CHAPTER THREE:

MEASUREMENT APPROACHES

TOPICS **Page**

Overview of Performance Measures 3-3

Program Records 3-4

Records From Other Agencies 3-4

Questionnaires 3-5

Tests 3-8

Interviews 3-9

Focus Group Interviews 3-10

Observations 3-11

Goal Attainment Scaling and Portfolios 3-12

Case Studies and Other 3-13

Combining Measurement Strategies 3-14

Selecting a Data Collection Method 3-15

Finding A "Do-able" Method For The Information You Need 3-15

Choosing An Appropriate Method 3-17

Reliability 3-18

Validity 3-20

Communicating With Your Stakeholders 3-22

Chapter Three:

MEASUREMENT APPROACHES

Once you have a logic model to specify your program outputs and outcomes, it is time to look at methods for measuring progress toward your goals and objectives. Performance measures provide evidence of inputs, outputs, and outcomes.

OVERVIEW OF PERFORMANCE MEASURES

Records	***Interviews***	***Goal Attainment Scaling***
Questionnaires	***Focus Groups***	***Case Studies***
Scales	***Observations***	***Journals, Logs, and Diaries***
Tests	***Portfolios***	***Videos and Photographs***

Many measurement approaches can be applied to assessing outcomes at community levels as well as in individual programs or initiatives. For example, records from the most common source of data on Oregon's Benchmarks. Records are also invaluable in assessing outcomes for individual programs.

> ***NOTE***
>
> *Similar measurement approaches can be applied to assessing outcomes at community levels as well as in individual programs or initiatives.*

Similarly, surveys, such as the Search Institute Profiles of Student Life and the Oregon Public School Drug Use Survey can be adapted to describe community wide conditions *or* outcomes of specific intervention or prevention efforts.

Whether you want to assess community wide conditions or the outcomes of an individual program, the most work in performance assessment usually centers on data collection. It is critical to carefully choose the measurement approaches that fit your needs and resources. Because the range of performance measures is enormous, the next sections of this chapter will review the major types of performance measures and the information that these various methods, techniques, and tools can supply.

Program Records

Program records can provide a wealth of information for describing your outputs: what you actually did, how much of it, and with whom. Records have the double advantage of being accessible and available.

MEASURE OUTPUTS FROM PROGRAM RECORDS

How many individuals participated?

What services did participants receive?

How long did individuals participate?

What activities were conducted?

Planning is the first step. The value of your records will depend on whether you are collecting appropriate information to measure the outputs and outcomes you seek. Also, program records must be kept carefully and thoroughly if they are to be useful.

If current records do not provide needed information, revise them to better monitor outputs and outcomes in the future. Involving program staff in decisions about record-keeping is crucial. Staff will offer valuable insights, and their involvement in planning will make it easier to get their cooperation in actual record-keeping.

With forethought and careful planning, program records can also furnish information on participant outcomes such as:

- number of children who are fully immunized by two years;
- proportion of child care providers using developmentally appropriate practices; and
- school attendance of youth in a mentoring program.

Records From Other Agencies

Records from other agencies and institutions offer a different perspective on the effects of your program. For example, crime records, showing the number of first-time offenders from a particular school district before and after the establishment of a recreation program might show the effectiveness of after-school activities. Similarly, records of telephone calls for information on obtaining quality childcare before and after a media campaign could show the effectiveness of the campaign in increasing parent awareness of quality elements.

Records provide a ready source of information and may be utilized with less effort than other data collection methods. Many types of records are in the public domain and are published by many government agencies in aggregate form.

More and more public records are available over the internet on a "web page." For example, the Oregon Health Division's Center for Health Statistics has a web page at http://www.ohd.hr.state.or.us that shows statistics for a variety of health outcomes. One table describes the annual number of mothers receiving first trimester prenatal care in each Oregon county. Another table shows the annual number of teen mothers by county.

> ***KEEP IN MIND***...
>
> Public records may not be in a form that is useful to you.
> Confidentiality issues may limit your access to certain types of information.

Public records are a ready source of information but they may not be kept in a form that is useful to you or appropriate for your program. For instance, records may not cover the period of time a program has operated, or it may be impossible to identify program participants in aggregate records. Certain types of records, such as child maltreatment reports, may not be available due to issues of privacy and confidentiality.

Programs can use *Record Extraction Forms* to collect information from their own written records or from the records of other agencies. See the appendix for samples and further information on developing record extraction forms.

Questionnaires

Questionnaires are one of the most commonly used survey methods. Typically used for self-assessment, questionnaires require answers to written questions. Data may be collected in a variety of ways:

- mail
- telephone
- in person, face-to-face
- groups of individuals

Because questionnaires can be mailed or handed out to people, you can cover a wide geographical area or contact a large number of people fairly easily. But mailing lists must be up-to-date since people move frequently.

If questionnaires are not "user-friendly" the numbers of people who actually respond are likely to be very low. Providing a self-addressed, stamped envelope increases the likelihood that a mailed questionnaire will be returned. See Chapter 4 for some guidelines on creating effective questionnaires.

Questionnaires use either fixed choice or free response question formats. Each format has particular advantages and disadvantages.

QUESTION TYPES

FIXED CHOICE

Overall, would you say your child is in good health?

___ Yes ___ Not sure ___ No

Overall, how do you rate your child's health?

1 Poor
2 Fair
3 Good
4 Very good
5 Excellent

FREE RESPONSE

How would you describe your child's health over the past 3 months?

_____

_____

Fixed-choice questions. Single, direct questions with preset answer choices are termed closed or *fixed-choice*. These questions require people to choose among offered alternatives instead of allowing them to respond in their own words.

Fixed choice questions are quick to answer and simple to score, but it is essential that the choices fully reflect most of the responses that participants might actually have.

The simplest fixed-choice question offers two possible responses, most commonly either *yes/no* or *true/false*. Questions that include a third category for *don't know*, or *not sure* tend to yield more accurate results because people are not forced to choose an answer when they aren't sure.

Other fixed choices involve rating or ranking. The choices should cover the *full range* of possible responses. For example, on a question about the impact of a program on family communication, response choices should include the negative impacts (things got worse), neutral impact (nothing changed) as well as positive impacts (things got better). Also each choice should be *mutually exclusive* of all others, that is, response choices should not overlap. (Chapter 4 provides more information on these issues).

Rating and ranking are common questionnaire assessment techniques that allow a number to be obtained for everything that is judged:

- *Rating techniques* ask the individual to make a judgment and place an answer on a continuum that is negative at one end and positive at the other. A number is then assigned to each point on the continuum.

- Asking a parent to rank four discipline approaches from one to four in *order* of preference is an example of *ranking*.

Rating scales can vary in length from two choices such as *yes/no* or *true/false* up to ten or more. Survey researchers generally advise 4- or 5-points for self-administered questionnaires.

Free-response questions. Other survey questions may be presented without a set of possible answers. Individuals supply their own answers to these open-ended questions:

> "What is the most important thing you've learned from this program?"
>
> "How do you discipline your child when s/he misbehaves?"
>
> "Is there anything else you want to tell us?"

Free-response questions tend to provide better measurement of sensitive or disapproved behavior. For example, asking a parent "When are you most likely to use physical punishment?" will give you more complete information than asking a yes/no question such as "Do you use physical punishment when disciplining your child?"

On the other hand, on written questionnaires, free response questions demand more time and a higher educational level than most fixed choice questions. The results also take longer to code and analyze.

Scales are a particular form of a questionnaire where a series of three or more questions are asked about the same issue. Scales have statistical information to show that the individual questions are closely related, giving the scale internal consistency.

Thus, scales offer a more reliable way to measure a person's attitudes and ideas about an issue than simply depending on a single question.

SCALE

Set of three or more questions about the same topic.

Statistical data show items have internal consistency.

Standardized instruments may have several sub-scales, each measuring a different aspect of the particular issue. For example, the Adult-Adolescent Parenting Inventory (see Chapter 5, Section 6) measures parenting attitudes and includes four sub-scales:

- empathy toward child
- unrealistic expectations for child's development
- role reversal (parent expects child to take care of parent)
- attitudes toward corporal punishment.

Tests

Tests measure abilities and knowledge. A wide variety of standardized tests can be purchased from publishers. *Standardized tests*, such as achievement tests or intelligence tests, have been developed over a period of years and pre-tested on large samples of people. These tests have specific requirements for their administration and scoring.

NORM-REFERENCED TEST

Measure of mastery or competency relative to the performance of others on the same test.

Most standardized tests generate *norm-referenced scores* that compare the test-taker's ability to a norm group tested during the standardization process. These comparisons often focus on specifically defined populations such as 11th grade students or children whose primary language is English.

When you interpret norm-referenced test scores, you compare scores to those for the norm group. Therefore, if you use a standardized test, be sure that the norm group is similar to your population. This is an especially important consideration when second languages and varying socioeconomic and sub-cultural patterns are involved.

Performance tests are often *criterion-referenced*. A mastery level is defined and individual scores are interpreted as below or above that criterion. Criterion-referenced tests are useful when the evaluation question focuses on whether the participant met minimum standards of performance. For example, criterion-referenced tests are used to screen children for age-appropriate development. These tests show whether development falls below or within the normal range for similar-aged children.

CRITERION-REFERENCED TEST

Measure of individual mastery or competence relative to the content of the test.

Tests can be devised to assess specific information or competencies gained through program participation such as knowledge of healthy eating practices or conflict resolution skills. The comparison will be the individual's knowledge or skill level before and after your intervention or you may wish to define criteria for minimum performance levels.

Tests are validated by asking experts in the field to review the items or questions and to determine if they are representative of the knowledge or skills you are measuring. By doing this, you establish *content validity* for your test.

Interviews

Interviews are carefully planned "spoken" questionnaires that are designed around explicit goals. Interviews are conducted in person or over the telephone, with one individual speaking to another.
Formally *structured* interviews use questions that are direct with clarification and elaboration allowed only within specific limits. More typically, interviews are *semi-structured* around a core of specific questions with follow-up probes that branch off to explore issues and responses in greater depth.

Open-ended questions often require probing. A skillful interviewer can listen carefully to a participant's answer, and ask follow-up questions to draw out feelings and opinions.

An interview question such as "What approach works best for you when your children misbehave?" could be followed up with probes to gain factual information about the frequency and use of various discipline techniques in that family. Probes should be neutral, offering no additional information to an already asked question.

NEUTRAL PROBING DRAWS OUT MORE COMPLETE ANSWERS

"I'm not sure exactly what you mean."

"Can you tell me more about that?"

"What happened next?"

"Why do you think that happened?"

Interviews are particularly useful when you are seeking in-depth information about an individual family or when you are dealing with sensitive areas, such as child maltreatment or substance abuse. However, interviewers require substantial training to develop skill in probing at significant points and avoidance of biasing tendencies. Interviews are also labor intensive, both to conduct and to analyze the results.

Interview data is typically analyzed through content analysis. Readers review interview protocols, looking for themes or categories in the responses. Responses within themes can be counted to quantify data. Interviews can provide a rich source of interesting and important results.

The success of interviews depends on the ability of the interviewer to develop rapport with the person being interviewed. Appearance and opening conversations create the critical initial impression. Personality also plays a role. Successful interviewers are warm, empathetic, and genuinely interested in the individual's responses.

SUCCESSFUL INTERVIEWERS carefully prepare and rehearse in advance:

- *Actions*
- *Lines*
- *Roles*
- *Routines*

Focus Group Interviews

Focus groups are a special type of interview involving planned discussion groups. Approximately 7-10 individuals, led by a trained interviewer, share their ideas and perceptions on a particular topic or area of interest. The discussion is recorded and the content reviewed at a later date.

USING A FOCUS GROUP

What do potential participants think about a new proposal or program?

How well is the current program working?

What are the strengths and weaknesses of the current program?

For many years, focus groups have been the mainstay of marketing research. More recently, focus group interviews are being used by non-profit organizations for needs assessment, program evaluation and to look at changes across systems (Krueger, 1988).

Focus groups are particularly useful when the area of interest is sensitive. For example, youth might be asked to participate in a focus group interview to assess the availability of drugs in a neighborhood or opportunities for after-school involvement in extra-curricular activities.

Focus groups are generally *not* a good source for outcome information for several reasons. They tend to involve only a small, non-representative sample of the program participants. Individual responses may be biased by the comments of other members in the group. Group discussions rarely yield the kind of individual-level information that can be tabulated and reported as outcomes. Adding a questionnaire to be completed by individuals either before or after the focus group session can provide outcome information.

Focus group interviews can provide insights about the meaning and interpretation of survey findings. Issues, concerns, and ideas overlooked in questionnaires are likely to come up in a group discussion.

Analysis of the focus group information centers on identifying patterns and trends that arose in the sessions. Analysis can be conducted for both individual focus groups and across several groups.

For further information and directions for conducting focus groups, see the ***Focus Group Interviews***, a publication from the Ohio State University Cooperative Extension Service, which is reproduced in the appendix.

In addition, several guidebooks to focus groups interviews are available. One of the most readable and applied is *Focus Groups: A Practical Guide* by Richard Krueger. 1990. Sage Publishers, Newbury Park, California.

Observations

Participant behavior and performance, environments, and events can be systematically observed and recorded by trained observers to provide descriptive or evaluative information. Observations are a good alternative or supplement to self-report information on behavioral skills and practices.

OBSERVATIONS
are useful for outcomes relating to behavior, facilities, environments

- Use of verbal instead of physical means to resolve conflicts.
- Developmentally appropriate qualities of a child care facility.
- Condition of neighborhood parks and play areas before and after a local clean-up campaign.

Observations, kept in the form of field notes or anecdotal records, can provide a rich source of descriptive information about events and behavior. Records of this type can be particularly useful if you are tracking system change activities and outcomes. For example, factual and detailed observations of meetings and participant responses can show the extent to which a collaborative group has improved decision-making processes or mobilized group resources.

Pre-determined guides (such as checklists or rating scales) can further focus information gathering on the outcomes being measured and assist in quantifying the data:

- Checklists show the presence or absence of a specific characteristic.
- Rating scales provide specific criteria for categorizing the observation.

TRAINED OBSERVER RATINGS

more objective than self-reports

more consistent than casual observations

require careful training and skills

improve with clear rating scales

Specific training in using rating scale criteria is necessary to avoid misinterpretation and insure that the criteria are being consistently applied. The value of observed information depends on the training and skill of the observers, and the clarity of the rating scale.

For guidelines on observations and training of raters, and discussion of raters, see ***Observation Rating Scales*** in the appendix for information on developing rating systems and training observers.

Goal Attainment Scaling

With an extensive history in human service contexts, goal attainment scaling is well suited to documenting outcomes. Goal attainment scaling is especially useful in programs where families select their own goals or when goals are highly idiosyncratic such as in mental health counseling.

FAMILY GOAL ATTAINMENT

Families and staff work together to identify *behavioral examples* of:

5 best anticipated outcome

4 more than expected success

3 expected success

2 less than expected success

1 most unfavorable outcome

In goal attainment scaling, families or other participants identify areas for action with assistance from program staff. Realistic goals are established and given a specified time period. The initial level provides a baseline that can be compared to succeeding levels in order to measure progress.

Each goal is weighted in terms of its overall importance to the family. Weighting places higher value on concerns that the family ranks as most significant and thus, provides a more accurate estimate of overall goal attainment.

Because goal attainment scaling involves the individual or the family in the process, it is one of the least obtrusive ways to collect information about outcomes. (See Chapter 5, Table 5-1, for a further description of goal attainment scaling).

Portfolios

Portfolios are powerful tools for learning, assessment, and self-discovery. In portfolio assessment:

- levels of mastery are clearly identified for a given set of skills
- evidence is gathered together in a portfolio to show developing competencies

PORTFOLIOS

- *collect information*
- *show mastery of skills*
- *compare skills to a set of standards*

The validity of portfolio assessment depends on having *clearly defined levels of mastery* available for judging progress.

Portfolio assessments are widely used in schools where, for example, children's work is collected to show the acquisition of language arts and mathematics skills. Rather than being sent home,

papers are placed in a portfolio.

By reviewing the portfolio contents, progress can be assessed relative to defined levels of mastery. Children can even evaluate their own progress and receive immediate feedback on their developing skills. Progress can be quantified as with goal attainment scaling.

Portfolio assessment can be helpful when individuals or programs wish to show they meet certain criteria. For example, child care centers who seek national accreditation, collect a portfolio of information and examples of practices in a number of specified areas to demonstrate competency levels.

Case Studies

Using a combination of assessment methods, case studies describe an individual case or multiple cases in intensive detail. Case studies are often used to examine:

- groups of people or organizations
- key decisions
- public programs
- changes within an organizational system or community

As a general rule, case study methodology is most appropriate when you seek to explain *how* and *why* events occur. Case studies use multiple sources of evidence, such as interviews, observations, records, and other types of documentation, to zero in on the "facts of the case."

An effective case study clearly defines the purpose at the outset and limits data collection to the evaluation questions under consideration. Otherwise, data collection can result in an overwhelming amount of descriptive information that is difficult to interpret and analyze (Yin, 1998).

Other

A variety of other methods can be used to show change over time. Some of these are as follows:

- *Journals*, *logs*, and *diaries* can record feelings and actions related to program activities.

- Individuals can make *testimonials* by describing personal experience in narrative fashion.

- *Anecdotes* in the form of "success stories" can be used to illustrate program outcomes.

- *Photographs* and *videos* depict program activities and or demonstrate change in behaviors.

Combining Measurement Strategies

Often, programs will find it useful, or necessary, to rely on several different strategies to get the information that is needed. For example, to measure outcomes, Oregon Healthy Start relies on several data sources including health records for immunization data, parent surveys, staff observation of home environments, and parents completion of a standardized developmental screening measure. These quantitative data are supplemented by anecdotal stories and case histories.

CONSIDER USING MULTIPLE MEASURES WHEN

- *there are multiple outcomes to assess,*
- *no one measure meets the needs of all participants,*
- *resources exist to track multiple measures.*

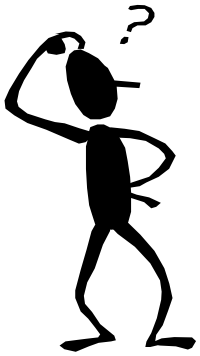
Not all programs have the resources to conduct multiple assessments but it is very important to consider all the ways to learn about a program and its outcomes. Pick those methods that best fit the outcomes to be assessed, the participants, and the staff and other resources of the program.

REMEMBER

A little accurate, valid information is better than a lot of poor and incomplete information.

SELECTING A DATA COLLECTION METHOD

Making a decision on which method (or methods) to use depends on your individual situation, the particular outputs and outcomes you are measuring, and how much time and resources you have to commit to data collection. Consider the following questions:



- Will the approach give you (and the intended users) the information you need?
- Given your available resources and time, is the method "do-able?"
- Will the method be appropriate for the participants in your program?
- Are the measures and procedures reliable and valid?

Finding A "Do-able" Method For The Information You Need

What do you want to find out? Focus your data collection on the information that will provide *credible* answers to your evaluation questions. These questions re-state the outputs and outcomes your program seeks to achieve.

Suppose you wanted to know: "Have conflict resolution skills improved among students in the mentoring program?" Begin by brainstorming all the methods that could provide you with that information.

All of these techniques *could* provide you with the information you want. But each method has different requirements in terms of the time and resources needed for data collection and analysis.

What information will satisfy your stakeholders? Remember that stakeholders include not only your funders, but also staff, administration, and the participants themselves.

Have conflict resolution skills improved among students in the mentoring program?

- *Ask students to rate their proficiency in using conflict resolution skills*
- *Send staff a questionnaire about students' conflict resolution skills*
- *Interview students for their opinions about their progress*
- *Give parents a rating scale to assess their children's conflict resolution skills*
- *Observe students as they role-play conflict situations, and rate their skills*
- *Have students keep a journal of their progress, including a checklist of skills they have learned*
- *Ask the teachers how often they resolve peer-related conflicts*
- *Check school records for disciplinary incidents or classroom disturbances*

Data collection takes time, effort and financial resources. Unless you have funding available to pay outside data collectors, time and effort will come from program staff and participants and dollars from the program budget.

Your task is to balance what is necessary to satisfy your stakeholders with the time and resources you can devote to data collection. Some methods are less expensive and time consuming than others.

- Questionnaires are generally inexpensive to produce and easy to use when you are working with a small group of people. Community-wide surveys tend to be more costly.
- Observations and individual interviews are the most time-intensive strategies.
- Focus groups take less time than individual interviews and yield much information about program implementation but less about individual outcomes.

BE SELECTIVE
BE REALISTIC

Can the method be built into your current program activities?

Can staff collect information without detracting from service?

What resources do you have for expenses?

- Goal attainment scaling fits naturally within a service model focusing on goal-oriented outcomes, such as are contained in individualized treatment or family support plans.
- Program records can provide an inexpensive and accessible source of information on activities, outputs, and outcomes.

OH NO.
Not more paperwork!!!!

What about paperwork? Data collection can add undue paperwork and detract from the time staff have to devote to the families they serve. Include staff in the decision making process as much as possible to choose an approach that requires the minimum amount of paperwork necessary to measure the outcomes you have selected.

Participating in data collection benefits staff in different ways. Discussing the feasibility of various approaches can clarify how a program is currently being implemented. Attention is focused on the outcomes that are expected from the intervention.

Staff gain a better understanding of the program participants' experience, skills, strengths and needs through the data collection process. This

information often proves to be invaluable in working with the individuals and understanding the dynamics of the program.

Choosing An Appropriate Method

Be sure the method you choose is appropriate for your population. Considerations include the following:

Language level You want participants to understand the questions you ask. Language should be clear and free from technical jargon. Use everyday language as much as possible.

Instruments must be translated if participants are not comfortable with English. Most Latino families speak Spanish but may use differing dialects. Translations should be reviewed by people familiar with the language, and dialect, of the group you are surveying.

Reading level In almost every group, some people will read very well while others will read poorly. Most instruments strive for a 4^{th} to 5^{th} grade reading level.

Reading level depends on the number of words in a sentence, the length of individual words, and the grammatical complexity of the sentence. Word processing programs can compute readability statistics.

Norm groups Standardized tests are developed by having different groups of people take them and then comparing the norms. Many standardized tests were developed for White middle-class populations. Be sure to consider the applicability of an instrument for the people your program serves. If necessary, find a different approach or carefully adapt an instrument to your participants.

Cultural sensitivity Assessment is made culturally sensitive by carefully meshing and adapting all aspects of the process with the cultural characteristics of the group being studied (Suzuki, Meller, & Ponterotto 1996). To reduce potentially biasing effects, ask a member of the cultural community being assessed to review the:

- development or adaptation of instruments, including translation;
- administration of the measures; and
- interpretation of the findings.

Building trust in the purpose of assessment and the benefits to the community will increase the success of your assessment efforts with minority populations.

Reliability

If your results are to be believable, you must use data collection methods that are technically sound and provide accurate and credible information. Evaluators use the terms reliability and validity to describe the quality and the technical properties of measurement tools and procedures.

RELIABILITY

shows that a measure is dependable and yields consistent results over time

Reliability is a statistical measure of how "reproducible" the results from an instrument are. In other words, if an individual's knowledge or characteristics have not actually changed, can you get similar results each time you use the instrument? If so, the measure is reliable.

Testing a baby's bath with your elbow will give you some information about the warmth of the water, but it probably will feel less warm if you dip your elbow in a second or third time. Also having two different people test the water with their elbows will not be reliable if they have different sensitivity to heat. These ways of measuring the temperature are less reliable than using a thermometer. No matter how many times you use it, a thermometer will give you similar (reliable) results, providing there has been no change in the temperature of the bath water.

Unfortunately, social and psychological measures are not as reliable as a thermometer. There are a variety of conditions that can affect responses at any given time. People may:

- Misunderstand the questions
- Not be interested
- Be distracted
- Not trust the questioning process
- Guess at answers

RELIABILITY COEFFICIENTS

Reliability	
.8 – 1.0	High
.7 - .8	Moderately high
.6 - .7	Moderate
.5 -.6	Moderately low
Below .5	Low

Statisticians call these chance events *random error* because they may occur during one time but not necessarily at a second time. Reliability analysis determines the extent to which differences among scores reflect true differences in characteristics or simply differences due to chance.

No instrument is perfectly reliable, but a carefully constructed scale or test can minimize this type of random measurement error. Statistics called reliability

coefficients are used to account for the effects of random measurement error. Values for reliability coefficients range from 0 (all random error) to 1.0 (no random error). Reliability coefficients for most instruments, such as surveys or tests, are calculated in one of two general ways:

- *Test-retest reliability* – Measures stability of scores on the same measure taken on two different days

- *Alternate-form reliability* – Measures consistency of scores across two different versions of the same measure

Another way to measure reliability is through *internal consistency*. Using several questions focused on the same topic provides a better measure of that behavior or characteristic than a single question. But it is important that the questions consistently measure the same concept.

INTERNAL CONSISTENCY RELIABILITY

Measures the extent to which a group of questions consistently measure a behavior or characteristic.

Internal consistency reliability shows how well a group of questions measure a characteristic or behavior. The statistic used to calculate internal consistency is called Cronbach's alpha. Values range from 0 (no consistency for the questions) to 1.0 (perfect consistency).

Inter-rater reliability is the final kind of reliability. This means that two or more raters observing the same behavior or event will assess that behavior or event in the same way. Inter-rater reliability is rated on the same 0-1.0 scale as other types of reliability.

INTER-RATER RELIABILTY

Measures the extent to which multiple observers or raters of the same event see it in the same way.

Inter-rater reliability is especially important in observations and in content analysis of qualitative data. Inter-rater reliability is increased by careful training and by clear and complete definitions of the behaviors to be rated.

For example, Chapter 9 presents several observation scales for assessing the quality of child care environments. All of these scales have high inter-rater reliability coefficients because clear, specific examples are provided to guide ratings and because raters are carefully trained.

Validity

Validity refers to the ability of a measurement approach to provide accurate, convincing information about the concept or skills being measured. In other words, are you measuring what you intend to measure? If so, your measurement approach is valid.

> ***VALIDITY***
>
> *shows that an instrument actually measures what it is supposed to measure.*

Amniocentesis is a valid way to determine the gender of an unborn child, for example, because it is almost always accurate. On the other hand, using folk wisdom -- "if it rides high, it's a girl – if it rides low, it's a boy" -- is not as valid because it is less likely to give you accurate information.

Validity is measured in a variety of ways (see Table 3-1). It is important to plan your data collection method to get the most valid, accurate information about your outcomes as possible given your time and resources.

Data collection methods involving trained observers, interviews, or case records usually provide a greater degree of accuracy than self-report measures. Observations and interviews are the most labor-intensive and time-consuming of the data collection methods, but if you have only a small group to sample, these methods may be right for you.

Self-reports in the form of questionnaires are easier to administer but typically provide less accurate information than observations and interviews. People have a tendency to overstate or understate their behavior, depending on what they perceive as the more socially acceptable response. For example, in self-assessments, teens often boast about their use of drugs and alcohol while adults downplay their use of these same substances.

Even though self-reports have greater threats to their validity than some other methodologies, they can provide valuable outcome information. Steps you can take to increase unbiased self-reporting include:

- Insure the confidentiality of the responses;
- Stress the need for honest answers;
- Let people know you use the information to improve program effectiveness, not to grade them; and
- Use clear and specific behaviors and time periods to guide people's thinking and responses (see Chapter 4 for more information).

Table 3-1: Types and Examples of Validity

Type of Validity	*Example*
Face validity reflects the judgment that an instrument appears, on the face of it, to measure what it is intended to measure.	An instrument measuring knowledge of specific concepts presented in a workshop, has face validity if the concepts are clearly spelled out in the questions.
Content validity depends on the judgment of experts in the field that the items or questions are representative of the skills, knowledge, or characteristics being measured.	The Knowledge of Infant Development Scale has content validity because child development experts agree that the items accurately reflect child development information.
Construct validity describes the extent to which the instrument measures a psychological factor such as depression or hostility in individuals whose behavior manifests that factor.	If people experiencing depression also scores high on a depression scale, the scale has construct validity.
Concurrent validity refers to the utility of the measure as an indicator of the individual's *current standing* on a criterion.	Developmental screening instruments have concurrent validity if they can successfully identify children whose development is, and is not, proceeding normally.
Predictive validity refers to the ability of the measure to predict an individual's *future standing* on a criterion.	Research has shown that individuals with high scores on the Child Abuse Potential Inventory are more likely to physically abuse their children *in the future* than people with lower scores.

Communicating With Your Stakeholders

Your choice of a measurement method will ultimately depend on the people outside your program who are likely to review the information in order to make decisions about support. This may include legislators, private foundations, donors, and the public. Will the resulting data be credible to these people?

Beware of looking *only* at the reliability and validity of a given instrument. If the tool doesn't adequately measure the outcome your program is seeking, the results will not be convincing. The best measure is one that is related directly to the goals of the program, provides useful, reliable information related to the outcomes, and is credible with decision-makers and other stakeholders.

Research has shown that decision makers are likely to judge results of performance assessments, evaluations, and research in three ways:

- Was the information gathered using sound methods?
- Are the results compatible with my experience, knowledge, and values?
- Does the information provide direction – either for immediate action or for considering alternative approaches to problems ? (Rossi & Freeman, 1989)

STAKEHOLDERS AND DECISION-MAKERS

want to know that your measurement approach was reliable, valid, and appropriate to your outcomes, participants, and resources.

When your findings are congruent with what decision-makers already believe (their prior experience, knowledge, and values), then the quality of your methods is not likely to be questioned.

But if the results are unexpected or counter-intuitive, the quality of your methods becomes critical. At that point, it is important to be able to demonstrate that the measurement methods used were reliable, valid, and appropriate to your outcomes, participants, and resources.