# Introduction to Measurement and Measures

In an evaluation or research project, **measurement** is the process of gathering and recording observations and information, also known as collecting data.

A **measure** is a tool used to collect this information.

Choosing and obtaining an appropriate measure(s) for each evaluation question will likely be one of the central challenges in writing and executing an evaluation plan.

Below is a brief discussion of several concepts related to measurement and measures, and how they are inter-related.

## A. Quality of Measurement and of Measures:

The quality of measurement is an important pillar of credible and useful evaluation results. Quality of measurement depends on the quality of the measure, the appropriateness of the measure in the context of the program and its evaluation, and the manner in which that measure is administered. What follows is a discussion of features of a high quality measure.

The quality of a measure is determined by three primary characteristics (1) accuracy ("validity," in evaluation terminology); (2) consistency ("reliability") and (3) fitness (to the program setting, the program lifecycle stage, target population, stakeholder needs, and staff resources.)

**(1) Validity** – The extent to which a measure yields information (data) that is accurate with respect to the construct being measured.

Measures are efforts to get at the "truth" of something – for example, did program participants actually change with regard to the expected outcome? The idea of the outcome may be quite clear to those conducting the evaluation, but translating the idea into the real world and measuring it can be difficult.

For example, the expected outcome might be "leadership skills". The extent to which a particular measure accurately captures the essence of "leadership skills", and allows for strong statements about the program's effect on "leadership skills" and not some other, related idea, like "confidence", is called "construct validity". [1]

A measure might have weak validity for any of a number of reasons. In a survey, for example, the questions might simply not capture what they are intended to capture. This weakness might be apparent to a reasonable outsider or it might only be evident to a specialist in the field. It could also be that the wording of the survey obscures the underlying truth. For example, a measure that is intended to assess youth self-esteem might be more revealing of participants' level of openness or capacity for self-disclosure because of the language used in the measure's questions. A measure designed for adults

[1] See "Idea of Construct Validity" in Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/considea.php.

that is used with children may yield data that are heavily influenced by respondents' reading skills rather than by the actual construct being measured. A measure with vague or misleading language that is open to multiple interpretations will yield a range of responses, leading to biased results.

**Assessing validity:** Unfortunately, there is no standardized objective measure of validity. This is an inherently negotiated dimension of measure quality. But there are systematic and rigorous approaches to assessing and demonstrating validity, and these are important factors to consider (or establish) when using or developing a measure

In practice, evaluation researchers approach the assessment of validity in a number of ways. "Face validity" might be viewed as the bare minimum, and refers to the idea that, "on its face", the measure appears to do a reasonable job of capturing what it is intended to capture. A more stringent criterion would be "expert validity", in which the measure has been reviewed by an appropriate set of experts, and has been deemed to be a valid measure for the construct in question (of course, "appropriate" and "experts" can be a matter of judgment as well). Additional efforts to establish validity take the form of systematically checking how well a measure performs. For example, do scores on the measure correlate well with other outcomes that it would be expected to predict? These and other aspects of validity are beyond the scope of this paper but there are excellent resources with more information.[2]

For current purposes, the main points are: (1) that published research on measures should provide a description of what was done to assess the validity of the measure in other studies, and this information should be taken into account on the positive side of the ledger when selecting a measure; (2) when using a home-grown or adapted measure, it will be important to take some steps to assess its validity and to document choices made in constructing the measure;[3] and (3) in either case, reports of evaluation results should provide information on the measure's validity, as this can (rightly) be part of what makes the evaluation report credible.

[2] See "Measurement Validity Types" in Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/measval.php.

[3] Reviewing the literature on a construct, even if these don't include actually measures, can be helpful in assessing and demonstrating how well the new measure meets accepted definitions of interpretations of the construct. Pilot-testing a newly-developed measure is an essential minimum level of validity assessment. For guidance on pilot-testing a measure, see "Planning and Conducting a Pilot Test" from the Corporation for National and Community Service, at http://www.nationalserviceresources.org/node/19498.

**(2) Reliability** – Reliability is the ability of a measure to yield consistent or repeatable results.

In any situation, an observation is a mixture of whatever the "true" thing is, plus some "noise" or measurement error. For example, if a test is intended to measure math skills, the score a student receives on a given day is a combination of his or her "true" math ability (as captured by the measure), plus some positive or negative variation such as whether the person slept well the night before, had a distracting student near them, got lucky when guessing, etc. Reliable measures are ones that can be expected, on average, to yield the same results over time, in different contexts, when administered by different people, and across different sample groups.

**Assessing reliability:** Statistically, the reliability of a measure can be estimated in a number of ways, using various correlation statistics.[4] Reliability coefficients range from 0 (meaning that the results of the measurement are completely due to random error) to 1 (indicating that the measure has perfect reliability.) Therefore, a reliability coefficient of .5 indicates that about half of the variance of the observed score is attributable to the true level of the construct in the participant, and the other half is attributable to random factors, such as whether they had breakfast. A reliability coefficient of .8 means the score is about 80% based in true level and about 20% random error.

If possible, choose a measure that has already been reviewed for reliability by researchers in this field (peer-reviewed literature is a good source of information.) There is no firm rule for what constitutes acceptable reliability, as this varies by field. Moreover even measures that have been found to be reliable in a range of circumstances still need to be considered carefully with respect to the particular target population and setting in the program being studied.

***Note: Validity and reliability are related but different.*** It's possible to have a measure that is reliable (yielding similar scores each time) but not valid (not measuring the intended construct) or valid (questions are on topic) but not reliable (results are too sensitive to the participant's mood, the time of day, etc.)[5]. It's best to have both.

**(3) Fitness –** Appropriateness of the measure given the evaluation and program context, lifecycle and stakeholders

[4] Evaluators use a number of standardized types of reliability estimates, such as internal consistency, test-retest, inter-rater reliability, and others. (One commonly used statistic is "Cronbach's alpha", which is an estimate of the internal consistency of a measure.) For detailed explanations and a discussion of how to compare reliability statistics, see "Types of Reliability" in Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/reltypes.php.

[5] For a good discussion and nice graphic of the possible relationships between reliability and validity, see "Reliability & Validity" in Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/measval.php.

How well a measure fits a particular program and the specific constructs intended in the evaluation is also an essential element of measure quality, separate from the validity and reliability that may have been established by researchers in other study settings. For example, even if a measure has been tested by researchers and found to be valid in the context of their study, if the construct it measures doesn't truly match the construct being measured in the current context, then it's not a valid measure for the current purpose. Similarly, if a high-quality measure is not well suited to the current program's context (program fidelity or lifecycle stage, participants, setting, stakeholders, etc.), then the results that would be obtained from it would not be as accurate or reliable as the published indicators of quality might suggest.

If it's not feasible to find a pre-screened measure that captures the construct at hand and is well suited to the participants and context of the current evaluation, then candidate measures may need to be modified, combined with another measure (or measures), or it may be necessary to start from scratch. In each of these cases, it is essential to pilot the new measure with the intent to assess its validity, reliability, and fit.

**Assessing fitness:** There are no standardized quantitative assessments of "fitness" in the way it is defined here. However careful attention to program context and the construct in the evaluation question can help ensure fit, and should certainly be described in a report of evaluation results, and ideally in the measures section of the evaluation plan itself.

The purpose of the evaluation needs to be kept in mind when choosing a measure and weighing its strengths and weaknesses. Be aware, for example, that in too closely tailoring a measure to the current program (something akin to "rigging the test"), it will be difficult or even impossible to demonstrate how the program fares relative to externally accepted definitions of the construct of interest, or to compare results to those in other programs with similar goals that are using consensus measures.

## B. Forms of Data

In program evaluation, data collected using measures generally takes three forms: 1) demographic or descriptive data (describing participants or program); 2) process data (assessing program and its implementation); and 3) outcome data (assessing participants, communities, etc.). Each form may come in two types – either qualitative or quantitative. Qualitative data is generally text-based rather than numeric, while quantitative date generally refers to numerical representations of observations. It is important to be cautious when making this distinction, because all evidence has dimensions of both.[6]

Below are some examples:

[6] Encyclopedia of Evaluation, Mathison, Sandra (editor), Sage Publications Inc., 2005 pp. 345-50 and p. 351.

| | quantitative | qualitative[7] |
|---|---|---|
| descriptive | demographic characteristics of a group, rendered in percentages | information about the quality of participant skills, rendered in text form. |
| process | counts of how many participated in an activity | assessments of the rapport between program staff and program participants, rendered in text form |
| outcome | test scores for participants who completed a program | participant reflections on how their behavior has changed, rendered in text form |

Depending on the evaluation question and design, both qualitative and quantitative data may need to be coded or scored and analyzed in appropriate ways.

## C. Levels of measurement[8]

The distinctions in following categories are important. They have implications for how to interpret data from the variable in question, and for what type of analysis can be done with each kind of data. For example, it is not uncommon for a report to summarize Likert scale responses by reporting the "average" rating … but since the rating scale only assigns meaning to the integers used in the scale (1, 2, 3, 4, for example), a 3.7 is not a defined answer. It is better, in such cases, to report results in terms of the percentage of responses in each of the four categories.

Nominal (or "categorical", or "name") data is collected using response options that are labeled, and is frequently based on a quality or trait – boys or girls; blonde, brunette or redhead. The labels are helpful for organizing and extracting meaning from data, but the labels themselves don't necessarily imply order, ranking, or relative value.

Ordinal ("order" or "rank") is used to differentiate between logical order or degrees of something such as first, second and last; or Associate's, Bachelor's, Master's and Doctorate. Likert scales

[7] For references on analyzing qualitative data, see http://learningstore.uwex.edu/Analyzing-Qualitative-Data-P1023C0.aspx
For references on content analysis and thematic "coding" of qualitative data, see http://www.ischool.utexas.edu/~palmquis/courses/content.html and http://academic.csuohio.edu/kneuendorf/content/

[8] For more information, see "Levels of Measurement in Trochim, William M. The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/measlevl.php; also "Data Levels and Measurement" by David Garson, NC State University, at http://faculty.chass.ncsu.edu/garson/PA765/datalevl.htm.

are a type of ordinal data. Nominal and Ordinal data are commonly summarized using percentages.

Interval (or "continuous") data is ordered data based upon a consistent scale. Fahrenheit temperature is a common example – the difference between 65 degrees Fahrenheit and 66 degrees Fahrenheit is the same as the difference between 100 degrees Fahrenheit and 101 degrees Fahrenheit. The difference is interpretable, but there is no absolute zero (0 degrees does not mean the absence of temperature) and therefore you can't say that 100 degrees is twice as hot as 50 degrees. Dates are another type of interval data.

Ratio data is interval data that has an absolute zero, and a score of 100 is twice that of 50. Height, weight, age, etc. are ratio data. Interval and ratio data are typically summarized as averages.

In cases where audience for an evaluation requires a certain level of analysis, it may make sense to choose the level of measurement accordingly.

## D. Types of Measurement Methods

Different methods of data collection are used in different circumstances or for different goals. For example, consider the drivers' test at the DMV. The capacity to be a good driver includes knowledge of rules of the road. The DMV issues a driving permit test for this. But they also want to measure skills, so the tool they use is an observational checklist which is completed during the driving test. If they only used a paper test, they would be missing important data crucial to the success of their Program (and the safety of the roadways). If someone wanted to look at group trends they may do secondary analysis of test scores across the country.

> **1) Direct measurement** includes those methods that solicit direct feedback from participants - such as a test, survey, or interview, or that require the researcher to be physically present (participant observation). Direct measures include pencil and paper instruments (test, survey), in-person or phone interviews, or electronic assessments (Survey Monkey or other on-line instruments), and simulations. They may also include a direct observation while completing a checklist, or collecting data such as height and weight.
>
> Below are some of the most common types of direct measurement[9]:

[9] For more information on measurement types, see
Powell, Ellen-Taylor "Collecting Evaluation Data; an Overview of Sources and Methods"
http://learningstore.uwex.edu/Collecting-Evaluation-Data-An-Overview-of-Sources-and-Methods-P1025C237.aspx.
For more on the advantages and limitations of common measurement types, see
Creswell, John (2003) Research design; qualitative, quantitative and mixed methods approaches, 2cnd ed., p. 186.

| | observer is: | items are: | mode is: | use to: |
|---|---|---|---|---|
| surveys | participant | standardized questions, standard or open-ended response options | researcher administered or self-administered | assess individual attitudes, knowledge, opinions, easily compare across individuals |
| interviews | participant | structured, semi-structured or unstructured questions, follow-up prompts | researcher verbal prompts, or projective techniques such as participant drawing | assess individual attitudes, knowledge, opinions in depth; develop survey questions from themes |
| observations | researcher | checklist | researcher as unobtrusive observer or researcher as participant | assess individual or group knowledge, attitudes, and especially behavior or skills. |
| focus groups | participant and researcher | semi-structured questions, facilitation techniques | researcher as facilitator in group discussion | assess individual attitudes, knowledge, opinions; especially useful for revealing group dynamics and interactions and calling up hidden information elicited by group process. |

**2) Unobtrusive measurement** refers to methods of data collection that don't require the researcher to intrude in the research context. It consists of several different methods – indirect measures, document analysis and secondary analysis.

> ***Indirect measures*** are those measures that occur naturally within a research context – such as through video or photographs, attendance and registration records, and analysis of other outputs, such as drawing and projects (i.e. birdhouses). It is important to be aware of the ethics of collecting information without the participants' knowledge. Some researchers refer to the use of teacher and parent interviews to learn about a child as indirect measures, but since that requires the presence of the researcher, we would call those activities direct measures, even though it isn't direct interaction with the object of study.

***Document analysis*** is the analysis of documents, typically to look for themes and major ideas. Examples include news articles, field notes, reports, and memos. The study may result in the text being broken down by word, phrase, sentence, or theme.

***Secondary analysis*** is the analysis of already existing data, and typically refers to the re-analysis of data collected from one or more other projects or resources (databases). Some of the most common data sources for already available information:

- Historical/archival records
- Administrative records
- Outputs (attendance records, registration forms, diagrams, posters, etc)
- Notes/photos (process information)