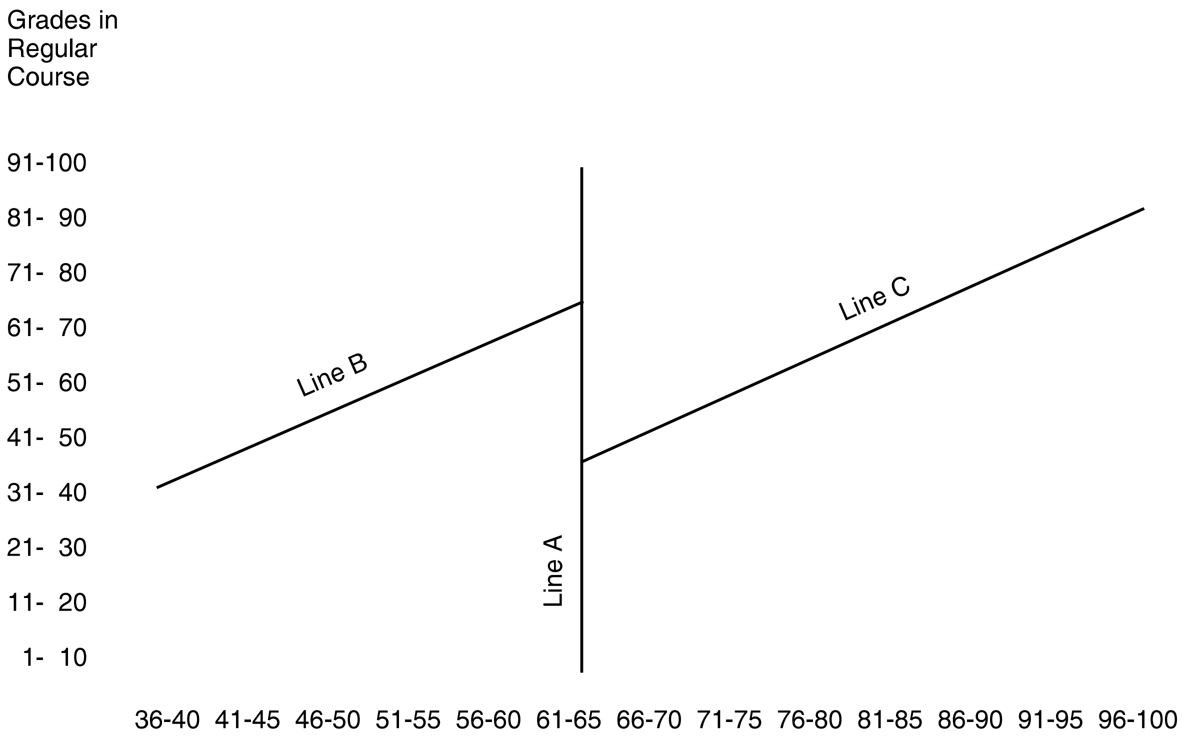
The regression-discontinuity design is similar to the time-series design in that it involves measurement of a number of groups and is amenable to both graphical presentation and statistical analysis. However, it is used under different circumstances. The regression-discontinuity design is especially useful when virtually all applicants must be given the training for a certain job.

Let us consider an example in which an industrial organization is being encouraged to employ and train personnel whose aptitude level normally would be predictive of failure to complete the required training. A preliminary training program is devised to prepare the low-aptitude applicants for the regular training, and it is decided that the effectiveness of this program should be evaluated. In addition, an aptitude test used by the organization is highly predictive of performance in the regular training; the score on this test that is traditionally used as a cutoff to determine entrance into the training program is 66. Let us assume that there are at least ten applicants in each five-point interval of scores on this test from the 36-40 interval to the 96-100 interval. Thus, there are prospective trainees in six 5-point intervals on the test below the conventional cutoff score and in seven 5-point intervals above the cutoff score.

The preliminary program is conducted for the prospective trainees with aptitude scores below the cutoff score of 66, and on completion of the program these trainees participate in the regular training. The prospective trainees having aptitude scores of 66 or above on the test also complete the regular training. It is then possible to plot the mean score for each 5-point aptitude interval from 36-40 to 96-100 (on the horizontal axis) against mean program grades earned by trainees in each aptitude interval (on the vertical axis).

If the preliminary program has been effective, there should be a “discontinuity” between the line connecting the 5-point intervals including the trainees who participated in the preliminary program (the low-aptitude trainees) and the trainees who did not participate in the preliminary program (the high-aptitude trainees). In other words, an upward extension of the trend line for the trainees who participated in the preliminary program should be higher than the trend line for the trainees who did not participate in the preliminary program. If there is no discontinuity between the trend line of those who participated and the trend line of those who did not, the data fail to support the contention that the preliminary program was effective.

An illustration of this discontinuity is shown in the figure that follows. Line A indicates the aptitude level dividing the trainees who participated in the preliminary program from all other trainees. Line B shows grades made in the regular training program plotted against aptitude-test scores for the trainees who participated in the preliminary program (low-aptitude trainees). Line C shows grades in the regular training program plotted against the aptitude-test scores for the trainees who did not participate in the preliminary program (high-aptitude trainees). The discontinuity between Line B and Line C suggests that the preliminary program was effective. Lines B and C may be plotted from mean grades made by the trainees in each interval on the aptitude test, or they may be regression lines resulting from correlational analysis.



Hypothetical Situation Illustrating the Regression-Continuity Design

THE RELIABILITY AND VALIDITY OF DEPENDENT VARIABLES

Reliability

A dependent variable must have reliability. As indicated earlier, the reliability of a measure is concerned with its stability and consistency. Thus, if a device does not provide a similar assessment of individuals at different points in time (when nothing has intervened to alter their performance), one cannot have much confidence in it. Similarly, consider a printed test that is made up of item pairs, one for each instructional objective; and let each pair consist of an *a* item and a *b* item, both measuring the same objective. If, when the test is given at the end of the training, the scores on the *a* items are substantially different from those on the *b* items, the test as a whole is not reliable.

In a training study, reliability is best defined as the consistency between two forms of a device designed to measure the dependent variable. Such consistency could be indicated by a substantial correlation (.55 or more) between the scores on the *a* items and the scores on the *b* items, provided that the two scores are added to constitute the total dependent variable.

If an objective instrument is being used as a criterion and it is possible to divide it into two equivalent halves, it is possible to obtain an estimate of its reliability from data obtained during the course of the study. If r_{ab} is the correlation between the *a* and *b* half scores, then the reliability of the whole can be estimated as $2r_{ab}/(1 + r_{ab})$. For example, if the correlation between the two half tests is .55, then the reliability of the sum can be estimated as .70.

Low reliability is a signal that further development is indicated, perhaps with the assistance of a specialist in psychological measuring devices (a “psychometrician”). With multiple-choice items, distractor counts can lead to the replacement of decoys that fail to attract responses; and decisions can be made about items that are too easy or too difficult as measured by the proportion of correct answers. Modifications of the instrument based on such item-analysis information can lead to improved reliability, which, in turn, can lead to higher validity.

All of this, however, applies only to achievement measures in programs that are targeted to increase skills and that do not insist that all graduates successfully complete all aspects of the training, as is likely to be the case when safety is of paramount importance. When safety is the primary concern, there is a special consideration with regard to reliability: If it can be determined that an instrument actually matches job requirements, reliability can be assumed.

Validity

A dependent variable also must have validity. Validity has to do with the degree to which correct inferences can be drawn from the information provided by a measure. For example, if there is a high correlation between scores on an end-of-training instrument and subsequent performance on the tasks that the training was designed to teach, one may infer that the end-of-training instrument is valid. One also may infer that the training is effective in fostering the desired knowledge and skills. Of course, in order to make these inferences, one must ensure that the trainees had not attained proficiency prior to entering the training. There are two closely related (but not identical) varieties of validity of interest to trainers: (a) content validity and (b) criterion-related validity.

Content Validity

Content validity has to do with the degree to which a measure represents the knowledge and skills that a program undertakes to teach. When a program is constructed to achieve specified objectives, those objectives are normally derived from operational requirements identified by means of job-task analysis.

Content validity is the primary type of validity that is “built into” the training-development model. In devising an instrument that will possess content validity, one often bases the items directly on training objectives. It is not unusual for an item to be essentially the same as the corresponding objective, except that the format is that of a test item.

The implications of content validity for evaluation of training are clear-cut. If an end-of-training instrument has been demonstrated to have a high degree of content validity and a particular group of trainees performs poorly on the instrument, this outcome indicates that something is wrong. It may be that the training for that particular group was not up to par; it may be that the trainees who made up the group did not have the requisite qualifications for, or interest in, the training; or there may be some other explanation. It becomes the task of the training manager to identify the source of the problem and to correct it.

To a large extent, content validity is internal. In other words, the items that make up the instrument are *constructed* to measure the training objectives. In this situation, there is always the possibility that the person who wrote the items was not completely successful in constructing items that adequately measured the objectives. This is one of the reasons that it is desirable to demonstrate that an instrument also possesses criterion-related validity.

Criterion-Related Validity

Criterion-related validity is external and exists whenever there is a definite relationship between any type of instrument and performance on the job. In training research, it is important to study relationships between training success, however measured, and subsequent work effectiveness. Lack of such a relationship throws doubt on the usefulness of the training. The criterion may be performance on a job sample, supervisor ratings of the quality of work, the number of acceptable units produced, or other measures of how well an individual performs assigned tasks. Measurement of job performance should be a byproduct of the job-task analysis used in determining operational requirements and specific training objectives. Development of evaluation procedures should be an integral part of planning a training-research project. In discussing training outcomes, it is convenient to assume that the dependent variable is measured in units and that it can be described with conventional statistics: the mean as a measure of central tendency and the standard deviation as a measure of dispersion or variability.

THE RECORD SYSTEM

Essential in any scientific investigation is a good record system. All data need to be identified as to the source, the nature and date of the observations, and the individuals involved. Identification should be sufficiently complete that a person who encounters the records later without prior knowledge of the project can understand all entries. If instruments have been used, they should be precisely identified. With subjective evaluations or ratings, methods and observers should be carefully stated.

It is important that all manual records be legible and more or less permanent. When data are entered into a computer bank, complete details should be available as to format and how the data may be called up, not only for analyses that have already been designed but also for use in studies that may be conceived later. Modern computers are versatile and can summarize large masses of information very efficiently, but using them for record keeping requires good planning. Published programs are extremely useful, but occasionally new programming is necessary. Sometimes reports compiled by computer are more or less in final form, but data summarized by machine are often processed further by manual methods.

STATISTICAL SIGNIFICANCE

In our previous descriptions of experimental designs, we used the phrase “statistical significance” several times. The term merits explanation.

Consider two groups that were equivalent at the beginning of a study but when measured at the end of training have different means on the dependent variable, perhaps a performance measure of the skill acquired. Except in the rarest of instances, there will be a difference in the means of two groups on any variable. In the case of a training study, the question is whether or not the difference is meaningful, that is, whether or not it is an actual reflection of the way in which the two groups performed.

On the basis of probability theory, mathematical statisticians have worked out the distribution of the differences that would be found between pairs of means drawn at random from an unlimited population in which the differences between pairs of means average zero. The distribution takes into account the number of cases and the distribution of scores within each group.

When an observed difference between two means is statistically significant, it cannot be considered a chance occurrence. In general, the greater the number of cases involved in the comparison of the two means, the smaller the difference need be to be considered an event not governed by chance. Accordingly, with large numbers of cases, a very small difference can meet the test of significance. On the other hand, the size of the difference needed for significance increases as variability internal to the groups increases.

The appropriate method of determining significance is a technical matter that depends, in part, on whether or not the means are inherently related. If the two samples are unmatched and come from independent sources, the means have no inherent relationship and are treated by the procedures for uncorrelated means. If, on the other hand, one is dealing with two matched samples or with the same group on two different occasions, the means have a built-in relationship that must be taken into consideration by using procedures for correlated means.

Because larger random samples are more stable than smaller samples, the difference between two means required for significance decreases as the samples become larger.

One-Sided and Two-Sided Tests

Another consideration is the contrast between “one-sided” and “two-sided” instruments of significance. If one is interested in any difference at all between the means of two groups, a two-way or two-sided statistic is appropriate. Such would be the case very rarely in training studies. Generally, training research is designed to evaluate an innovation believed likely to be more effective than the approach that is currently in use.

When there is an experimental group using an innovation and a control group using a conventional approach, the “null” hypothesis is that the two means on the dependent variable (in the parent population represented by the sample) are identical. The experimental hypothesis is that the mean of the group participating in or using the innovation is higher than the mean of the control group. To test whether the innovation is effective, one first inspects the means to determine whether the difference is in the expected direction; if it is, one then applies a “one-sided t-test” of significance. If the difference is not in the expected direction, the innovation is considered to be ineffective.

The difference is divided by the “standard error of the difference,” which is the estimated standard deviation of an unlimited number of pairs of means of same-sized groups drawn at random from an unlimited population in which the variability is approximately the same as in the samples. The formula for the standard error of the difference, its logical basis, and the details of its application in research situations are given in many texts on statistics, together with special procedures to be used when groups are small (fewer than thirty members). A variation of the formula must be used when the means are inherently related, as when the same group is measured before and after training.

When a difference divided by its standard error is 1.645 or more, it is significant at the .05 level. In other words, there is one chance in twenty (or less) that the difference is the result of sampling; and there are at least nineteen chances in twenty that something other than chance has produced the difference. If the ratio of the difference to its standard error is 2.326 or more, significance is at the .01 level, meaning that there is only one chance in one hundred (or less) that only sampling is operative.

It must be pointed out that a difference can be reliable and still not amount to much in the real world. A small difference based on a large number of cases might be statistically significant—perhaps at the .01 level—while a large difference between small groups might not meet any statistical test of significance. Statistical significance is a necessary but not sufficient condition for a research finding to be considered important. Instead, importance is dependent on two requirements:

1. The dependent variable must be important in the world of work; and
2. The gain ascribed to training must be large enough to have real value at work.

Relation of Statistical Instruments to Decision Making

A training innovation may yield a significantly higher mean in an experiment and yet not be appropriate to introduce in preparing people for the actual work situation. If the study is based on large numbers of cases, a difference that may not mean very much in the practical situation may be in the expected direction and statistically significant. Considerations beyond statistical significance include the cost effectiveness of the innovation, public relations, acceptance by trainees, and availability of the resources needed for the proposed change in the training program.

The study itself usually provides only an incomplete basis for a decision. A crucial question is the effect of the obtained difference between the means in actual operations. One must consider whether or not the superiority of the experimentally trained individuals can be translated into reduction of training costs or greater effectiveness on the job as indicated by increased quality or quantity of output. The decision regarding what to do about the innovation may be hard or easy to make, but it is important to remember that it rests in the hands of administrators who are paid to make decisions. The role of the investigators is to provide the facts.

END PRODUCTS OF EVALUATIVE STUDIES

This chapter deals with the following issues regarding the reporting of results from evaluative studies:

1. Answers to questions posed by the investigation;
2. The audience for the answers;
3. The way in which findings are to be presented; and
4. The actions that logically follow.

The alternative actions are as follows:

1. Confirmation of the present program as satisfactory;
2. Rejection of the present program in favor of an alternative; or
3. Changes in the present program as indicated by specific results.

Although the ultimate purpose of a organization-sponsored training program is the improvement of the organization, intermediate considerations include the following:

1. Acceptability to trainees, an issue that is related to motivation;
2. Coverage of needed skills;
3. Identification of program elements that are redundant or superfluous;
4. Transferability of the skills developed during the training to work situations; and
5. The degree to which the results of the training endure over time.

With any evaluative study, the more specifically the questions posed have been answered, the more useful the results. In addition, the findings of such a study occasionally provide answers to questions that were not originally envisioned. The primary means of conducting an evaluation are through the use of experiments and the administration of questionnaires or similar instruments.

EXPERIMENTS

Experimental investigations, as discussed previously, have both advantages and disadvantages over studies based on instruments such as questionnaires. The advantages include the following:

From *The Complete Book of Training: Theory, Principles, and Techniques* (Chapter 8), by G. Douglas Mayo and Philip H. DuBois (1987), San Diego, CA: Pfeiffer & Company. Used with the permission of the authors.

1. Conclusions are based on objective results. In general, one can depend on the findings.
2. Success on the job can be used as a criterion (dependent variable). This sometimes makes it possible to estimate the dollar value of the training.
3. An experimental study often can be conducted without the subjects' awareness. This eliminates a possible source of bias.
4. The independent, dependent, and control variables are all carefully defined. Subjects and experimental conditions are described. Accordingly, the area of applicability of the findings can be ascertained and the study can be replicated.

The experimental approach also presents several disadvantages:

1. The experimental variable typically has limited variability. Often there are only two variations. Consequently, it may be difficult to pinpoint aspects that need emphasis or reduction.
2. It is difficult to estimate changes over time because each time interval requires a new set of measurements.
3. Attitudes and motivation can be assessed only indirectly or by means of a separate set of measurements.
4. Experiments are costly in terms of the time required for planning and administration. Some experiments also are expensive in terms of trainee time required.

INTERPRETING RESULTS OF EXPERIMENTS

When a training experiment yields positive results, the benefits to the sponsoring organization should be made clear through written reports and oral presentations to management. Verbal descriptions, numerical findings, and charts reinforce one another. Emphasis should be on the usefulness of the results and the clarity of presentation. The fact that audiences vary greatly in their ability to understand technical findings and to perceive applications should not obscure the requirement that the first objective in reporting a study is to present the big picture as clearly as possible and then fill in the details as the need for more information becomes apparent.

Consider the case in which a difference in a dependent variable clearly is not the result of chance and is of considerable practical importance. For example, suppose that an organization has been able to identify two groups of salespeople (thirty in each group) who were judged to be approximately equivalent in sales experience, aptitude for selling, product knowledge, and assigned territory. Also assume that the organization's practice has been to assign each newly appointed salesperson to a two-week program covering detailed information about the product line, markets in which the organization has been successful, and suggestions from top producers. All of this is "conventional training." Without the knowledge of those not selected for "special training," one of the groups is given an additional one-week program in sales strategies and techniques, emphasizing analysis of clients' psychological needs and ways of inducing action based on knowledge of those needs. Then assume that in the six months after training, gross sales for the two groups were as shown in the table.

Comparison of Gross Sales for Two Sales Groups

Comparison of Gross Sales for Two Sales Groups

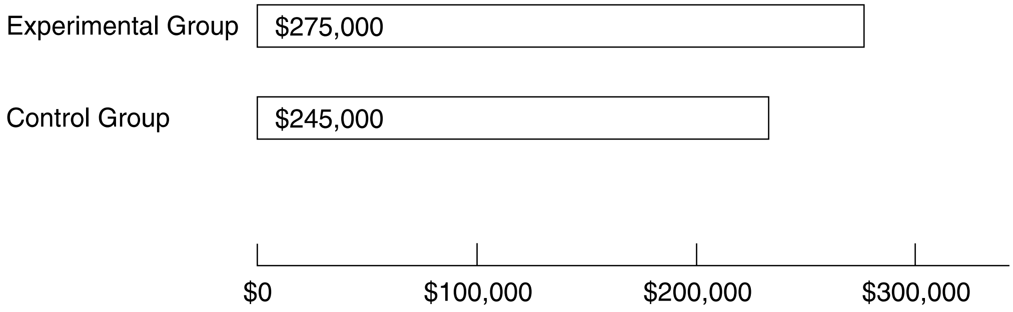
Group	Number	Thousands of Dollars	
		Mean	*Standard Deviation*
Experimental Group*	30	275	65
Control Group**	30	243	50

*Received both conventional and special training

**Received conventional training only

Further assume that when the study was designed, it was decided that a difference favoring the experimental group would be considered conclusive if it were significant at the .05 level. In terms of the usual one-sided t-test, this difference is significant at the .05 level (although not at the more rigorous .01 level).

In this example a simple bar chart is useful, showing the difference between the means graphically (see figure).



Average Sales During the Six Months After Training

The line under the graph provides the scale for interpretation. Management will be interested in the fact that the members of the experimental group averaged $32,000 more (approximately 13 percent more) in sales than the control group. Whether this means that the experimental training actually represents a gain in dollars is a matter to be studied. The net profit on the $32,000-average increment must be compared with the total cost of the training, which may include tuition, living expenses, and travel for the trainees as well as overhead costs. Another pertinent concern may be loss to the organization resulting from the absence of the trainees from active selling.

Still another matter to consider is the length of time needed by the organization to recover training costs. Some organizations may be willing to recover training costs over a period of years. However, assume that in this case the organization decided that a satisfactory program should pay off in six months. If gross organizational profit on its sales were 8 percent (that is, $2,560 on sales of $32,000) and the cost of training were estimated at $2,200 per individual, this training would have shown a modest profit for the organization.

In real-life situations, payoff needs careful study. Appropriate criteria that can be translated into monetary terms must be sought. Sometimes a program can be shown to be highly beneficial in monetary terms. But there are programs that may lead to no monetary gain or even a loss.

PREPARING REPORTS ON RESULTS OF EXPERIMENTS

After an experiment has been conducted and interpreted, two reports should be prepared: (a) a complete report for those who are fully aware of research procedures and terminology and (b) a summary for the managers who may not be interested in checking the details but who are in a position to use the information in making decisions. By providing all the specifics of the study, the complete report substantiates the summary and can be consulted whenever clarification is needed.

The following outline is suggested for the complete report:

1. The general purpose of the investigation;
2. The specific hypotheses;
3. The subjects: who they were, how they were recruited for or assigned to the study, and their backgrounds (pertinent personal information, including prior training and related experience);
4. Identification of experimental, dependent, and control variables and how they were defined and measured;
5. The procedures: the instructions given to the administrators of the research, the instructions to subjects, the instruments used, implementation, details about time, and the conditions under which the dependent variable(s) were measured;
6. Statistical analyses of the data: detailed findings and graphic presentations if appropriate;
7. Interpretation of the statistical analyses; and
8. Conclusions and suggested applications.

The complete report should have the following characteristics:

1. The information presented should be such that at a later time an investigator could reproduce the study with essentially identical variables and method and with a comparable group of subjects.

2. The information should be sufficiently detailed that it provides answers to almost all, if not all, questions that might be raised about what was done, the reasoning behind what was done, and possible applications of the conclusions.
3. The format of the report should be such that all parts can be readily comprehended by the intended audience. The presentation always includes verbal descriptions and often includes photographs, charts, graphs, and tables.
4. If there is a possibility that a different type of analysis would result in different conclusions, the original data should be made available either in an appendix or by stated arrangement.

Preparing Bar Charts

The previous figure is a simple example of the result of a study reported in graphic form. The essential element of a bar chart is that the dependent variable is depicted in a linear fashion, with variation in length representing variation in amount. The effectiveness of the display depends largely on whether the dependent variable is important to those who are reviewing the results. In sales training, the dependent variable might be number of units sold, gross value of the units, or net profits to the organization. In managerial training, the criteria might be increase in production, reduction of manufacturing time or cost, smoother internal operations, or increases in employee morale. All such dependent variables lend themselves to presentation through the use of bar charts.

Criteria such as those listed in the preceding paragraph present practical and theoretical difficulties. In evaluating sales training, for example, it may be impossible to assume that each salesperson has an equal opportunity to achieve sales measured on a common scale. Products or services assigned for sale may vary widely; territories differ; and the supervision received may vary from one individual to another and may have a definite effect on production. Probably the best ways to neutralize to some degree the effect of extraneous variables on a criterion are to define and measure the criterion carefully, to adjust it by statistical methods when feasible, and to use a substantial number of subjects.

In evaluating managerial training, the difficulties that one normally encounters with criteria are compounded because, almost by definition, each managerial job is unique. The manager directs an essential, distinct portion of the total enterprise and, at least in a progressive organization, is expected to improve ways of accomplishing objectives. One way of evaluating success in managerial training is to rely on the judgment of the trainees' superiors with regard to the degree that the training has improved job performance.

In some business and industrial operations, it is occasionally possible to identify sizable groups of individuals who have more or less identical assignments and for whom a defined criterion is appropriate. This simplifies the task of evaluating training. In other cases, reliance on the judgment of superiors is essential.

FURTHER CONSIDERATION OF THE QUESTIONNAIRE METHOD

A second—and very powerful—way to evaluate training is to administer an instrument to trainees either at the conclusion of the program or after an interval of time, which permits evaluation in relation to actual application at work. The use of questionnaires and other instruments as a basic method was discussed previously in this volume. Here the emphasis is on using the results.

Questionnaires have a number of attractive features:

1. A wide range of topics can be covered.
2. Correctable training deficiencies often can be identified.
3. There is no limitation on the degree of specificity of the evaluation; either an assessment of the overall program or reactions to specific details can be elicited.
4. Trainees generally find questionnaires acceptable and readily cooperate in completing them.

Questionnaire Requirements

In developing a questionnaire, one should address the following issues:

1. Is the language very clear?
2. Is all the information requested germane to the inquiry? (Only when the response has the possibility of being the basis for some sort of action should a question be included.)
3. When a multiple-choice format is used, do the suggested replies cover the range of appropriate answers? Are the different replies assigned to appropriate steps?
4. Is the length reasonable? Can the questionnaire be completed in a few minutes?
5. Will the questionnaire fulfill its purpose without making undue demands on the respondent? (Generally speaking, no information should be requested unless the respondent is likely to have it readily available.)
6. How will the questionnaire be accepted? (Even in a questionnaire that protects the respondent's anonymity, some questions could be objectionable. Good taste is important!)

The End-of-Training Questionnaire

An excellent time to administer an evaluative instrument is during the final session of a program. The group is likely to be intact, and sampling may be 100 percent. Details of both the program content and the training situation are clearly in mind. Simple steps can ensure the respondents' anonymity so that opinions are expressed freely. Because this is a good time for trainees to "have their say," cooperation is likely to be excellent.

Good information can be obtained about:

1. The trainer and the impression that he or she makes (the degree to which the trainer is perceived as hard working, informative, helpful, and successful in creating a good learning environment);
2. Trainee acceptance of any special techniques or training aids introduced into the program;
3. Perception of the adequacy of the physical arrangements (space, freedom from interruptions, heating, lighting, ventilation, furniture, audibility, and visibility);
4. The best feature of the program as evaluated by the trainees;
5. The least satisfactory aspect(s) of the program; and
6. Whether the program seems likely to result in a permanent gain in a desired skill.

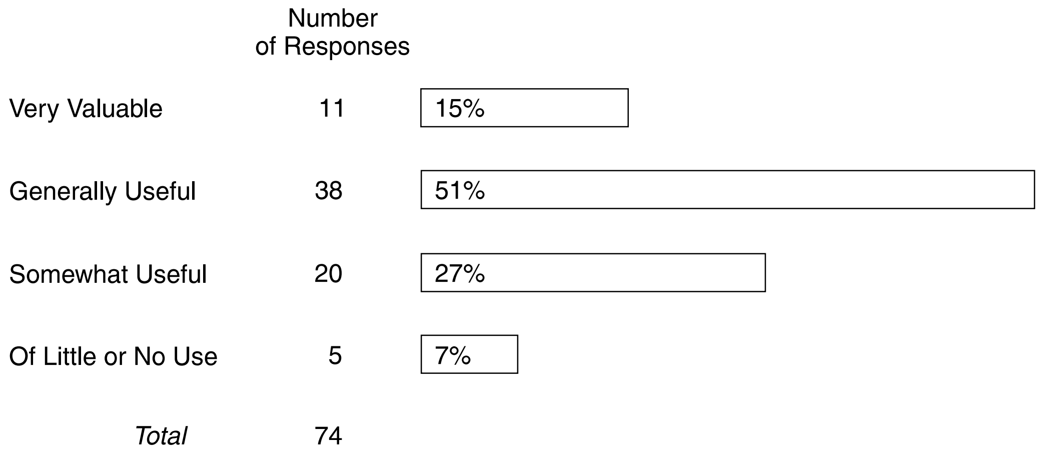
When the trainees are employed by a particular organization and have participated in an organization-sponsored program, two questions provide good overall evaluations:

1. Does the trainee consider the program a good investment of the organization’s money?
2. Would the trainees recommend the program to other employees as a worthwhile investment of their time and effort?

In addition to the structured items, space should be provided in which the trainees can print comments or suggestions. Providing such an opportunity sometimes evokes reactions to topics that were not considered when the questionnaire was designed.

Information from the structured items of an end-of-training questionnaire is easy to tabulate. (Comments are sometimes reported verbatim, but are more intelligible to the reader if grouped or summarized.) Most multiple-choice items present degrees of response that lend themselves to display in a bar chart such as the one shown in the figure that follows. When presenting such a chart, one can point out the impact of the results by making a comment such as “Of the seventy-four respondents, sixty-nine or 93 percent reported that they believed that their newly acquired skills would be at least ‘somewhat useful’ when they returned to their jobs.”

Question: How useful will the skills that you developed in this course be to you in your job?



Bar Chart for Responses to Multiple-Choice Item

A suggested format for a partial end-of-training questionnaire is shown in the next figure. In addition to the items provided in the figure, items should be constructed to cover other aspects of the program for which reactions would be useful. Some of these items might include ones about the trainer (whose name should be listed on the questionnaire), the specific training aids or procedures that were used, the physical facilities (training rooms, living quarters, and so forth), meals, coffee breaks, and the schedule of the program.

End of Training Questionnaire[1]

Instructions: The organizers of this program want your frank evaluation of its value and how it was conducted. You need not write your name on this questionnaire; no attempt will be made to identify the responses of any individual. It is hoped that your replies will be useful in improving future programs. For each multiple-choice item, write a check mark in the box next to the response that is most appropriate for you. When comments are called for, please print.

1. I regard this program as:

   □ Very valuable for my work.

   □ Definitely useful for my work.

   □ Somewhat useful for my work.

   □ Of little or no use for my work.

Comments:____

____

____

2. The instruction in this program was:

   □ Very interesting and highly effective.

   □ Fairly interesting and reasonably effective.

   □ Marginally satisfactory.

   □ Boring—should be improved.

Comments:____

____

____

3. The best feature of this program was____________________________________

____

4. The feature of this program that I found least satisfactory was____________________

____

5. The investment made by my organization in my training in this program:

   □ In the long run will pay big dividends.

   □ Should be considered worthwhile.

   □ Is neither good nor bad.

   □ Should be considered a waste of money.

6. To someone in my situation, I would recommend this program:

   □ Enthusiastically.

   □ As very good.

   □ With reservations.

   □ As something to avoid.

[1] The authors thank Drs. Robert E. Lefton and V. Ralph Buzzotta of Psychological Associates, Inc., in St. Louis, Missouri, for permission to use this adaptation of their questionnaire.

The Delayed Evaluation

An administrator of a training program may find it useful to administer an instrument some months after the program has been completed. Trainees with prior work experience will have had time to determine whether their experiences in the program have resulted in gains in work proficiency. How long an interval is a matter of judgment; it must be long enough so that a good estimate of the value of the skills resulting from the training can be made and not so long that the details of the training experience have been forgotten. Some investigators like an interval of six to twelve months.

An instrument used at this time can produce important information on coverage, both for trainees without prior experience on the job as well as for trainees with considerable job experience. The instrument can elicit comments on omissions in the program or areas that should have been covered or that should have had greater emphasis in order to be most useful.

With this type of instrument, the respondents can be asked specifically what they are able to do better as a result of the training. Specific examples are exceedingly useful, and sometimes estimates of the monetary value to the organization can be obtained.

An advantage of the delayed instrument over the one administered during the final session of the program is that a lapse of time may moderate negative reactions. An individual who at the end of the program may have expressed annoyance with a trainer because of what seemed to be excessive demands may have come to realize that high standards of trainee performance actually produce lasting gains.

The following figure presents a suggested format for the questionnaire to be administered after experience on the job, including two questions that are applicable if goal setting is part of the program.

ORAL PRESENTATION OF RESULTS

Meeting with Managers

The results of training research—whether the findings are positive or negative—are reported to those who need to know how the study has turned out in order to use it as a basis for making decisions. In the business world, training research is supported by management, which needs to know whether the particular program involved is meeting the needs of the organization and can be considered a good investment.

Suggested Format for Questionnaire

To Be Administered Some Months After Training[2]

Instructions: Sometime ago you participated in (give title and place of program). The organizers of this program are interested in your present evaluation of the program and the degree to which it has been helpful to you in your work. Please print all comments.

1. Read the four responses listed below and check the box next to the response that best indicates how useful the program turned out to be in terms of helping you in your work:

   □ Indispensable.

   □ Valuable but not indispensable.

   □ Somewhat useful.

   □ Of no value.

2. The most useful aspect was____

   ____

3. The part that was least useful or least satisfactory was____________________________________

____

4. During the training did you establish a goal to achieve at work? Check the appropriate box.

   □ Yes

   □ No

5. If you answered yes, what was the goal, how did you attempt to achieve it, and what were the results?

____

____

____

____

____

6. If you achieved your goal satisfactorily, try to estimate the dollar value of your success to the organization over a period of one year.__________

7. What suggestions do you have for improving this training?

____

____

8. List topics or areas that should have been included or emphasized more strongly so that the program would have had greater value for you.

____

____

9. What topics might be dropped or given less emphasis?

____

____

[2] The authors thank Drs. Robert E. Lefton and V. Ralph Buzzotta of Psychological Associates, Inc., St. Louis, Missouri, for permission to use this adaptation of their questionnaire.

Although a formal written report is often of considerable interest and constitutes a permanent record of the study, managers frequently request an oral presentation. Visual aids, such as charts that can be displayed on an easel or shown on a screen, are useful; but the great advantage of an oral presentation is the give-and-take involved. The researchers can take measures to ensure that the importance of the project and the meaning of the results are understood. The managers can take steps to verify the applicability of the findings to the work situation. The managers probably know in a general way what has been happening, but the research staff should prepare the presentation in sufficient detail that answers will be available for all questions that may come up and that important findings will be clearly presented. Evidence includes results from any experimental study as well as summaries from instruments. The degree to which trainees have found the program valuable is generally of interest to managers, as are reports of incidents that demonstrate the economic usefulness of the training.

During such a presentation, a discussion in which everyone participates is useful. Often the future of a program will be decided on the basis of the interaction of the managers and the researchers. Answers to a number of questions need to be established:

1. Is the program effective in its assigned function?
2. Are there better options?
3. Are improvements needed? Are there portions of the program that should be modified, dropped, or replaced? Is the coverage adequate?
4. As far as costs are concerned, is the training as efficient as can reasonably be expected?
5. In terms of money invested, what is the dollar yield of the program?

These questions cannot be answered in a hurried atmosphere. Prudence requires that all pertinent facts be taken into consideration and that a generally satisfactory approach be adopted. Possible actions include the following:

1. Continuation of the present program with more evidence to be collected later;
2. Modification of the current program on the basis of available evidence;
3. Consideration of a radical change in the training program; and
4. Acceptance of the current program with no further study at the present time.

Meetings with Trainers and Training Designers

It is also useful to deliver an oral presentation to the training staff as well as to the designers of the program (if the trainers were not the designers). Although the outcome of an experimental study would be of interest to both trainers and designers, findings from instrumentation merit particular scrutiny because they cover topics such as the following:

1. How the trainer or trainers were perceived during the program. Appraisal is on an individual basis, and sometimes corrective action is indicated.

2. Annoyances resulting from correctable conditions, such as poor lighting, poor audibility, bad ventilation, improper heating or cooling, and interruptions. Although trainers should detect undesirable situations without the use of an instrument, it often happens that such conditions are missed until reported formally.
3. Acceptability of identified training aids and special training techniques.
4. Perceived redundancies and omissions in the program.
5. The degree of enthusiasm of the trainees, as evidenced by their expressed willingness to recommend the training to others and their perception of the program as a good investment on the part of the sponsors.

The decisions that can be made in such meetings depend on how the program is organized. Modifications by training designers are subject to review by the managers to whom they report, regardless of whether the designers are employees of the organization or external contractors. Because trainers are directly involved in training operations, their recommendations are especially important; nevertheless, these recommendations are subject to the approval of the trainers' supervisors.

WRITTEN PRESENTATION OF RESULTS

A formal written report of a project has several uses:

1. It is a permanent record of what has been done and what has been discovered.
2. It generally includes recommendations for action.
3. It can be circulated to all individuals with an interest in the training of the particular target group or the overall training program of the organization.
4. It serves to familiarize a wide audience with the usefulness of systematic research in training. This can lead to an expanded role of the research staff and to improvement in operations.

Sometimes written reports are merely filed; but when constructive recommendations seem likely to enhance operations, the research staff can often take steps to encourage the implementation of these recommendations. It is important to perceive training research as merely one phase of improving ongoing operations.

PUBLICATION IN PROFESSIONAL JOURNALS

Occasionally an evaluative study is reported in the professional literature. Other studies, some of which are just as well executed, are recorded only in organizational files. There are three categories of professional publications in which a report on the results of a training investigation might be published:

1. A journal addressed to the managers in the industry;
2. A journal read by the growing group of full-time, professional trainers; and

3. A journal in educational or industrial and organizational psychology.

There are reasons to publish as well as reasons not to publish. Certainly the professional worker in training evaluation obtains much of his or her knowledge from publications. Thus, the worker has an obligation to share results with other professionals so that the field can continue to grow. It is only through sharing that this can happen.

A first question to ask is whether the study can be written in such a fashion that it makes a genuine contribution to knowledge. If the answer is yes, publication certainly should be considered.

Some organizations may believe that any discoveries resulting from sponsored research should be considered confidential so as not to give an advantage to the competition. In most cases, this is not a valid argument because it would be almost impossible to hide a real secret in the field of training. Any innovation is likely to be known by a wide circle of people. What a research study uncovers is usually a new field of application, a somewhat different approach, or—occasionally—a new method.

Publication can result in recognition by colleagues in other organizations and may encourage communication with them. It is an important means of furthering professional development and, ultimately, of improving one’s own training program.

GRAPHICS

Both in written reports and in oral presentations, graphs and diagrams are often helpful. Recommendations regarding the preparation of graphic aids are as follows:

1. Because their purpose is to clarify, graphs should be simple and easy to read.
2. Graphs should focus on important information, but not too much information should be included in any single display.
3. In a training study, a graph is often designed to show the relationship between an independent variable, such as the use or lack of use of a particular training aid, and a dependent variable, such as an achievement score or success on the job. Both variables should be clearly indicated. When the dependent variable is in units, the scale should be shown.
4. The caption should be short and informative and should present the relationship or distribution displayed as well as the source of the data. Somewhere in the diagram—in the caption or elsewhere—the number of trainees participating in the study should be shown. When a measure of relationship, such as a correlation coefficient, has been computed for a relationship that has been graphed, it is appropriate to show it.
5. Different colors or cross-hatching may be used to indicate the different groups involved or the different training approaches that have been dealt with. In such cases, a key must be provided.

As one is preparing the findings of a study for presentation, one should consider ways in which graphs might be useful. Facts that can be expressed in percentages can be graphed, either in bar-chart form or, when the percentages add up to 100 percent, as a pie chart. Experimental studies of learning often yield a learning curve, in which the amount learned is in units on the vertical axis and trials or time on the horizontal axis. It may require creative thinking to develop a graph that successfully summarizes an important finding.

Both graphs and descriptive statistics represent summaries of data. In all cases, a numerical organization of the information (often in terms of frequencies and percentages or means and perhaps standard deviations) precedes the development of the visual presentation. Although graphic methods can be used to show fairly complicated findings, their chief use in training research is to display distributions and relationships between variables: an experimental or independent variable on the one hand and the dependent variable on the other.

STATISTICS

In a program of training research, statistics are important in describing the reliability and validity of measuring instruments, in determining the degree to which confidence can be placed in an obtained difference between means, and in controlling the effects of variables not of direct interest. In addition to control methods that include matching and random assignment to groups, there are statistical techniques designed to accomplish much the same result through the subtraction of unwanted variance from statistics indicating relationship. Although techniques of multivariate analysis of variance and the numerous varieties of multivariate correlation are beyond the scope of this discussion, a statistical consultant may occasionally suggest a multivariate method useful to incorporate in the design of a specific training study, especially when control of unwanted variance by a nonstatistical method is impractical. In a complete report of a project, statistical results have their place as intermediate findings that bolster conclusions—the real end products of a study.

REFERENCES AND BIBLIOGRAPHY

Alreck, P.L., & Settle, R.B. (1986, January). The survey research handbook. *Journal of the Market Research Society, 94.*

American Psychological Association. (1982). *Ethical principles in the conduct of research with human participants.* Washington, DC: Author.

The *Annual* series for HRD practitioners. (1972-1994). J.W. Pfeiffer, J.E. Jones, & L.D. Goodstein (Eds.). San Diego, CA: Pfeiffer & Company.

Bowers, D.G., & Franklin, J.L. (1974). Basic concepts of survey feedback. In J.W. Pfeiffer & J.E. Jones (Eds.), *The 1974 annual handbook for group facilitators.* San Diego, CA: Pfeiffer & Company.

Bowers, D.C., & Franklin, J.L. (1977). *Survey-guided development I: Data-based organizational change.* San Diego, CA: Pfeiffer & Company.

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

Cook, J.D., Hepworth, S.J., Wall, T.D., & Warr, P.B. (1981). *The experience of work: A compendium and review of 249 measures and their use.* Orlando, FL: Academic Press.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Downie, N.M., & Heath, R.W. (1970). *Basic statistical methods* (3rd ed.). New York: Harper & Row.

Dunham, R.B., & Smith, F.J. (1979). *Organizational surveys: An internal assessment of organizational health.* Glenview, IL: Scott, Foresman.

Farrow, D.L., & Sample, J.A. (1994). BARS: Developing behaviorally anchored rating scales. In J.W. Pfeiffer (Ed.), *Pfeiffer & Company Library, 18* (pp. 113-119). San Diego, CA: Pfeiffer & Company.

Farrow, D.L., & Sample, J.A. (1986). BARS: Developing behaviorally anchored rating scales. In J.W. Pfeiffer & L.D. Goodstein (Eds.), *The 1986 annual: Developing human resources.* San Diego, CA: Pfeiffer & Company.

Fowler, F.J., Jr. (1984). *Survey research methods.* Beverly Hills, CA: Sage.

Frame, R.M., Hess, R.K., & Nielsen, W.R. (1982). Survey-guided development. In *The OD source book: A practitioner's guide.* San Diego, CA: Pfeiffer & Company.

Franklin, J.L., Wissler, A.L., & Spencer, G.J. (1977). *Survey-guided development III: A manual for concepts training.* San Diego, CA: Pfeiffer & Company.

Gavin, J.F. (1984). Survey feedback: The perspectives of science and practice. *Group & Organization Studies,* 9(1), 29-70.

Hanson, P.G. (1981). *Learning through groups: A trainer's basic guide.* San Diego, CA: Pfeiffer & Company.

Harrison, R. (1970). Choosing the depth of organizational intervention. *Journal of Applied Behavioral Science*, 6, 181-202.

Hausser, D.L., Pecorella, P.A., & Wissler, A.L. (1977). *Survey-guided development II: A manual for consultants.* San Diego, CA: Pfeiffer & Company.

Hersey, P., & Blanchard, K.H. (1973). *LEAD self (other): Leader effectiveness & adaptability description.* Escondido, CA: Leadership Studies, Inc.

Hillway, T. (1956). *Introduction to research.* Boston, MA: Riverside Press Cambridge.

Hogarth, R.M. (Ed). (1982). *Question framing and response consistency.* San Francisco, CA: Jossey Bass.

Kesselman-Turkel, J., & Peterson, F. (1982). *Research shortcuts.* Chicago: Contemporary Books.

Labaw, P. (1980). *Advanced questionnaire design.* Cambridge, MA: Abt Books.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology,* 140, 1-55.

Likert, R. (1971). *The human organization.* New York: McGraw-Hill.

Lord, F.M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin,* 72, 336-337.

Mahler, W.R. (1974). *Diagnostic studies.* Reading, MA: Addison-Wesley.

Maiorca, J.J. (1991). Basic statistics for the HRD practitioner. In J.W. Pfeiffer (Ed.), *The 1991 annual: Developing human resources.* San Diego, CA: Pfeiffer & Company.

Mayo, G.D., & DuBois, P.H. (1987). *The complete book of training: Theory, principles, and techniques. San* Diego, CA: Pfeiffer & Company.

Morgan, J.P., Jr. (1981). Central West Virginia cultural-awareness quiz. In L. Thayer (Ed.), *50 strategies for experiential learning: Book two.* San Diego, CA: Pfeiffer & Company.

Mouton, J.S., & Blake, R.R. (1975). *Instrumented team learning: A behavioral approach to student-centered learning.* Austin, TX: Scientific Methods.

Nadler, D.A. (1977). *Feedback and organization development: Using data-based methods.* Reading, MA: Addison-Wesley.

Oppenheim, A.N. (1966). *Questionnaire design and attitude measurement.* New York: Basic Books.

Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning.* Urbana, IL: University of Illinois Press.

Patton, M.Q. (1980). *Qualitative evaluation methods.* Beverly Hills, CA: Sage.

Peters, D. (1985). *Directory of human resource development instrumentation.* San Diego, CA: Pfeiffer & Company.

Pfeiffer, J.W., Heslin, R., & Jones, J.E. (1976). *Instrumentation in human relations training: A guide to 92 instruments with wide application to the behavioral sciences* (2nd ed.). San Diego, CA: Pfeiffer & Company.

Rao, T.V. (1985). The entrepreneurial orientation inventory: Measuring the locus of control. In L.D. Goodstein & J.W. Pfeiffer (Eds.), *The 1985 annual: Developing human resources.* San Diego, CA: Pfeiffer & Company.

Rotter, J.B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs,* 80 (1, Whole No. 609).

Sample, J.A. (1986). The use of behaviorally based scales in performance appraisal. In J.W. Pfeiffer & L.D. Goodstein (Eds.), *The 1986 annual: Developing human resources.* San Diego, CA: Pfeiffer & Company.

Sample, J.A. (1994). The use of behaviorally based scales in performance appraisal. In J.W. Pfeiffer (Ed.), *Pfeiffer & Company Library, 20* (pp. 167-178). San Diego, DA: Pfeiffer & Company.

Schostrom, E.L. (1974). *Personal orientation inventory.* San Diego, CA: Educational and Industrial Testing Service (EDITS).

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments in question form, wording and context.* New York: Academic Press.

Schutz, W. (1977). *FIRO awareness scales manual.* Palo Alto, CA: Consulting Psychologists Press.

Schutz, W. (1982). *The Schutz measures: An integrated system for assessing elements of awareness.* San Diego, CA: Pfeiffer & Company.

Sudman, S., & Bradburn, N.M. (1982). *Asking questions. A practical guide to questionnaire design.* San Francisco, CA: Jossey-Bass.

Supervisory Attitudes: The X-Y Scale. (1972). In J.W. Pfeiffer & J.E. Jones (Eds.), *The 1972 annual handbook for group facilitators.* San Diego, CA: Pfeiffer & Company.

Supervisory Attitudes: The X-Y Scale. (1994). In J.W. Pfeiffer (Ed.), *Pfeiffer & Company Library 19,* (pp. 187-192). San Diego, CA: Pfeiffer & Company.

The Survey of Organizations (SOO). (1980). Ann Arbor, MI: Institute for Social Research, The University of Michigan, and Rensis Likert Associates, Inc.

Taylor, J.C., & Bowers, D.G. (1972). *The Survey of Organizations: A machine-scored standardized questionnaire instrument.* Ann Arbor; MI: Institute for Social Research.

Terborg, J.R. (1979). Women as managers scale (WAMS). In J.E. Jones & J.W. Pfeiffer (Eds.), *The 1979 annual handbook for group facilitators.* San Diego, CA: Pfeiffer & Company.

Van de Ven, A.H., & Ferry, D.L. (1980). *Measuring and assessing organizations.* New York: John Wiley.