www.statsoft.com

- Products
- Solutions
- Buy
- Trials
- Support

What can we help you find? Search

TextbookReliability and Item Analysis

Looking for info about statistics?
We wrote the book on it.
And you can read it for free!

Elementary Concepts
Statistics Glossary
Basic Statistics
ANOVA / MANOVA
Association Rules
Boosting Trees
Canonical Analysis
CHAID Analysis
C & R Trees
Classification Trees
Cluster Analysis
Correspondence Analysis
Data Mining Techniques
Discriminant Analysis
Distribution Fitting
Experimental Design
Factor Analysis
General Discrim. Analysis
General Linear Models
Generalized Additive Mod.
Generalized Linear Mod.
General Regression Mod.
Graphical Techniques
Ind.Components Analysis
Linear Regression
Log-Linear Analysis
MARSplines
Machine Learning
Multidimensional Scaling
Neural Networks
Nonlinear Estimation
Nonparametric Statistics
Partial Least Squares
Power Analysis
Process Analysis
Quality Control Charts
Reliability / Item Analysis
SEPATH (Structural eq.)
Survival Analysis
Text Mining
Time Series / Forecasting

Reliability and Item Analysis

This topic discusses the concept of *reliability of measurement* as used in social sciences (but not in industrial statistics or biomedical research). The term *reliability* used in industrial statistics denotes a function describing the probability of failure (as a function of time). For a discussion of the concept of *reliability* as applied to product quality (e.g., in industrial statistics), please refer to the section on Reliability/Failure Time Analysis in the *Process Analysis* topic (see also the section Repeatability and Reproducibility and the topic Survival/Failure Time Analysis). For a comparison between these two (very different) concepts of *reliability*, see Reliability.

General Introduction

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

Variance Components
Statistical Advisor
Distribution Tables
References Cited
Send Comments
Business Solutions
Free Resources
About Textbook

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. If you are not familiar with the *correlation coefficient* or the variance statistic, we recommend that you review the respective discussions provided in the Basic Statistics section.

The classical testing theory model of scale construction has a long history, and there are many textbooks available on the subject. For additional detailed discussions, you may refer to, for example, Carmines and Zeller (1980), De Gruijter and Van Der Kamp (1976), Kline (1979, 1986), or Thorndyke and Hagen (1977). A widely acclaimed "classic" in this area, with an emphasis on psychological and educational testing, is Nunnally (1970).

Testing hypotheses about relationships between items and tests. Using Structural Equation Modeling and Path Analysis (*SEPATH*), you can test specific hypotheses about the relationship between sets of items or different tests (e.g., test whether two sets of items measure the same construct, analyze multi-trait, multi-method matrices, etc.).

To index

Basic Ideas

Suppose we want to construct a questionnaire to measure people's prejudices against foreign-made cars. We could start out by generating a number of items such as: "Foreign cars lack personality," "Foreign cars all look the same," etc. We could then submit those questionnaire items to a group of subjects (for example, people who have never owned a foreign-made car). We could ask subjects to indicate their agreement with these statements on 9-point scales, anchored at *1=disagree* and *9=agree*.

True scores and error. Let us now consider more closely what we mean by precise measurement in this case. We hypothesize that there is such a thing (theoretical construct) as "prejudice against foreign cars," and that each item "taps" into this concept to some extent. Therefore, we may say that a subject's response to a particular item reflects two aspects: first, the response reflects the prejudice against foreign cars, and second, it will reflect some esoteric aspect of the respective question. For example, consider the item "Foreign cars all look the same." A subject's agreement or disagreement with that statement will partially depend on his or her general prejudices, and partially on some other aspects of the question or person. For example, the subject may have a friend who just bought a very different looking foreign car.

Testing hypotheses about relationships between items and tests. To test specific hypotheses about the relationship between sets of items or different tests (e.g., whether two sets of items measure the same construct, analyze multi- trait, multi-method matrices, etc.) use *Structural Equation Modeling (SEPATH)*.

To index

Classical Testing Model

To summarize, each measurement (response to an item) reflects to some extent the true score for the intended concept (prejudice against foreign cars), and to some extent esoteric, random error. We can express this in an equation as:

$X = tau + error$

In this equation, *X* refers to the respective actual measurement, that is, subject's response to a particular item; *tau* is commonly used to refer to the *true score*, and *error* refers to the random error component in the measurement.

To index

Reliability

In this context the definition of *reliability* is straightforward: a measurement is reliable if it reflects mostly true score, relative to the error. For example, an item such as "Red foreign cars are particularly ugly" would likely provide an unreliable measurement of prejudices against foreign- made cars. This is because there probably are ample individual differences concerning the likes and dislikes of colors. Thus, this item would "capture" not only a person's prejudice but also his or her color preference. Therefore, the proportion of true score (for prejudice) in subjects' response to that item would be relatively small.

Measures of reliability. From the above discussion, one can easily infer a measure or statistic to describe the reliability of an item or scale. Specifically, we may define an *index of reliability* in terms of the proportion of true score variability that is captured across subjects or respondents, relative to the total observed variability. In equation form, we can say:

$$\text{Reliability} = \sigma^2_{\text{(true score)}} \;/\; \sigma^2_{\text{(total observed)}}$$

To index

Sum Scales

What will happen when we sum up several more or less reliable items designed to measure prejudice against foreign-made cars? Suppose the items were written so as to cover a wide range of possible prejudices against foreign-made cars. 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. Perhaps a somewhat more practical example will further clarify this point. Suppose you want to measure the height of 10 persons, using only a crude stick as the measurement device. Note that we are not interested in this example in the absolute correctness of measurement (i.e., in inches or centimeters), but rather in the ability to distinguish reliably between the 10 individuals in terms of their height. If you measure each person only once in terms of multiples of lengths of your crude measurement stick, the resultant measurement may not be very reliable. However, if you measure each person 100 times, and then take the average of those 100 measurements as the summary of the respective person's height, then you will be able to make very precise and reliable distinctions between people (based solely on the crude measurement stick).

Let's now look at some of the common statistics that are used to estimate the reliability of a sum scale.

To index

Cronbach's Alpha

To return to the prejudice example, 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*.

To index

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.

To index

Correction for Attenuation

Let us now consider some of the consequences of less than perfect reliability. Suppose we use our scale of prejudice against foreign-made cars to predict some other criterion, such as subsequent actual purchase of a car. If our scale correlates with such a criterion, it would raise our confidence in the *validity* of the scale, that is, that it really measures prejudices against foreign-made cars, and not something completely different. In actual test design, the *validation* of a scale is a lengthy process that requires the researcher to correlate the scale with various external criteria that, in theory, should be related to the concept that is supposedly being measured by the scale.

How will validity be affected by less than perfect scale reliability? The random error portion of the scale is unlikely to correlate with some external criterion. Therefore, if the proportion of true score in a scale is only 60% (that is, the reliability is only .60), then the correlation between the scale and the criterion variable will be *attenuated*, that is, it will be smaller than the actual correlation of true scores. In fact, the validity of a scale is always limited by its reliability.

Given the reliability of the two measures in a correlation (i.e., the scale and the criterion variable), we can estimate the actual correlation of true scores in both measures. Put another way, we can *correct* the correlation *for attenuation*:

$r_{xy,corrected} = r_{xy} /(r_{xx}*r_{yy})^{½}$

In this formula, $r_{xy,corrected}$ stands for the corrected correlation coefficient, that is, it is the estimate of the correlation between the true scores in the two measures x and y. The term r_{xy} denotes the uncorrected correlation, and r_{xx} and r_{yy} denote the reliability of measures (scales) x and y. You can compute the attenuation correction based on specific values, or based on actual raw data (in which case the reliabilities of the two measures are estimated from the data).

To index

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 prejudices against foreign-made cars. 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. For example, we could ask a small group of highly committed American car buyers to express their general thoughts and feelings about foreign-made cars. 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 prejudice 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 prejudice against foreign-made cars, and few esoteric aspects we consider random error. To do so, we would look at the following:

STATISTICA RELIABL.	Summary for scale: Mean=46.1100 Std.Dv.=8.26444 Valid n:100 Cronbach alpha: .794313 Standardized alpha: .800491

ANALYSIS	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.

Tetrachoric correlations. In educational and psychological testing, it is common to use *yes/no* type items, that is, to prompt the respondent to answer either yes or no to a question. An alternative to the regular correlation coefficient in that case is the so-called *tetrachoric* correlation coefficient. Usually, the tetrachoric correlation coefficient is larger than the standard correlation coefficient, therefore, Nunnally (1970, p. 102) discourages the use of this coefficient for estimating reliabilities. However, it is a widely used statistic (e.g., in mathematical modeling).

To index