Measurement Properties

Teresa Steffen, PhD, PT

Scaling

Nominal Scale: Numerals represent category levels only; classification **Examples: sex, dx, nationality Often use mode (frequency)**

Nominal categories are mutually exclusive and exhaustive so that no object or person can be assigned to more than one and no one is left out.

Ordinal Scale: Numbers indicate rank order of observations. The intervals between ranks may not be consistent or even known. Ordinal Scales may or may not contain a true zero. **Examples: MMT, satisfaction (1-7), Likert scales Often use median—ordered count**

Ordinal scales are appropriate for descriptive analysis only.
Ordinal data can undergo arithmetic operations, such as obtaining an average rank for a group BUT these manipulations can only be interpreted in the context of ranking.

Interval Scale: Equal intervals between numbers, but not related to true zero, not representing true quantity. You can have negative values. **Examples; calendar years, IQ, degrees centigrade, Often use arthrimetic mean**

Interval values can be added, subtracted and transformed but a ratio cannot be developed.

Ratio Scale: Numbers represent units with equal intervals, measured from true zero. **Examples; distance, time, age, weight, blood pressure, force-----most powerful use arthrimetic mean**

All mathematical and statistical operations are permissible with ratio level data.

Reliability

Reliability can be conceptualized as dependability, predictability or consistency. If a patient's responses are reliable, we can predict how that person will respond under given conditions. A reliable therapist is one who will be able to measure the same variables with consistent scores.

Reliability may be defined as the degree to which a measure is free from random error. The reliability of a measure is usually quantified in terms of the degree to which it renders consistent or reproducible results when properly administered under similar circumstances. The issue which must be addressed for each type of reliability is how much is good enough. Reliability is usually considered "good enough" with correlations of 0.65 to 0.94.

Measurement Properties

Teresa Steffen, PhD, PT

Types of Reliability

There are several types of reliability which can be evaluated. These include *internal consistency, interrater, intrarater, and test-retest.* The selection depends on the purpose of the measure and the type of random error that one wants estimated.

Internal consistency refers to the way individual items of the instrument group together to form a unit. Consistent total scores are more likely with high internal consistency even if there is measurement error in one (or more) of the items. **Cronbach's coefficient alpha** can be used to assess internal consistency.

Test-retest reliability, "or stability over time, is an important aspect of reliability, particularly with outcome measures." The degrees to which the scores change on repeated administration in the so called "stable" state should be taken into account when assessing "true" change over time.

Interrater reliability is the degree to which scores on a measure obtained by one trained observer agree with scores obtained by another trained observer. If trained individuals cannot agree, the assessment procedure is of doubtful use.

Intrarater reliability is the degree to which scores on a measure obtained by one trained observer agree with the scores obtained when the same observer reapplies the measure at another time.

Kappa (K) should be used in assessments yielding multiple nominal categories, because it corrects for chance.

Intra-class correlation coefficient (ICC) is usually used to determine the reliability of a test when ratings are on an *ordinal* scale. Some authors suggest that with all quantitative scales - ordinal, interval or ratio, an analysis of variance (ANOVA) model and estimates of *intra-class correlation coefficients* (ICC) are possible and desirable.

Validity

Validity concerns the extent to which an instrument measures what it intended to measure. Are we measuring what we think we are measuring? Validity addresses what we are able to say about the test results. Reliability is a necessary prerequisite to validity but does not automatically suggest validity.

Examples:

1. Does the test discriminate among individuals with or without certain traits?
2. Can we use it to evaluate change over time?
3. Can we use it to predict about a patient's potential ability or function based on the outcome of the test?

Measurement Properties

Teresa Steffen, PhD, PT

The researcher is responsible for presenting evidence to support the validity of an instrument he/she is using. Developing validity is not as straightforward as establishing reliability. The documentation of test validity is rarely satisfactory.

Types of Validity

The most common types of validity reported are *content*, *construct* and *criterion*.

I. **Content Validity**: Indicates that the items that make up an instrument adequately sample the universe of content that defines the variable being measured. This is useful with questionnaires and inventories.

II. **Construct Validity:** The degree to which the scores obtained concur with the underlying theories related to the content - the theoretical constructs. Constructs are concepts with multiple attributes and are embedded in theory. Establishing adequate construct validity requires piecing together a network of relationships. Specifically, construct validity is tested by (1) seeing whether a measure displays the pattern of converging or predictive relationships it should (convergent validity); (2) distinguishing the construct from confounding factors (divergent or discriminate validity); and (3) measuring with variations in settings, populations, and even details in measurement procedure so that generalization can be made beyond a narrow application.

III. **Criterion-Related Validity:** Indicates the outcomes of one instrument, the target test, can be used as a substitute measure for an established gold standard criterion test. This can be tested as concurrent or predictive validity.

A. **Concurrent Validity:** Establishes validity when two measures are taken at relatively the same time. Often used when the target test is considered more efficient than the gold standard and, therefore, can be used instead of the gold standard.

B. **Predictive Validity:** Establishes that the outcome of the target test can be used to predict a future criterion score.

IV. **Responsiveness:** When the specific purpose of a measure is to evaluate outcome, a new type of validity- *responsiveness* - is emerging in the measurement literature. It deals with the notion of providing evidence of the ability of the measure to detect true change over time.

- individual test items should be responsive to clinically important changes overtime
- the scale applied should have sufficient gradations to register change
- variations between replicate assessments should be small
- a strong relationship in change scores between a measure itself and external measures should exist
- instructions on interpretation of change scores should be adequately documented

Measurement Properties

Teresa Steffen, PhD, PT

There is not yet agreement on the best method for describing responsiveness for a given test. The following methods are used in our literature:

1) Change scores: pre- to post-test scores (is there a significant difference?)
2) Controlled trails: is there a significant difference between a treatment group vs. a control group after intervention?
3) Effect size (ES) [<0.4=small; 0.5=moderate; 0.8=large].

Cohen's *d* $$d = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{(n_t - 1)s_t^2 + (n_c - 1)s_c^2}{n_t + n_c}}}$$

4) Clinical significant difference (CDS) can be calculated as Minimal detectable change (MDC) or Reliable change index (RCI) or Minimal clinically important difference (MCID) or minimal important difference (MID) This concept is explained in an article by Dr. Haley in Physical Therapy, 86, 5 May 2006 p. 735-743.

Minimally detectable change (MDC) = criterion amount of change that must occur for a clinician to conclude that there is "genuine change". MDC= $z\text{-score}_{\text{level of confidence}} \times SD_{\text{baseline}} \times \sqrt{(2[1-r_{\text{test-retest}}])}$.
z-score for MDC(95) is 1.96 and for MDC(90) it is 1.64.

V. **Sensitivity:** A measure of validity of a screening procedure. Sensitivity tests the ability to obtain a positive test when the target condition is really present, or a true positive.

VI. **Specificity:** A measure of validity of a screening procedure. Specificity tests the ability to obtain a negative test when the condition is really absent, or a true negative.

Likelihood Ratios (LR): A LR combine a test's sensitivity and specificity to indicate the shift of probability give the test results. A positive LR favors the existence of a disorder (or symptom like falls).

The formula for LR are: Positive LR=Sensitivity/(1-Specificity); Negative LR is Negative=(1-Sensitivity)/Specificity.

Measurement Properties

Teresa Steffen, PhD, PT

Diagnostic Accuracy of Likelihood Ratio:

Positive LR	Negative LR	Interpretation Ratio
Greater than 10	Less than 0.1	Generate large and often conclusive shifts in probability
5-10	0.1-0.2	Generate moderate shifts in probability
2-5	0.2-0.5	Generate small but sometimes important shifts in probability
1-2	0.5-1	Alter probability to a small and rarely important degree

VII**. Ceiling & Floor Effects:** The measurement scale is incapable of discriminating among subjects above or below a certain level.

( ) after a number means SD.

When you have a mean and a standard deviation you can report a range that incorporates 95.45% of the population by adding and subtracting two standard deviations to the mean.

Many articles give confidence intervals rather than standard deviations. This assumes there is a non-normal distribution. (The scores do not follow a bell-shaped curve but are skewed to the right or left.)