What we will cover-handouts available

Graphing
Correlation
Reliability
Validity

Presentation material taken from

http://info.wlu.ca/~wwwpsych/gebotys/
http://www.socialresearchmethods.net/kb/rel&val.htm
http://www.statsoft.com/textbook/stathome.html
http://www.websurveyor.com/resources/web-survey-best-practice.asp?c=49&LEAD_SOURCE=BESTPRAC

Correlation Coefficient (Ex. 2.2 Pg.126)

How strong is the linear relationship between quantitative variables X and Y? Researchers report the correlation coefficient to summarize the strength of association between two variables. The correlation is defined as follows:

$$r_{xy} = \textit{correlation coefficient}$$

$$= \frac{s_x}{s_y} b_{xy} \qquad b_{xy} \neq b_{yx}$$

$$= \frac{s_{xy}}{s_x s_y}$$

where

$$s_x = \sqrt{\frac{1}{(n-1)} \sum (x - \bar{x})^2}$$

$$s_y = \sqrt{\frac{1}{(n-1)} \sum (y - \bar{y})^2}$$

$$s_{xy} = \sqrt{\frac{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}{(n-1)}}$$

b_{xy} = slope of least square line for (x_i, y_i)
i=1,2 ... n

The correlation is a messy calculation. It is best calculated by computer. If the standard deviation of X and Y is equal to one then the correlation is equal to the slope of the line.

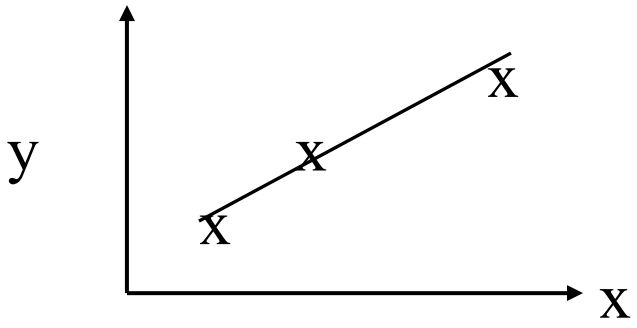
Facts

1) $r_{xy} = r_{yx}$ (the correlation of X and Y is equal to the correlation of Y and X)

2) $\mathbf{-1 \leq r_{xy} \leq 1}$
(the correlation lies between -1 and +1)

3) $r_{xy} = 1$ **if and only if** $Y_i = a + b X_i$
$i = 1, 2 \ldots n \quad b > 0$

In other words the correlation is 1 only if all of the points lie on a line whose slope is greater than 0.

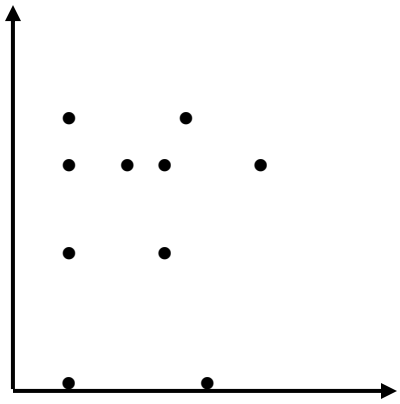


4) $r_{xy} = -1$ if and only if $Y_i = a + b X_i$
$i = 1,2 \ldots n \quad b<0$
The correlation is -1 only if all of the points lie on a line whose slope is less than 0.

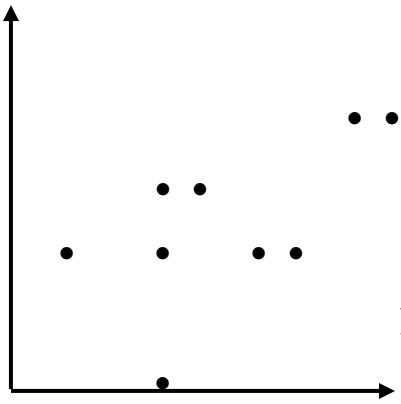
5) r^2_{xy} = proportion of observed variation in Y explained by a linear dependence on X.

Some examples are given below of data and the corresponding correlation. Note how the points cluster closer to a line as the

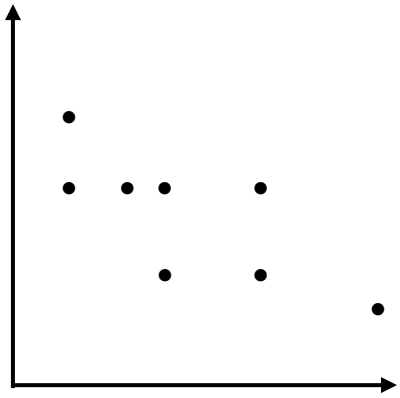
correlation moves away from 0 and closer to ± 1.



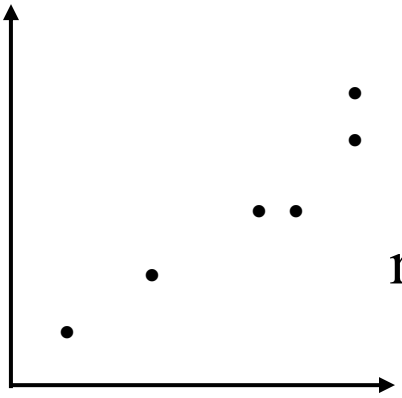
r=.02



r=.30



r= - .70



r=.90

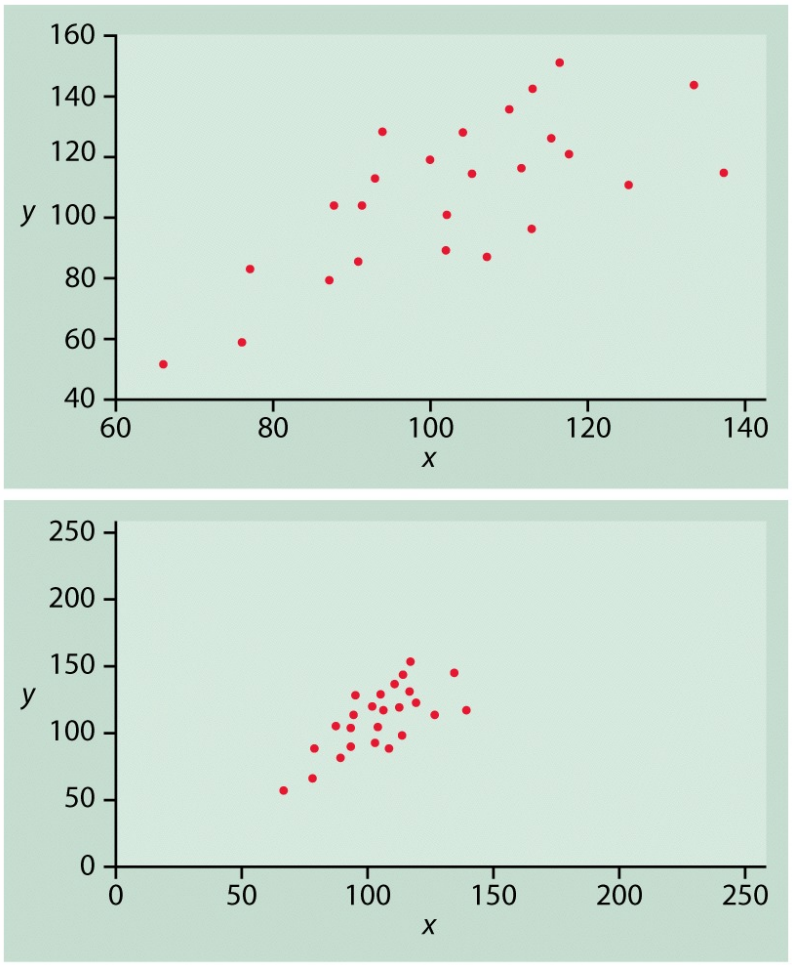


Figure 2.9 Two scatterplots of the same data; the linear pattern in the lower plot appears stronger because of the surrounding white space.

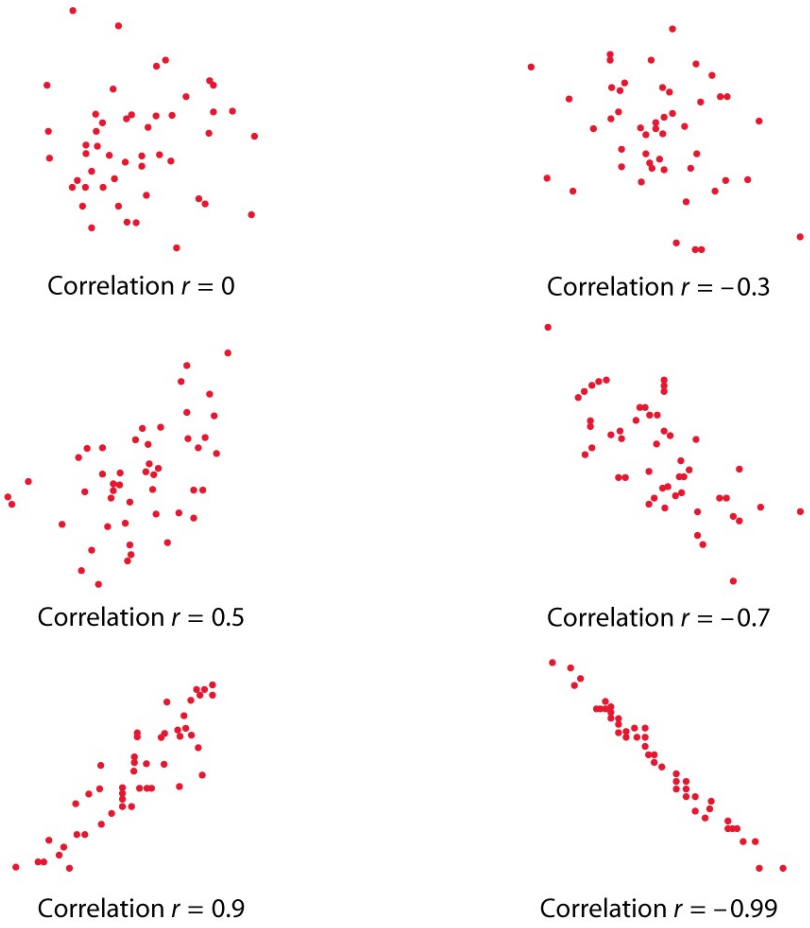


Figure 2.10 How the correlation *r* measures the direction and strength of linear association.

Example (cont): Age is (X)
Income is (Y)

The **correlation** between X and Y is:

r = .967

This is a very strong, positive relationship:

$r^2 = (.967)^2 =$ **.94 of the variation in Y is explained by X.**

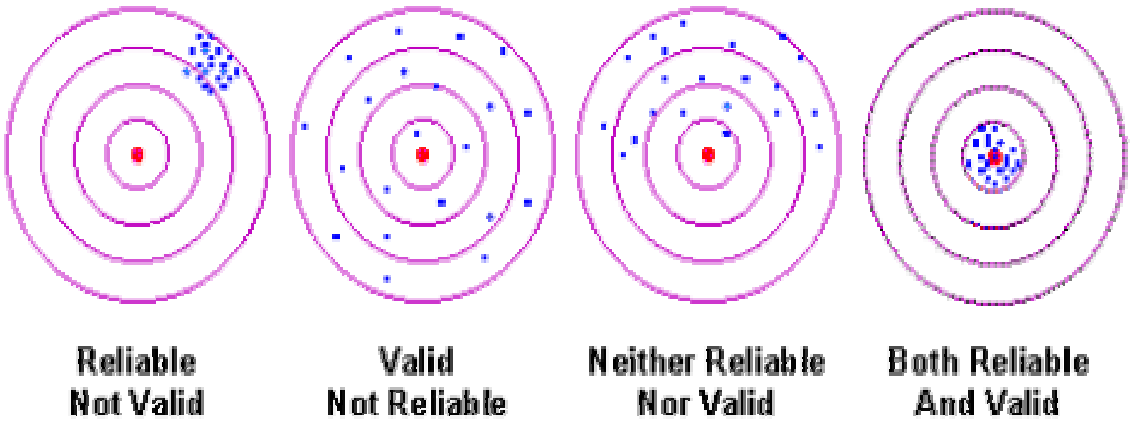
R squared is .94 or we can say that 94% of the variance in income is explained by the age.

Almost all of the variance in Y (income) is explained by X (age).

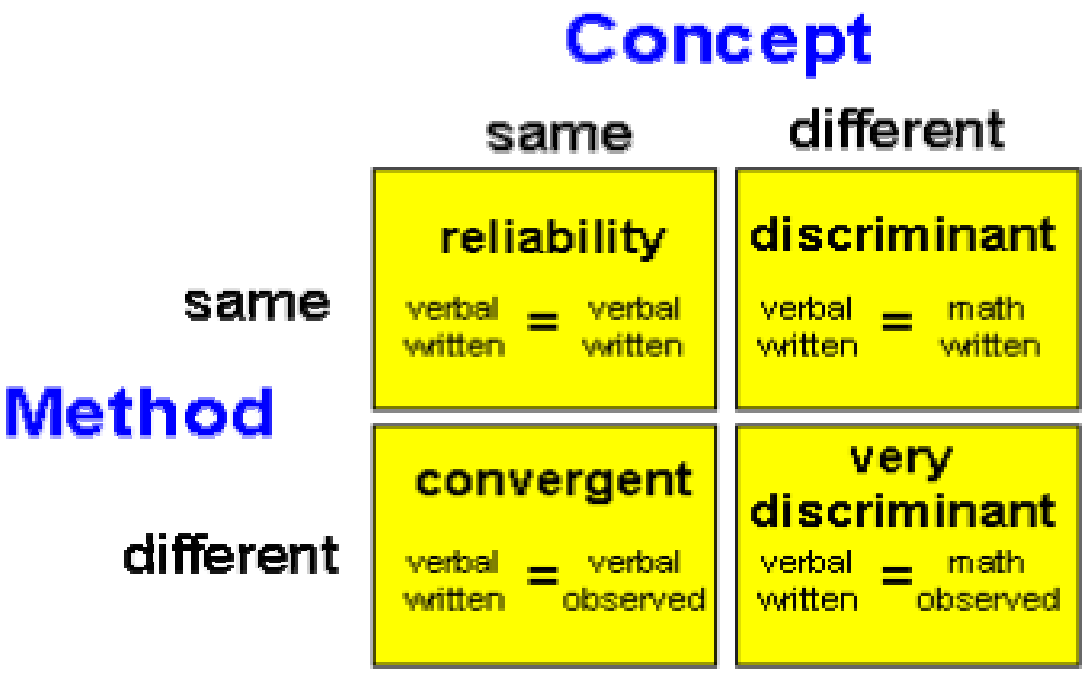
In other words there is very little variance in Y (6%) that X does not explain.

reliability and validity

We often think of reliability and validity as separate ideas but, in fact, they're related to each other. Think of the center of the target as the concept that you are trying to measure. Imagine that for each person you are measuring, you are taking a shot at the target.



Another way we can think about the relationship between reliability and validity is shown in the figure below. Imagine that we have two concepts we would like to measure, student verbal and math ability.



reliability

In many areas of research, the precise measurement of hypothesized processes or variables (theoretical *constructs*) poses a challenge by itself. For example, in psychology, the precise measurement of personality variables or attitudes is usually a necessary first step before any theories of personality or attitudes can be considered. In general, in all social sciences, unreliable measurements of people's beliefs or intentions will obviously hamper efforts to predict their behavior. The issue of precision of measurement will also come up in applied research, whenever variables are difficult to observe. For example, reliable measurement of employee performance is usually a difficult task; yet, it is obviously a necessary precursor to any performance-based compensation system.

In all of these cases, *Reliability & Item Analysis* may be used to construct reliable measurement scales, to improve existing scales, and to evaluate the reliability of scales already in use. Specifically, *Reliability & Item Analysis* will aid in the design and evaluation of *sum scales*, that is, scales that are made up of multiple individual measurements (e.g., different items, repeated measurements, different measurement devices, etc.). You can compute numerous statistics that allows you to build and evaluate scales following the so-called *classical testing theory* model.

The assessment of scale reliability is based on the correlations between the individual items or measurements that make up the scale, relative to the variances of the items. The classical testing theory model of scale construction has a long history, and there are many textbooks available on the subject. A widely acclaimed "classic" in this area, with an emphasis on psychological and educational testing, is Nunally (1970).

True Score Theory is a theory about measurement.

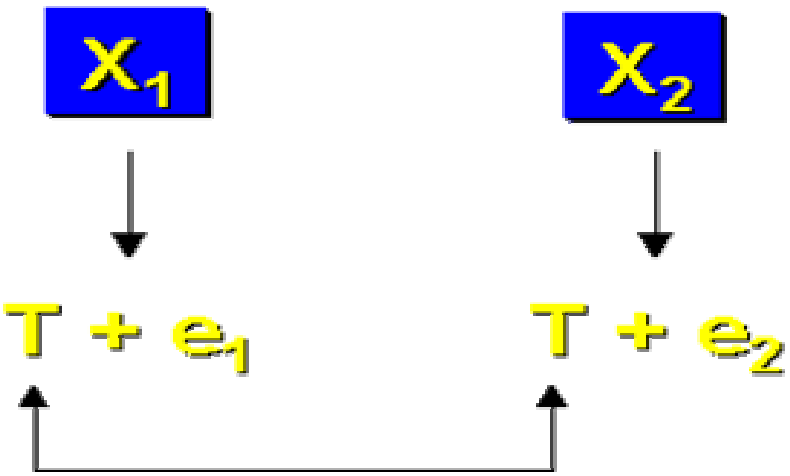
true score theory maintains that every measurement is an additive composite of two components: **true ability** (or the true level) of the respondent on that measure; and **random error**. We observe the measurement -- the score on the test, the total for a self-esteem instrument, the scale value for a person's weight. W

.

The simple equation of $\mathbf{X = T + e_X}$ has a parallel equation at the level of the variance or variability of a measure. That is, across a set of scores, we assume that:

$$\mathbf{var(X) = var(T) + var(e_X)}$$

What is **reliability**? For instance, we often speak about a car as reliable: "I have a reliable car." Or, news people talk about a "usually reliable source". In both cases, the word reliable usually means "dependable" or "trustworthy."



In research, the term reliability means "repeatability" or "consistency". A measure is considered reliable if it would give us the same result over and over again (assuming that what we are measuring isn't changing!).

We'll begin by defining a measure that we'll arbitrarily label **X**. It might be a person's satisfaction with child care survey ,score on a math achievement test or a measure of severity of illness. It is the value (numerical or otherwise) that we observe in our study. Now, to see how repeatable or consistent an observation is, we can measure it twice.

It's important to keep in mind that we observe the **X** score -- we never actually see the true (**T**) or error (**e**) scores. For instance, a student may get a score of **85** on a satisfaction with child care survey.

If our measure, **X**, is reliable, we should find that if we measure or observe it twice on the same persons that the scores are pretty much the same. If you look at the figure you should see that the only thing that the two observations have in common is their true scores, **T**. That the two observed scores, $\mathbf{X_1}$ and $\mathbf{X_2}$ are related only to the degree that the observations share true score. You should

remember that the error score is assumed to be random.the true score -- your true ability on that measure -- would be the same on both observations

Reliability is a **ratio** or fraction.

$$\frac{\text{true level on the measure}}{\text{the entire measure}}$$

You might think of reliability as the proportion of "truth" in your measure. Now, we don't speak of the reliability of a measure for an individual -- reliability is a characteristic of a measure that's taken across individuals. the definition above in terms of a set of observations. The easiest way to do this is to speak of the variance of the scores.

$$\frac{\text{the variance of the true score}}{\text{the variance of the measure}}$$

for the variance and our variable names:

$$\frac{\text{var(T)}}{\text{var(X)}}$$

how do we calculate the variance of the true scores.

we can't compute reliability because we can't calculate the variance of the true scores

the best we can do is to *estimate* it. Maybe we can get an estimate of the variability of the true scores. Remember our two observations, X_1 and X_2? We assume that these two observations would be related to each other to the degree that they share true scores. So, let's calculate the correlation between X_1 and X_2. Here's a simple formula for the correlation:

$$\frac{\text{covariance}(X_1, X_2)}{\text{sd}(X_1) * \text{sd}(X_2)}$$

where the 'sd' stands for the standard deviation (which is the square root of the variance). If we look carefully at this equation, we can see that the covariance, which simply measures the "shared" variance between measures must be an indicator of the variability of the true scores because the true scores in X_1 and X_2 are the only thing the two observations share! So, the top part is essentially an estimate of **var(T)** in this context. And, since the bottom part of the equation multiplies the standard deviation of one observation with the standard deviation of the same measure at another time, we would expect that these two values would be the same (it is the same measure we're taking) and that this is essentially the same thing as squaring the standard deviation for either observation. So, the bottom part of the equation becomes the variance of the measure (or **var(X)**). If you read this paragraph carefully, you should see that the correlation between two observations of the same measure is an estimate of reliability.

How big is an estimate of reliability? To figure this out, let's go back to the equation given earlier:

$$\frac{\text{var(T)}}{\text{var(X)}}$$

and remember that because X = T + e, we can substitute in the bottom of the ratio:

$$\frac{\text{var(T)}}{\text{var(T) + var(e)}}$$

we can easily determine the range of a reliability estimate.

Sum Scales

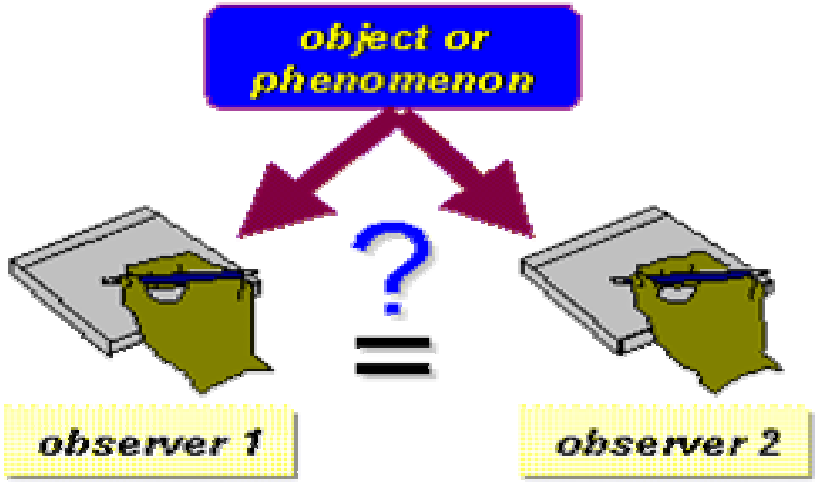
What will happen when we sum up several more or less reliable items designed to measure child care satisfaction? Suppose the items were written so as to cover a wide range of possible areas. If the error component in subjects' responses to each question is truly random, then we may expect that the different components will cancel each other out across items. In slightly more technical terms, the expected value or mean of the error component across items will be zero. The true score component remains the same when summing across items. Therefore, the more items are added, the more true score (relative to the error score) will be reflected in the sum scale.

Number of items and reliability. This conclusion describes a basic principle of test design. Namely, the more items there are in a scale designed to measure a particular concept, the more reliable will the measurement (sum scale) be.

general classes of reliability estimates

There are four *general classes of reliability estimates*, each of which estimates reliability in a different way. They are:

- **Inter-Rater or Inter-Observer Reliability**
  Used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon.
- **Test-Retest Reliability**
  Used to assess the consistency of a measure from one time to another.
- **Parallel-Forms Reliability**
  Used to assess the consistency of the results of two tests constructed in the same way from the same content domain.
- **Internal Consistency Reliability**
  Used to assess the consistency of results across items within a test.

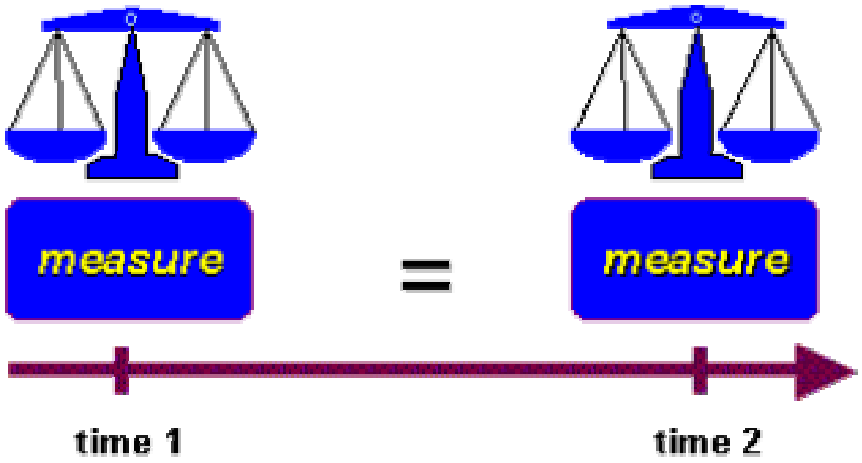


Inter-Rater or Inter-Observer Reliability

Whenever you use humans as a part of your measurement procedure, you have to worry about whether the results you get are reliable or consistent. People are notorious for their inconsistency. We are easily distractible. We get tired of doing repetitive tasks. We daydream. We misinterpret. The other major way to estimate inter-rater reliability is appropriate when the measure is a continuous one. There, all you need to do is calculate the correlation between the ratings of the two observers.

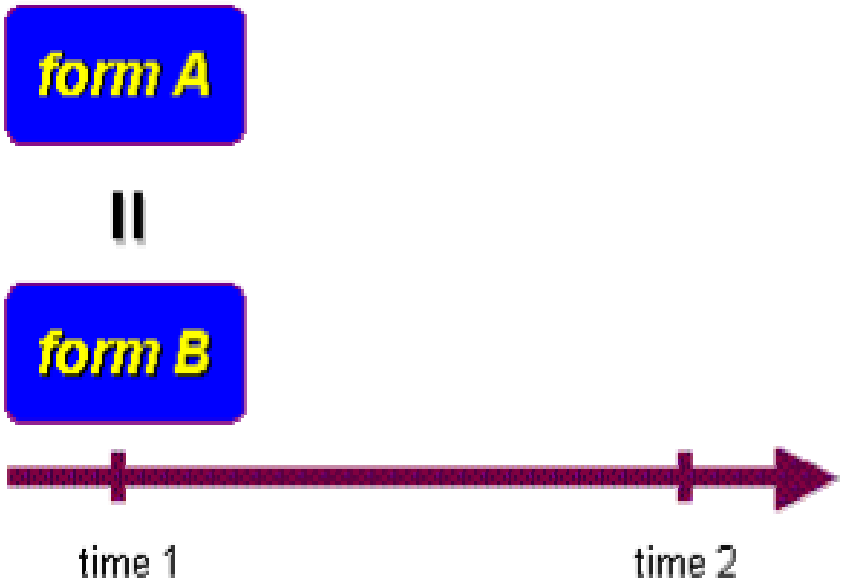
Test-Retest Reliability

We estimate test-retest reliability when we administer the same test to the same (or a similar) sample on two different occasions.



Parallel-Forms Reliability

In parallel forms reliability you first have to create two parallel forms. One way to accomplish this is to create a large set of questions that address the same construct and then randomly divide the questions into two sets. You administer both instruments to the same sample of people. The correlation between the two parallel forms is the estimate of reliability.

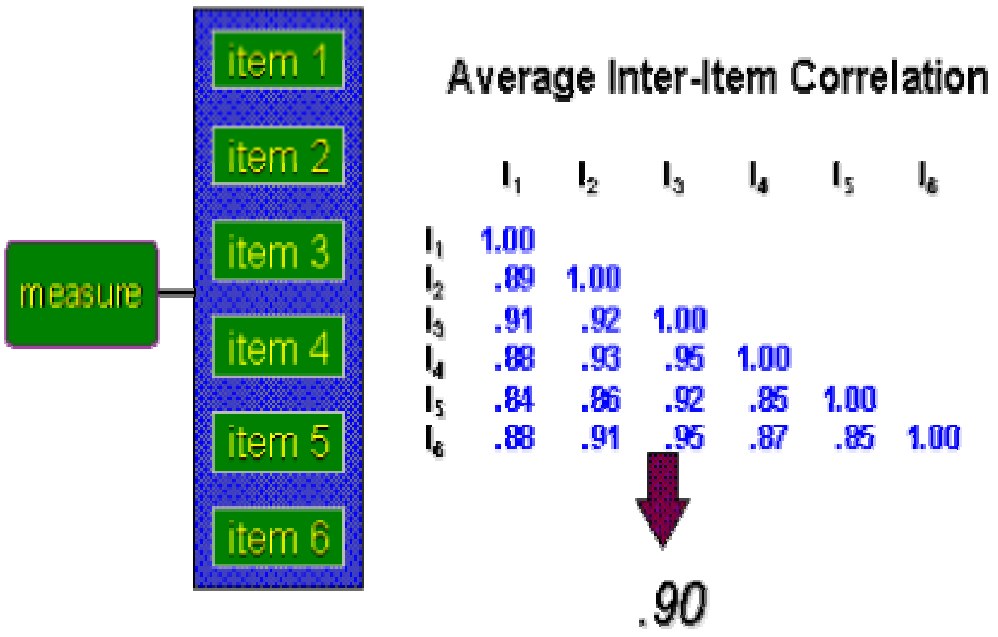


Internal Consistency Reliability

In internal consistency reliability estimation we use our single measurement instrument administered to a group of people on one occasion to estimate reliability. In effect we judge the reliability of the instrument by estimating how well the items that reflect the same construct yield similar results. We are looking at how consistent the results are for different items for the same construct within the measure. There are a wide variety of internal consistency measures that can be used.

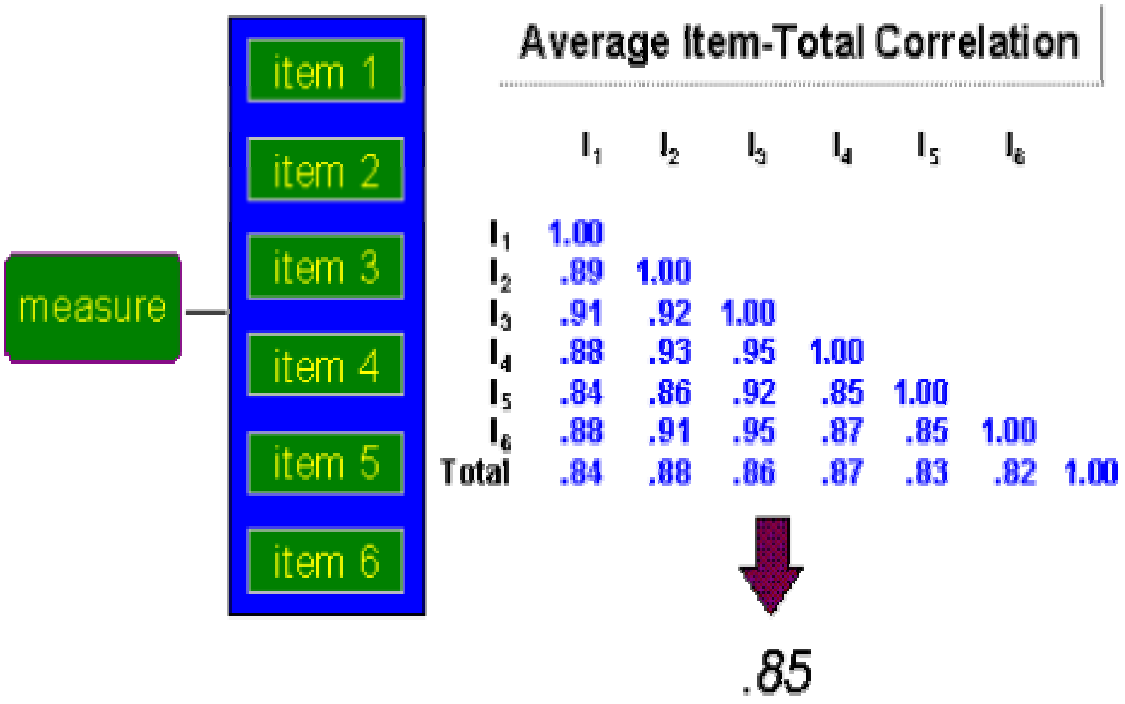
Average Inter-item Correlation

The average inter-item correlation uses all of the items on our instrument that are designed to measure the same construct. We first compute the correlation between each pair of items, as illustrated in the figure. For example, if we have six items we will have 15 different item pairings (i.e., 15 correlations). The average interitem correlation is simply the average or mean of all these correlations. In the example, we find an average inter-item correlation of .90 with the individual correlations ranging from .84 to .95.



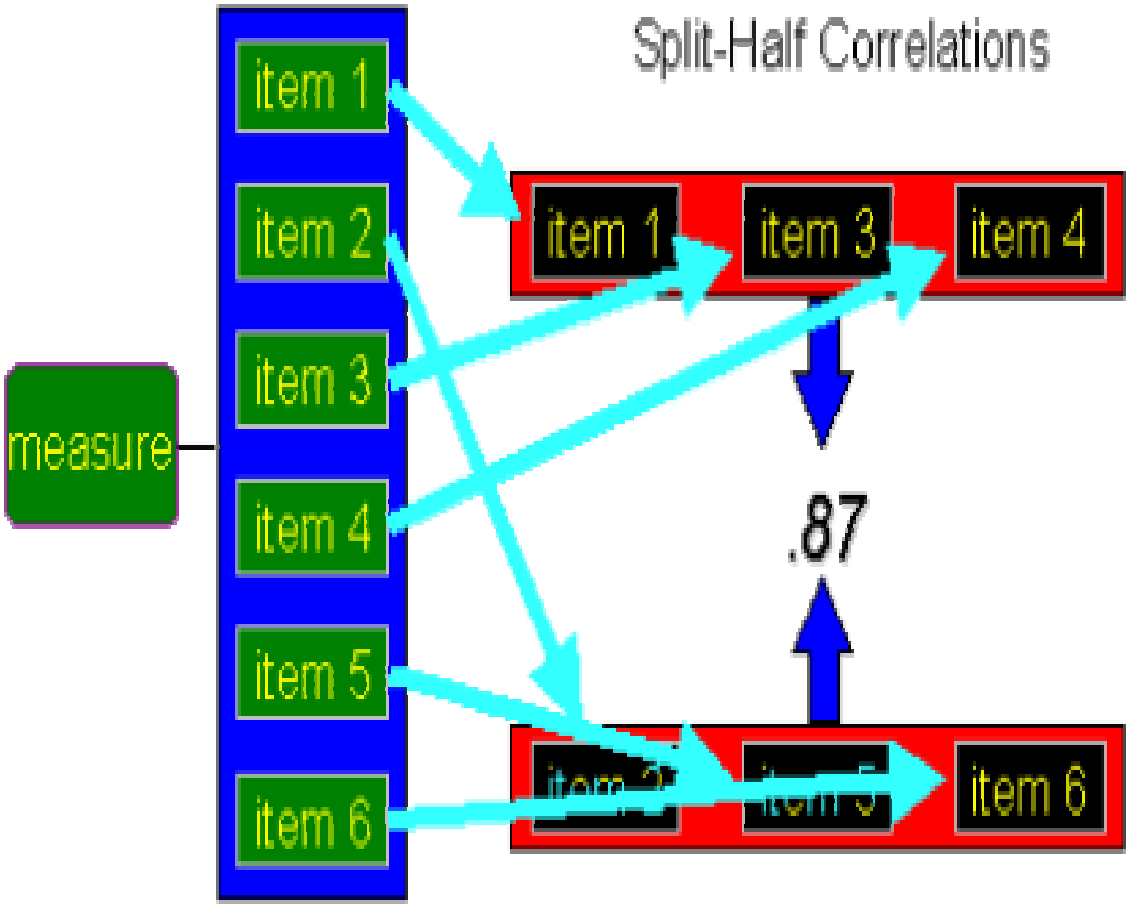
Average Itemtotal Correlation

This approach also uses the inter-item correlations. In addition, we compute a total score for the six items and use that as a seventh variable in the analysis. The figure shows the six item-to-total correlations at the bottom of the correlation matrix. They range from .82 to .88 in this sample analysis, with the average of these at .85.



Split-Half Reliability

In split-half reliability we randomly divide all items that purport to measure the same construct into two sets. We administer the entire instrument to a sample of people and calculate the total score for each randomly divided half. the split-half reliability estimate, as shown in the figure, is simply the correlation between these two total scores. In the example it is .87.



Split-Half Reliability

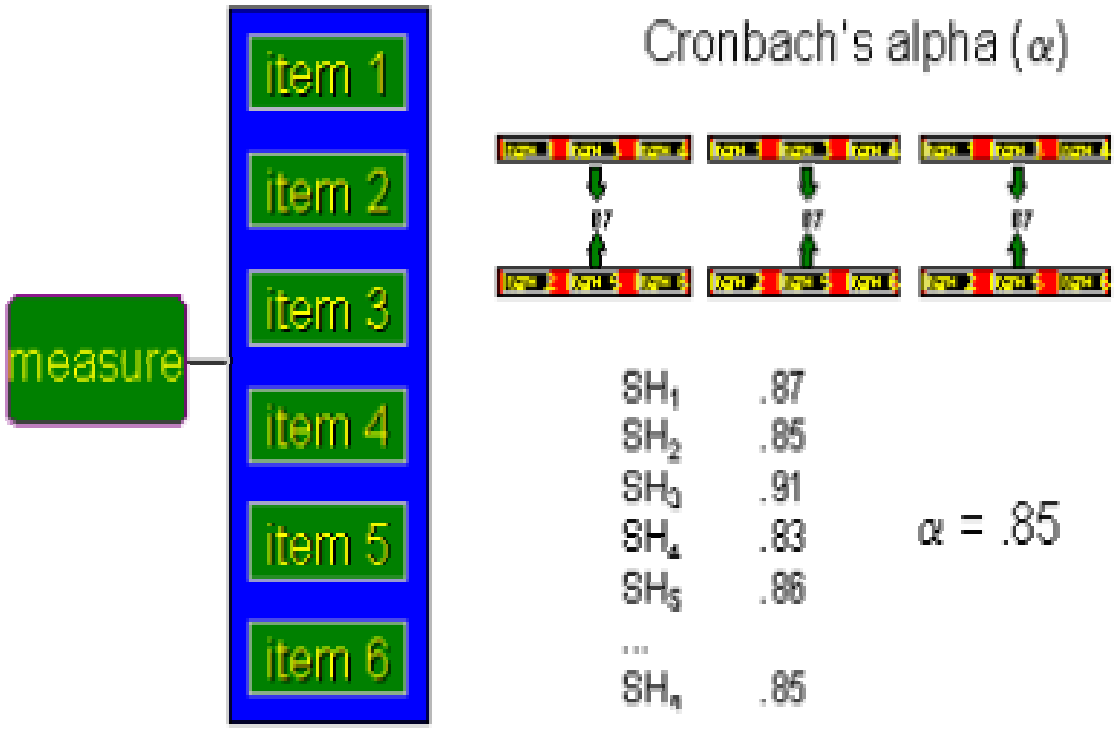
An alternative way of computing the reliability of a sum scale is to divide it in some random manner into two halves. If the sum scale is perfectly reliable, we would expect that the two halves are perfectly correlated (i.e., *r = 1.0*). Less than perfect reliability will lead to less than perfect correlations. We can estimate the reliability of the sum scale via the *Spearman-Brown split half* coefficient:

$r_{sb} = 2r_{xy} /(1+r_{xy})$

In this formula, r_{sb} is the split-half reliability coefficient, and r_{xy} represents the correlation between the two halves of the scale.

Cronbach's Alpha (α)

Imagine that we compute one split-half reliability and then randomly divide the items into another set of split halves and recompute, and keep doing this until we have computed all possible split half estimates of reliability. Cronbach's Alpha is mathematically equivalent to the average of all possible split-half estimates, The figure shows several of the split-half estimates for our six item example and lists them as SH with a subscript.



Cronbach's Alpha

if there are several subjects who respond to our items, then we can compute the variance for each item, and the variance for the sum scale. The variance of the sum scale will be smaller than the sum of item variances if the items measure the *same* variability between subjects, that is, if they measure some true score. Technically, the variance of the sum of

two items is equal to the sum of the two variances *minus* (two times) the covariance, that is, the amount of true score variance common to the two items.

We can estimate the proportion of true score variance that is captured by the items by comparing the sum of item variances with the variance of the sum scale. Specifically, we can compute:

$\alpha = (k/(k-1)) * [1- \Sigma(s^2_i)/s^2_{sum}]$

This is the formula for the most common index of reliability, namely, Cronbach's coefficient *alpha* (α). In this formula, the s_i**2's denote the variances for the k individual items; s_{sum}**2 denotes the variance for the sum of all items. If there is no true score but only error in the items (which is esoteric and unique, and, therefore, uncorrelated across subjects), then the variance of the sum will be the same as the sum of variances of the individual items. Therefore, coefficient *alpha* will be equal to zero. If all items are perfectly reliable and measure the same thing (true score), then coefficient alpha is equal to 1. (Specifically, $1-\Sigma(s_i**2)/s_{sum}**2$ will become equal to $(k-1)/k$; if we multiply this by $k/(k-1)$ we obtain 1.)

Alternative terminology. Cronbach's *alpha*, when computed for binary (e.g., true/false) items, is identical to the so-called *Kuder-Richardson-20* formula of reliability for sum scales. In either case, because the reliability is actually estimated from the consistency of all items in the sum scales, the reliability coefficient computed in this manner is also referred to as the *internal-consistency reliability*.

Comparison of Reliability Estimators

Each of the reliability estimators has certain advantages and disadvantages. Inter-rater reliability is one of the best ways to estimate reliability when your measure is an observation. However, it requires multiple raters or observers. As an alternative, you could look at the correlation of ratings of the same single observer repeated on two different occasions. For example, let's say you collected videotapes of child-mother interactions and had a rater code the videos for how often the mother smiled at the child. To establish inter-rater reliability you could take a sample of videos and have two raters code them independently. To estimate test-retest reliability you could have a single rater code the same videos on two different occasions. You might use the inter-rater approach especially if you were interested in using a team of raters and you wanted to establish that they yielded consistent results. If you get a suitably high inter-rater reliability you could then justify allowing them to work independently on coding different videos. You might use the test-retest approach when you only have a single rater and don't want to train any others. On the other hand, in some studies it is reasonable to do both to help establish the reliability of the raters or observers.

The parallel forms estimator is typically only used in situations where you intend to use the two forms as alternate measures of the same thing. Both the parallel forms and all of the internal consistency estimators have one major constraint -- you have to have multiple items designed to measure the same construct. This is relatively easy to achieve in certain contexts like achievement testing (it's easy, for instance, to construct lots of similar addition problems for a math test), but for more complex or subjective constructs this can be a real challenge. If you do have lots of items, Cronbach's Alpha tends to be the most frequently used estimate of internal consistency.

The test-retest estimator is especially feasible in most experimental and quasi-experimental designs that use a no-treatment control group. In these designs you always have a control group that is measured on two occasions (pretest and posttest). the main problem with this approach is that you don't have any information about reliability until you collect the posttest and, if the reliability estimate is low, you're pretty much sunk.

Each of the reliability estimators will give a different value for reliability. In general, the test-retest and inter-rater reliability estimates will be lower in value than the parallel forms and internal consistency ones because they involve measuring at different times or with different raters.

Designing a Reliable Scale

After the discussion so far, it should be clear that, the more reliable a scale, the better (e.g., more valid) the scale. As mentioned earlier, one way to make a sum scale more valid is by adding items. You can compute how many items would have to be added in order to achieve a particular reliability, or how reliable the scale would be if a certain number of items were added. However, in practice, the number of items on a questionnaire is usually limited by various other factors (e.g., respondents get tired, overall space is limited, etc.). Let us return to our prejudice example, and outline the steps that one would generally follow in order to design the scale so that it will be reliable:

Step 1: Generating items. The first step is to write the items. This is essentially a creative process where the researcher makes up as many items as possible that seem to relate to topic of interest(ie chid care satisfaction). In theory, one should "sample items" from the domain defined by the concept. In practice, for example in marketing research, *focus groups* are often utilized to illuminate as many aspects of the concept as possible. In educational and psychological testing, one commonly looks at other similar

questionnaires at this stage of the scale design, again, in order to gain as wide a perspective on the concept as possible.

Step 2: Choosing items of optimum difficulty. In the first draft of our questionnaire, we will include as many items as possible. We then administer this questionnaire to an initial sample of typical respondents, and examine the results for each item. First, we would look at various characteristics of the items, for example, in order to identify *floor* or *ceiling* effects. If all respondents agree or disagree with an item, then it obviously does not help us discriminate between respondents, and thus, it is useless for the design of a reliable scale. In test construction, the proportion of respondents who agree or disagree with an item, or who answer a test item correctly, is often referred to as the *item difficulty*. In essence, we would look at the item means and standard deviations and eliminate those items that show extreme means, and zero or nearly zero variances.

Step 3: Choosing internally consistent items. Remember that a reliable scale is made up of items that proportionately measure mostly true score; in our example, we would like to select items that measure mostly the topic of interest(child care satisfaction), and few esoteric aspects we consider random error. To do so, we would look at the following:

STATISTICA RELIABL. ANALYSIS	Summary for scale: Mean=46.1100 Std.Dv.=8.26444 Valid n:100 Cronbach alpha: .794313 Standardized alpha: .800491 Average inter-item corr.: .297818					
variable	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Squared Multp. R	Alpha if deleted
ITEM1	41.61000	51.93790	7.206795	.656298	.507160	.752243
ITEM2	41.37000	53.79310	7.334378	.666111	.533015	.754692
ITEM3	41.41000	54.86190	7.406882	.549226	.363895	.766778
ITEM4	41.63000	56.57310	7.521509	.470852	.305573	.776015
ITEM5	41.52000	64.16961	8.010593	.054609	.057399	.824907
ITEM6	41.56000	62.68640	7.917474	.118561	.045653	.817907
ITEM7	41.46000	54.02840	7.350401	.587637	.443563	.762033
ITEM8	41.33000	53.32110	7.302130	.609204	.446298	.758992
ITEM9	41.44000	55.06640	7.420674	.502529	.328149	.772013
ITEM10	41.66000	53.78440	7.333785	.572875	.410561	.763314

Shown above are the results for 10 items. Of most interest to us are the three right-most columns. They show us the correlation between the respective item and the total sum

score (without the respective item), the squared multiple correlation between the respective item and all others, and the internal consistency of the scale (coefficient *alpha*) if the respective item would be deleted. Clearly, items *5* and *6* "stick out," in that they are not consistent with the rest of the scale. Their correlations with the sum scale are *.05* and *.12*, respectively, while all other items correlate at *.45* or better. In the right-most column, we can see that the reliability of the scale would be about *.82* if either of the two items were to be deleted. Thus, we would probably delete the two items from this scale.

Step 4: Returning to Step 1. After deleting all items that are not consistent with the scale, we may not be left with enough items to make up an overall reliable scale (remember that, the fewer items, the less reliable the scale). In practice, one often goes through several rounds of generating items and eliminating items, until one arrives at a final set that makes up a reliable scale.

VALIDITY

A measure (e.g. a test, a questionnaire or a scale) is useful if it is reliable and valid. A measure is valid if it measures what it purports to measure. Validity can be assessed in several ways depending on the measure and its use.

1. Content Validity

Content validation is employed when it seems likely that test users will want to draw references from observed test scores to performances on a larger domain of tasks similar to items on the test. Typically, it involves asking expert judges to examine test items and judge the extent to which these items sample a specified performance domain. There are two types of content validity: face validity and logical validity. A test has face validity if an examination of the items leads to the conclusion that the items are measuring what they are supposed to be measuring. Logical or sampling validity is based on a careful comparison of the items to the definition of the domain being measured.

2. Criterion Related Validity

Criterion-related validation is a study of the relationship between test scores and a practical performance criterion that is measurable. The criterion is the thing of interest or the outcome we are concerned about. When a test score, X, can be related to a criterion score, Y, criterion-related validity can be determined. The validity coefficient, ρ_{XY} can be based on a predictive or a concurrent study. A predictive-validity coefficient is obtained by giving the test to all relevant people, waiting a reasonable amount of time, collecting criterion scores, and calculating the validity coefficient. When a test is used to predict future behaviour, predictive validity should be established. A concurrent-validity coefficient is a correlation between test and criterion scores when both measurements are obtained at the same time. Concurrent-validity coefficients are appropriate when the test scores are used to estimate a concurrent criterion rather than to predict a future criterion.

3. Construct Validity

Construct validation is appropriate whenever the test user wants to draw inferences from test scores to a behaviour domain which cannot be adequately represented by a single criterion or completely defined by a universe of content. A test's construct validity is the degree to which it measures the behaviour domain or other theoretical constructs or traits that it was designed to measure. More specifically, construct validity can be understood as the extent to which the behaviour domain or the constructs of theoretical interest have been successfully operationalized. For example, a researcher may be interested in determining clients' satisfaction with health care services. Since "satisfaction with health care services" is a construct which cannot be adequately represented by a criterion or defined by a universe of content, the researcher chooses to develop a questionnaire of 20 items in order to tap the construct "satisfaction" and proceeds to collect the data. The question is how does the researcher know that what he/she is measuring through the questionnaire is actually and purely clients' satisfaction with health care services and not something else nor a mixture with other constructs such as clients' degree of confidence in the medical profession? In this case, a construct validation is appropriate.

Establishing construct validity is an ongoing process that involves the verification of predictions made about the test scores. Procedures for construct validation may include correlations between test scores and designated criterion variables, differentiation between groups, factor analysis, multitrait-multimethod matrix analysis, or analysis of variance components within the framework of generalizability theory. The following pages will contain introductions and explanations of one of the procedures for determining construct validity: the factor analysis.

a. ***Factorial Validity***

Factorial validity is a form of construct validity that is established through a factor analysis. Factor analysis is a term that represents a large number of different mathematical procedures for analyzing the interrelationships among a set of variables and for explaining these interrelationships in terms of a reduced number of variables, called factors. A factor is a hypothetical variable that influences scores on one or more observed variables. To go back to the example on "satisfaction with health care services" cited earlier, it is not difficult to envisage that if the 20 - item questionnaire is really a valid measure of the construct "satisfaction with health care services", a factor analysis on the scores of the 20 - item questionnaire should result in one factor that can explain most of the variances in these 20 items. But if the 20 - item questionnaire is instead measuring two different behaviour domains (e.g. "Satisfaction with health care services" and "confidence in the medical profession"), factor analysis on the scores of the 20 - item questionnaire should result in two factors, with items

measuring “satisfaction” having high factor loadings[1] on one factor and items measuring “confidence” loading highly on the remaining factor.

To conclude, factorial validity is one form of construct validity. Factorial validity is assessed by the process of factor analyzing the correlations of scores from selected tests (or individual items in a single test) and obtaining a predicted factor-loading pattern.

Factotial Validity-General Purpose

The main applications of factor analytic techniques are: (1) to *reduce* the number of variables and (2) to *detect structure* in the relationships between variables, that is to *classify variables*. Therefore, factor analysis is applied as a data reduction or structure detection method (the term *factor analysis* was first introduced by Thurstone, 1931).

Basic Idea of Factor Analysis as a Data Reduction Method

Suppose we conducted a (rather "silly") study in which we measure 100 people's height in inches and centimeters. Thus, we would have two variables that measure height. If in future studies, we want to research, for example, the effect of different nutritional food supplements on height, would we continue to use both measures? Probably not; height is one characteristic of a person, regardless of how it is measured.

Let us now extrapolate from this "silly" study to something that one might actually do as a researcher. Suppose we want to measure people's satisfaction with child care. We design a satisfaction questionnaire with various items; among other things we ask our participants how satisfied they are with their child care (item 1) and how important child care is to them(item 2). Most likely, the responses to the two items are highly correlated with each other.

Combining Two Variables into a Single Factor. One can summarize the correlation between two variables in a scatterplot. A regression line can then be fitted that represents the "best" summary of the linear relationship between the variables. If we could define a variable that would approximate the regression line in such a plot, then that variable would capture most of the "essence" of the two items. participants single scores on that new factor, represented by the regression line, could then be used in future data analyses

---

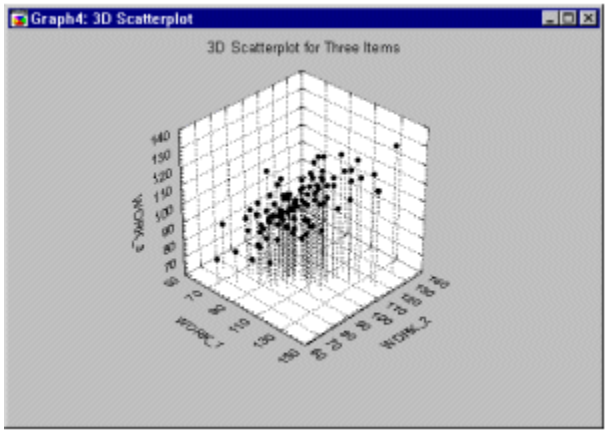
[1] The meaning of factor loadings will be discussed in greater detail in a later section. In the mentime, just imagine that a factor loading is a number which is very much like a correlation coefficient in size and meaning. When a factor analysis is conducted on a correlation matrix, tests that are influenced by certain factors are said to have high factor loadings or to load highly on those factors.

to represent that essence of the two items. In a sense we have reduced the two variables to one factor. Note that the new factor is actually a linear combination of the two variables.

Principal Components Analysis. The example described above, combining two correlated variables into one factor, illustrates the basic idea of factor analysis, or of principal components analysis to be precise (we will return to this later). If we extend the two-variable example to multiple variables, then the computations become more involved, but the basic principle of expressing two or more variables by a single factor remains the same.

Extracting Principal Components. We do not want to go into the details about the computational aspects of principal components analysis here, which can be found elsewhere (references were provided at the beginning of this section). However, basically, the extraction of principal components amounts to a *variance maximizing (varimax) rotation* of the original variable space. For example, in a scatterplot we can think of the regression line as the original *X* axis, rotated so that it approximates the regression line. This type of rotation is called *variance maximizing* because the criterion for (goal of) the rotation is to maximize the variance (variability) of the "new" variable (factor), while minimizing the variance around the new variable

Generalizing to the Case of Multiple Variables. When there are more than two variables, we can think of them as defining a "space," just as two variables defined a plane. Thus, when we have three variables, we could plot a three- dimensional scatterplot, and, again we could fit a plane through the data.



With more than three variables it becomes impossible to illustrate the points in a scatterplot, however, the logic of rotating the axes so as to maximize the variance of the new factor remains the same.

Multiple orthogonal factors. After we have found the line on which the variance is maximal, there remains some variability around this line. In principal components analysis, after the first factor has been extracted, that is, after the first line has been drawn through the data, we continue and define another line that maximizes the remaining variability, and so on. In this manner, consecutive factors are extracted. Because each

consecutive factor is defined to maximize the variability that is not captured by the preceding factor, consecutive factors are independent of each other. Put another way, consecutive factors are uncorrelated or *orthogonal* to each other.

How many Factors to Extract? Remember that, so far, we are considering principal components analysis as a data reduction method, that is, as a method for reducing the number of variables. The question then is, how many factors do we want to extract? Note that as we extract consecutive factors, they account for less and less variability. The decision of when to stop extracting factors basically depends on when there is only very little "random" variability left. The nature of this decision is arbitrary; however, various guidelines have been developed, and they are reviewed in *Reviewing the Results of a Principal Components Analysis* under *Eigenvalues and the Number-of- Factors Problem*.

Reviewing the Results of a Principal Components Analysis. Without further ado, let us now look at some of the standard results from a principal components analysis. To reiterate, we are extracting factors that account for less and less variance. To simplify matters, one usually starts with the correlation matrix, where the variances of all variables are equal to 1.0. Therefore, the total variance in that matrix is equal to the number of variables. For example, if we have 10 variables each with a variance of 1 then the total variability that can potentially be extracted is equal to 10 times 1. Suppose that in the satisfaction study introduced earlier we included 10 items to measure different aspects of satisfaction with child care at home and at work. The variance accounted for by successive factors would be summarized as follows:

STATISTICA FACTOR ANALYSIS	Eigenvalues (factor.sta) Extraction: Principal components			
Value	Eigenval	% total Variance	Cumul. Eigenval	Cumul. %
1	6.118369	61.18369	6.11837	61.1837
2	1.800682	18.00682	7.91905	79.1905
3	.472888	4.72888	8.39194	83.9194
4	.407996	4.07996	8.79993	87.9993
5	.317222	3.17222	9.11716	91.1716
6	.293300	2.93300	9.41046	94.1046
7	.195808	1.95808	9.60626	96.0626
8	.170431	1.70431	9.77670	97.7670
9	.137970	1.37970	9.91467	99.1467
10	.085334	.85334	10.00000	100.0000

Eigenvalues

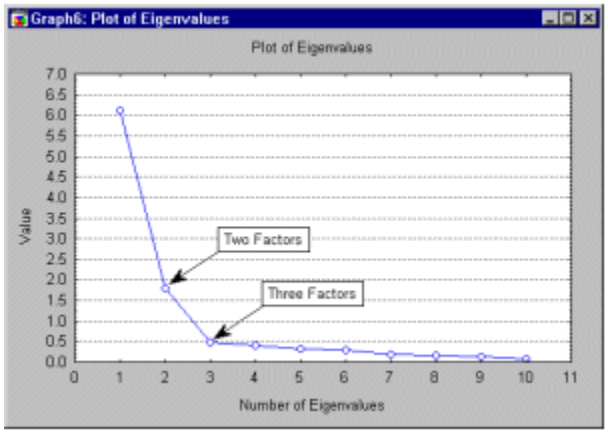
In the second column (*Eigenvalue*) above, we find the variance on the new factors that were successively extracted. In the third column, these values are expressed as a percent of the total variance (in this example, 10). As we can see, factor 1 accounts for 61 percent of the variance, factor 2 for 18 percent, and so on. As expected, the sum of the

eigenvalues is equal to the number of variables. The third column contains the cumulative variance extracted. The variances extracted by the factors are called the *eigenvalues*. This name derives from the computational issues involved.

Eigenvalues and the Number-of-Factors Problem
Now that we have a measure of how much variance each successive factor extracts, we can return to the question of how many factors to retain. As mentioned earlier, by its nature this is an arbitrary decision. However, there are some guidelines that are commonly used, and that, in practice, seem to yield the best results.

The Kaiser criterion. First, we can retain only factors with eigenvalues greater than 1. In essence this is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, we drop it. This criterion was proposed by Kaiser (1960), and is probably the one most widely used. In our example above, using this criterion, we would retain 2 factors (principal components).

The scree test. A graphical method is the *scree* test first proposed by Cattell (1966). We can plot the eigenvalues shown above in a simple line plot.



Cattell suggests to find the place where the smooth decrease of eigenvalues appears to level off to the right of the plot. To the right of this point, presumably, one finds only "factorial scree" -- "scree" is the geological term referring to the debris which collects on the lower part of a rocky slope. According to this criterion, we would probably retain 2 or 3 factors in our example.

Principal Factors Analysis
Before we continue to examine the different aspects of the typical output from a principal components analysis, let us now introduce principal factors analysis. Let us return to our satisfaction questionnaire example to conceive of another "mental model" for factor analysis. We can think of participants responses as being dependent on two components. First, there are some underlying common factors, such as the "satisfaction-with-child care" factor we looked at before. Each item measures some part of this common aspect of satisfaction. Second, each item also captures a unique aspect of satisfaction that is not addressed by any other item.

Communalities. If this model is correct, then we should not expect that the factors will extract all variance from our items; rather, only that proportion that is due to the common factors and shared by several items. In the language of factor analysis, the proportion of variance of a particular item that is due to common factors (shared with other items) is called *communality*. Therefore, an additional task facing us when applying this model is to estimate the communalities for each variable, that is, the proportion of variance that each item has in common with other items. The proportion of variance that is unique to each item is then the respective item's total variance minus the communality. A common starting point is to use the squared multiple correlation of an item with all other items as an estimate of the communality

Factor Analysis as a Classification Method

Let us now return to the interpretation of the standard results from a factor analysis. We will henceforth use the term *factor analysis* generically to encompass both principal components and principal factors analysis. Let us assume that we are at the point in our analysis where we basically know how many factors to extract. We may now want to know the meaning of the factors, that is, whether and how we can interpret them in a meaningful manner. To illustrate how this can be accomplished, let us work "backwards," that is, begin with a meaningful structure and then see how it is reflected in the results of a factor analysis. Let us return to our child care satisfaction example; shown below is the correlation matrix for items pertaining to satisfaction at work and items pertaining to satisfaction at home.

STATISTICA FACTOR ANALYSIS	Correlations (factor.sta) Casewise deletion of MD n=100					
Variable	**WORK_1**	**WORK_2**	**WORK_3**	**HOME_1**	**HOME_2**	**HOME_3**
WORK_1	1.00	.65	.65	.14	.15	.14
WORK_2	.65	1.00	.73	.14	.18	.24
WORK_3	.65	.73	1.00	.16	.24	.25
HOME_1	.14	.14	.16	1.00	.66	.59
HOME_2	.15	.18	.24	.66	1.00	.73
HOME_3	.14	.24	.25	.59	.73	1.00

The work satisfaction items are highly correlated amongst themselves, and the home satisfaction items are highly intercorrelated amongst themselves. The correlations across these two types of items (work satisfaction items with home satisfaction items) is comparatively small. It thus seems that there are two relatively independent factors reflected in the correlation matrix, one related to satisfaction at work, the other related to satisfaction at home.

Factor Loadings. Let us now perform a principal components analysis and look at the two-factor solution. Specifically, let us look at the correlations between the variables and the two factors (or "new" variables), as they are extracted by default; these correlations are also called factor *loadings*.

STATISTICA FACTOR ANALYSIS	Factor Loadings (Unrotated) Principal components	
Variable	**Factor 1**	**Factor 2**
WORK_1	.654384	.564143
WORK_2	.715256	.541444
WORK_3	.741688	.508212
HOME_1	.634120	-.563123
HOME_2	.706267	-.572658
HOME_3	.707446	-.525602
Expl.Var	2.891313	1.791000
Prp.Totl	.481885	.298500

Apparently, the first factor is generally more highly correlated with the variables than the second factor. This is to be expected because, as previously described, these factors are extracted successively and will account for less and less variance overall.

Rotating the Factor Structure. We could plot the factor loadings shown above in a scatterplot. In that plot, each variable is represented as a point. In this plot we could rotate the axes in any direction without changing the *relative* locations of the points to each other; however, the actual coordinates of the points, that is, the factor loadings would of course change. In this example, if you produce the plot it will be evident that if we were to rotate the axes by about 45 degrees we might attain a clear pattern of loadings identifying the work satisfaction items and the home satisfaction items.

Rotational strategies. There are various rotational strategies that have been proposed. The goal of all of these strategies is to obtain a clear pattern of loadings, that is, factors that are somehow clearly marked by high loadings for some variables and low loadings for others. This general pattern is also sometimes referred to as *simple structure* (a more formalized definition can be found in most standard textbooks). Typical rotational strategies are *varimax*.

We have described the idea of the varimax rotation before (see *Extracting Principal Components*), and it can be applied to this problem as well. As before, we want to find a rotation that maximizes the variance on the new axes; put another way, we want to obtain a pattern of loadings on each factor that is as diverse as possible, lending itself to easier interpretation. Below is the table of rotated factor loadings.

STATISTICA FACTOR ANALYSIS	Factor Loadings (Varimax normalized) Extraction: Principal components	
Variable	**Factor 1**	**Factor 2**
WORK_1	.862443	.051643
WORK_2	.890267	.110351
WORK_3	.886055	.152603
HOME_1	.062145	.845786

HOME_2	.107230	.902913
HOME_3	.140876	.869995
Expl.Var	2.356684	2.325629
Prp.Totl	.392781	.387605

Interpreting the Factor Structure. Now the pattern is much clearer. As expected, the first factor is marked by high loadings on the work satisfaction items, the second factor is marked by high loadings on the home satisfaction items. We would thus conclude that satisfaction, as measured by our questionnaire, is composed of those two aspects; hence we have arrived at a *classification* of the variables.

Confirmatory Factor Analysis. Over the past 15 years, so-called confirmatory methods have become increasingly popular (e.g., see Jöreskog and Sörbom, 1979). In general, one can specify *a priori*, a pattern of factor loadings for a particular number of orthogonal or oblique factors, and then test whether the observed correlation matrix can be reproduced given these specifications.

Miscellaneous Other Issues and Statistics

Factor Scores. We can estimate the actual values of individual cases (observations) for the factors. These factor scores are particularly useful when one wants to perform further analyses involving the factors that one has identified in the factor analysis.

Reproduced and Residual Correlations. An additional check for the appropriateness of the respective number of factors that were extracted is to compute the correlation matrix that would result if those were indeed the only factors. That matrix is called the *reproduced* correlation matrix. To see how this matrix deviates from the observed correlation matrix, one can compute the difference between the two; that matrix is called the matrix of *residual* correlations. The residual matrix may point to "misfits," that is, to particular correlation coefficients that cannot be reproduced appropriately by the current number of factors.

Matrix Ill-conditioning. If, in the correlation matrix there are variables that are 100% redundant, then the inverse of the matrix cannot be computed. For example, if a variable is the sum of two other variables selected for the analysis, then the correlation matrix of those variables cannot be inverted, and the factor analysis can basically not be performed. In practice this happens when you are attempting to factor analyze a set of highly intercorrelated variables, as it, for example, sometimes occurs in correlational research with questionnaires. Then you can artificially lower all correlations in the correlation matrix by adding a small constant to the diagonal of the matrix, and then restandardizing

it. This procedure will usually yield a matrix that now can be inverted and thus factor-analyzed; moreover, the factor patterns should not be affected by this procedure. However, note that the resulting estimates are not exact.