An Introduction to Classical Test Theory and Reliability

Charlie Lewis, Educational Testing Service

clewis@ets.org

Prologue

This document provides a formal, model-based introduction to classical test theory and reliability. More specifically, it describes a parametric statistical model for observed and true test scores, and uses this model to define the reliability of a test. It extends the model, first to the case of two tests and then to the general case of multiple tests as well as to composite tests. In these extensions, we consider assumptions about various relationships that may exist among true scores for different tests and the implications that these relationships have for expressing reliability.

The traditional model for classical test theory (see, for instance, Lord & Novick, 1968) differs in two important ways from the model used here. First, the traditional model is non-parametric, in the sense that no specific probability distribution is assumed for the test scores. The model used here assumes (multivariate) Normal distributions for test scores. Second, the traditional model adopts a two-level hierarchical sampling approach, assuming both between-person and within-person (replication) sampling of test scores. The model used here assumes only between-person sampling of test scores. This follows Holland's (1990) treatment of probabilities in item response theory. Both of these differences with the traditional model allow a simpler treatment of classical test theory with the model used here.

Since we are using a formal, model-based approach to introducing reliability, it may be worthwhile to make a general observation about the use of models. Describing his own work in test theory, Rasch (1960, p. 37) stated the following:

> "That the model is not true is certainly correct, no models are … Models should not be true but it is important that they are applicable."

To put this another way, it is *not* important that the assumptions made by a model be correct. They certainly will not be. What *is* important is that the inferences we wish to make based on a model are robust to the (inevitable) violations of the model's assumptions. For instance, the model introduced here assumes that both true and error scores are Normally distributed, and that conditional error variances are constant for different values of the true score. These assumptions certainly do not hold, except for simulated data. However, the formulas that are derived based on these assumptions are useful in many settings. This is analogous to the robustness of inferences based on the standard linear regression model in statistics.

Finally, before beginning the actual introduction, let's compare our treatment of classical test theory with item response theory. The use of a parametric classical test theory model and only considering between-person sampling of test scores both parallel modern treatments of item response theory. This makes the point that classical test theory and item response theory have much more in common than is often acknowledged. It is a point that is also made by Lord (1980, p. 7) in his text on item response theory:

> "Nothing in this book will contradict either the assumptions or basic conclusions of classical test theory. Additional assumptions will be made; these will allow us to answer questions that classical test theory cannot answer."

A Parametric Classical Test Theory Model for a Single Test: Defining Reliability

We start with a population of people and a test that we think measures something about the people that's of interest to us. Let's imagine that each person in the population has what we'll call an *observed score* on the test. For a randomly selected person from the population, we may

treat their observed score as a random variable and denote it by *X*. It has a probability distribution with mean denoted by

$$Ave(X) = \mu_X$$

and variance denoted by

$$Var(X) = \sigma_X^2 \ .$$

To discuss the concept of the *reliability* of an observed score for a given test and a given population of people, we need to introduce something called a *true score*, also associated with each person in the population. This true score is sometimes referred to as a *latent variable* to emphasize the fact that it is not observable. True scores (and latent variables generally) only exist in the context of a model, so we need to make that model explicit. Now that we have two scores associated with each person, randomly sampling a person from the population creates a bivariate random variable that we may denote by (X,T).

Our model has two components. First, we specify the conditional distribution of *X* given *T*:

$$X|T \sim N(T, \sigma_E^2) \ .$$

This part of the model says that the conditional distribution of *X* given *T* is Normal, with mean *T* and variance σ_E^2 :

$$Ave(X|T) = T \text{ and } Var(X|T) = \sigma_E^2 \ .$$

In other words, if we take the mean observed score for all persons in our population who have the same true score, then that mean observed score is equal to the true score. Also, if we find the

variance of the observed scores for all persons in our population who have the same true score, that variance is equal to a constant (not dependent on the true score) that we denote by σ_E^2.

The second component of the model simply says that the true score has a Normal distribution with mean μ_T and variance σ_T^2:

$$T \sim N\left(\mu_T, \sigma_T^2\right) .$$

These two components taken together imply that $\left(X, T\right)$ has a bivariate Normal distribution.

They also imply that we may write the mean and variance of the observed score as

$$\mu_X = \mu_T$$

and

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 .$$

Thus we have

$$X \sim N\left(\mu_T, \sigma_T^2 + \sigma_E^2\right) .$$

Moreover, we may write the covariance between *X* and *T* as

$$Cov\left(X, T\right) = \sigma_T^2 .$$

Consequently, we may write the squared correlation between *X* and *T* as

$$Corr\left(X, T\right)^2 = \rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} .$$

The *reliability* of *X* for the given population of persons is defined as this squared correlation:

$$Rel(X) = \rho_{XT}^2 .$$

As indicated above, we may also write

$$Rel(X) = \frac{\sigma_T^2}{\sigma_X^2} \text{ and } Rel(X) = 1 - \frac{\sigma_E^2}{\sigma_X^2} .$$

A Classical Test Theory Model for Two Tests

Now suppose we wish to consider the relationship between two tests. We may denote the random variables defined by their observed scores as X_1 and X_2, with corresponding true score random variables denoted as T_1 and T_2. Any time we are working with two tests, we will extend our earlier model in one important way. Specifically, we assume

$$(X_1, X_2) \big| (T_1, T_2) \sim N_2 \left[(T_1, T_2); \begin{pmatrix} \sigma_{E1}^2 & 0 \\ 0 & \sigma_{E2}^2 \end{pmatrix} \right] .$$

The new assumption here is that, given their true scores, the two observed scores are uncorrelated. In the context of the bivariate Normal distribution, this means that the observed scores are *conditionally independent*, given their true scores.

We also make an assumption about the distribution of the true scores. Specifically, we assume that (T_1, T_2) has a general bivariate Normal distribution:

$$(T_1, T_2) \sim N_2 \left[(\mu_{T1}, \mu_{T2}); \begin{pmatrix} \sigma_{T1}^2 & \sigma_{T12} \\ \sigma_{T12} & \sigma_{T2}^2 \end{pmatrix} \right] .$$

These two assumptions imply the following form for the distribution of the observed scores:

$$\left(X_1, X_2\right) \sim N_2\left[\left(\mu_{T1}, \mu_{T2}\right); \begin{pmatrix} \sigma_{T1}^2 + \sigma_{E1}^2 & \sigma_{T12} \\ \sigma_{T12} & \sigma_{T2}^2 + \sigma_{E2}^2 \end{pmatrix}\right].$$

In particular, note that we may write the covariance between the two observed scores as equal to the covariance between their true scores:

$$Cov\left(X_1, X_2\right) = Cov\left(T_1, T_2\right) = \sigma_{T12}\,.$$

This is a consequence of the conditional independence assumption.

Congeneric Tests

Now suppose that the two tests we are considering "measure the same thing." Specifically, suppose there exists a "common" true score T with $Ave\left(T\right) = 0$ and $Var\left(T\right) = 1$ such that

$$T_1 = \sigma_{T1} T + \mu_1 \text{ and } T_2 = \sigma_{T2} T + \mu_2\,.$$

In this case, X_1 and X_2 are called *congeneric tests*.

Using our earlier results, this implies that

$$Ave\left(X_1\right) = Ave\left(T_1\right) = \mu_1 \text{ and } Ave\left(X_2\right) = Ave\left(T_2\right) = \mu_2\,.$$

The model also implies that

$$Var\left(T_1\right) = \sigma_{T1}^2,\ Var\left(T_2\right) = \sigma_{T2}^2 \text{ , and } Cov\left(T_1, T_2\right) = \sigma_{T1}\sigma_{T2} \text{ , so } Corr\left(T_1, T_2\right) = 1\ .$$

As a consequence of the perfect correlation between T_1 and T_2, we may write

$$\left[Corr\left(X_1 X_2\right)\right]^2 = \frac{\left[Cov\left(T_1, T_2\right)\right]^2}{Var\left(X_1\right) Var\left(X_2\right)} = \left(\frac{\sigma_{T1}^2}{\sigma_{X1}^2}\right)\left(\frac{\sigma_{T2}^2}{\sigma_{X2}^2}\right) = Rel\left(X_1\right) Rel\left(X_2\right)\ .$$

Essentially Tau Equivalent Tests

A special case of congeneric tests occurs when

$$\sigma_{T1} = \sigma_{T2} = \sigma_T \ .$$

In this case, the two tests are called *essentially tau equivalent*. (If it is also the case that $\mu_1 = \mu_2$, then the two tests are simply called *tau equivalent*.)

The essentially tau equivalent model implies that

$$Var(T_1) = Var(T_2) = Cov(T_1, T_2) = \sigma_T^2 .$$

Consequently, we have

$$Cov(X_1, X_2) = \sigma_T^2 \ .$$

In general, the reliability of X_1 is given by

$$Rel(X_1) = \frac{Var(T_1)}{Var(X_1)} \ .$$

If X_1 and X_2 are essentially tau equivalent, we may write

$$Rel(X_1) = \frac{Cov(X_1, X_2)}{Var(X_1)} .$$

Similarly,

$$Rel(X_2) = \frac{Cov(X_1, X_2)}{Var(X_2)} .$$

In other words, if X_1 and X_2 are essentially tau equivalent, we may express the reliability of each of these tests in terms of moments of the observed scores. Note that the two tests are not necessarily equally reliable, due to the fact that σ_{E1}^2 and σ_{E2}^2 need not be equal.

(Essentially) Parallel Tests

A special case of the essentially tau equivalent model occurs when

$$\mu_1 = \mu_2 \text{ and } \sigma_{E1}^2 = \sigma_{E2}^2 \ .$$

In this case,

$$T_1 = T_2 \ , \ Ave\left(X_1\right) = Ave\left(X_2\right) \text{ and } Var\left(X_1\right) = Var\left(X_2\right).$$

We say that these two tests are *parallel* and we have the result that

$$Rel\left(X_1\right) = Rel\left(X_2\right) = Corr\left(X_1, X_2\right) = \rho_{X12} \ .$$

This result says that the reliability of a test is given by its correlation with a parallel test. Note that this result does not require the test means to be equal. Suppose we call two tests with $\mu_1 \neq \mu_2$ and $\sigma_{E1}^2 = \sigma_{E2}^2$ *essentially parallel*. We may generalize our result to say that the reliability of a test equals its correlation with an essentially parallel test.

Composite Tests

Returning to our general model for two tests X_1 and X_2, we define a new composite observed score equal to their sum:

$$X_{1+2} = X_1 + X_2 \ .$$

We may define the true score for this composite observed score as

$$T_{1+2} = T_1 + T_2 \ .$$

Based on our general bivariate classical test theory model, we may write

$$X_{1+2} \left| T_{1+2} \sim N\left(T_{1+2}, \sigma_{E1}^2 + \sigma_{E2}^2\right) \right. \text{ and } T_{1+2} \sim N\left(\mu_{T1} + \mu_{T2}, \sigma_{T1}^2 + 2\sigma_{T12} + \sigma_{T2}^2\right) .$$

These results imply the following distribution for the composite observed score:

$$X_{1+2} \sim N\left(\mu_{T1} + \mu_{T2}, \sigma_{T1}^2 + \sigma_{E1}^2 + 2\sigma_{T12} + \sigma_{T2}^2 + \sigma_{E2}^2\right).$$

In general, the reliability of X_{1+2} may be written as

$$Rel\left(X_{1+2}\right) = \frac{Var\left(T_{1+2}\right)}{Var\left(X_{1+2}\right)} = \frac{\sigma_{T1}^2 + 2\sigma_{T12} + \sigma_{T2}^2}{\sigma_{T1}^2 + \sigma_{E1}^2 + 2\sigma_{T12} + \sigma_{T2}^2 + \sigma_{E2}^2} = 1 - \frac{\sigma_{E1}^2 + \sigma_{E2}^2}{\sigma_{T1}^2 + \sigma_{E1}^2 + 2\sigma_{T12} + \sigma_{T2}^2 + \sigma_{E2}^2} .$$

Now suppose that X_1 and X_2 are essentially tau equivalent. In this case, the composite true score may be written as

$$T_{1+2} = 2T + \mu_1 + \mu_2 \ .$$

The reliability of X_{1+2} is given by

$$Rel\left(X_{1+2}\right) = \frac{Var\left(T_{1+2}\right)}{Var\left(X_{1+2}\right)} = \frac{4\sigma_T^2}{Var\left(X_{1+2}\right)} \ .$$

Using the result for the covariance of essentially tau equivalent tests, we may write the reliability of X_{1+2} as

$$Rel\left(X_{1+2}\right) = \frac{4Cov\left(X_1, X_2\right)}{Var\left(X_{1+2}\right)} .$$

Thus we may write the reliability of the composite test using observed score moments.

An equivalent result was given by Rulon (1939). Define the *observed difference score* X_{1-2} as

$$X_{1-2} = X_1 - X_2 \ .$$

If we continue to assume that X_1 and X_2 are essentially tau equivalent, then the true difference score T_{1-2} is equal to a constant $(\mu_1 - \mu_2)$ and we have

$$Var(X_{1-2}) = \sigma_{E1}^2 + \sigma_{E2}^2 \ .$$

In other words, the difference score between two tau equivalent tests has zero reliability. This allows us to write

$$Rel(X_{1+2}) = 1 - \frac{Var(X_{1-2})}{Var(X_{1+2})}.$$

Another equivalent version of this result is given by

$$Rel(X_{1+2}) = 2\left[1 - \frac{Var(X_1) + Var(X_2)}{Var(X_{1+2})}\right].$$

We will return to a more general version of this expression (known as coefficient alpha) in the next section.

Now suppose X_1 and X_2 are (essentially) parallel tests. In this case, we may write the variance of X_{1+2} as

$$Var(X_{1+2}) = Var(X_1) + 2Cov(X_1, X_2) + Var(X_2) = 2Var(X_1)\left[1 + Rel(X_1)\right].$$

This lets us write the reliability of X_{1+2} as

$$Rel\left(X_{1+2}\right)=\frac{2Corr\left(X_1,X_2\right)}{1+Corr\left(X_1,X_2\right)}=\frac{2Rel\left(X_1\right)}{1+Rel\left(X_1\right)}.$$

This result is known as the *Spearman-Brown formula*. (See, for instance, Lord & Novick, 1968, p. 84.) A generalization of this formula to the case where a test score X_{1++k} is obtained as the sum of *k* parallel test scores is given by

$$Rel\left(X_{1++k}\right)=\frac{kRel\left(X_1\right)}{1+\left(k-1\right)Rel\left(X_1\right)}.$$

We should emphasize that the Spearman-Brown results only hold for composite tests made up of essentially parallel component tests. In other words, $X_1, X_2, ..., X_k$ must all be essentially parallel, i.e. have the same true scores. except for constant differences, and equal error variances. In the next section, we consider a general theory for multiple tests.

Multivariate Classical Test Theory

The purpose of this section is to generalize the formal framework provided for classical test theory to the case where we have many tests. Specifically, we continue to consider a population of people from which we may sample the people randomly. Now, instead of each person having an observed score on a single test, we imagine that they each have an observed score on each of *k* tests. For convenience, we may use vector notation to represent the associated random vector of observed scores resulting from random sampling of people from the population:

$$\mathbf{X}'=\left(X_1,...,X_k\right).$$

Similarly, we imagine a random vector of true scores associated with the observed scores in a sense to be made explicit:

$$\mathbf{T}' = \left(T_1, ..., T_k\right).$$

We assume that the conditional distribution of **X** given **T** is *k*-variate Normal:

$$\mathbf{X}\left|\mathbf{T}\right.: \; N_k\left(\mathbf{T}; \boldsymbol{\Sigma}_E\right) .$$

The conditional covariance matrix $\boldsymbol{\Sigma}_E$ is assumed to be diagonal, with the j^{th} diagonal element denoted by σ_{Ej}^2 . For a multivariate Normal distribution, this assumption implies conditional independence among the observed scores, given the true scores. It also implies the following univariate conditional distributions for X_j given T_j :

$$X_j \left| T_j \sim N\left(T_j, \sigma_{Ej}^2\right) \right. .$$

Finally, we assume that the distribution of the true score vector **T** is a general *k*-variate Normal:

$$\mathbf{T} \sim N_k\left(\boldsymbol{\mu}_T; \boldsymbol{\Sigma}_T\right) .$$

The j^{th} element of $\boldsymbol{\mu}_T$ is denoted by μ_{Tj}. The diagonal elements of $\boldsymbol{\Sigma}_T$ are the variances of the true scores. The j^{th} diagonal element is denoted by σ_{Tj}^2. The off-diagonal elements of $\boldsymbol{\Sigma}_T$ are the covariances among the true scores. The covariance between T_j and $T_{j'}$ is denoted by $\sigma_{Tjj'}$.

Combining the two distributional assumptions, we may obtain the distribution for **X**:

$$\mathbf{X} \sim N_k\left(\boldsymbol{\mu}_X; \boldsymbol{\Sigma}_X\right) .$$

Here we have

$$\boldsymbol{\mu}_X = \boldsymbol{\mu}_T \ .$$

The diagonal elements of $\boldsymbol{\Sigma}_X$ are the variances of the observed scores. They may be written as

$$\sigma^2_{Xj} = \sigma^2_{Tj} + \sigma^2_{Ej} \ .$$

The off-diagonal elements are the observed score covariances. They may be written as

$$\sigma_{Xjj'} = \sigma_{Tjj'} \ .$$

Using matrix notation, we may write

$$\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_T + \boldsymbol{\Sigma}_E \ .$$

Reliability for Composite Observed Scores

Now consider a composite observed score random variable defined as

$$X_{1++k} = \mathbf{1}'_k \mathbf{X} = \sum_{j=1}^{k} X_j \ .$$

(Here $\mathbf{1}_k$ denotes a vector with *k* elements, all equal to 1.)

X_{1++k} has a corresponding composite true score random variable given by

$$T_{1++k} = \mathbf{1}'_k \mathbf{T} = \sum_{j=1}^{k} T_j \ .$$

The conditional distribution of X_{1++k} given T_{1++k} is Normal:

$$X_{1++k} \left| T_{1++k} \sim N\left(T_{1++k}, \sum_{j=1}^{k} \sigma_{Ej}^2 \right) \right. .$$

The distribution of T_{1++k} is also Normal:

$$T_{1++k} \sim N\left(\sum_{j=1}^{k} \mu_{Tj}, \sum_{j=1}^{k} \sigma_{Tj}^2 + \sum_{j \neq j'}^{k} \sigma_{Tjj'} \right) .$$

Using matrix notation, we may write this as

$$\mathbf{1}_k' \mathbf{T} \sim N\left(\mathbf{1}_k' \boldsymbol{\mu}_T, \mathbf{1}_k' \boldsymbol{\Sigma}_T \mathbf{1}_k \right) .$$

Similarly, we may write

$$\mathbf{1}_k' \mathbf{X} \sim N\left(\mathbf{1}_k' \boldsymbol{\mu}_X, \mathbf{1}_k' \boldsymbol{\Sigma}_X \mathbf{1}_k \right).$$

These results allow us to write the reliability of X_{1++k} as

$$Rel\left(X_{1++k} \right) = Corr\left(X_{1++k}, T_{1++k} \right)^2 = \frac{Var\left(T_{1++k} \right)}{Var\left(X_{1++k} \right)} = \frac{\sum_{j=1}^{k} \sigma_{Tj}^2 + \sum_{j \neq j'}^{k} \sigma_{Tjj'}}{\sum_{j=1}^{k} \sigma_{Tj}^2 + \sum_{j=1}^{k} \sigma_{Ej}^2 + \sum_{j \neq j'}^{k} \sigma_{Tjj'}} .$$

We may write this expression as

$$Rel\left(X_{1++k} \right) = 1 - \frac{\sum_{j=1}^{k} \sigma_{Ej}^2}{Var\left(X_{1++k} \right)} .$$

It is always true that

$$\sigma_{Xj}^2 \geq \sigma_{Ej}^2 .$$

Consequently, we may write

$$Rel\left(X_{1++k}\right) \geq 1 - \frac{\sum_{j=1}^{k} \sigma_{Xj}^2}{Var\left(X_{1++k}\right)} .$$

Thus the right hand side of this expression is a general lower bound for the reliability of a composite score.

Essentially Tau Equivalent Component Scores

We say that the component tests with observed scores X_j and true scores T_j are essentially tau equivalent if there exists a random variable *T* (with mean 0 and variance denoted by σ_T^2) and constants μ_j such that

$$T_j = T + \mu_j \ .$$

In this case, the composite true score is given by

$$T_{1++j} = kT + \sum_{j=1}^{k} \mu_j \ .$$

Note that

$$\sigma_{Tj}^2 = \sigma_{Tjj'} = \sigma_T^2 \ .$$

We may write the variance of T_{1++k} as

$$Var\left(T_{1++k}\right) = k^2 \sigma_T^2 \ .$$

The variance of X_{1++k} becomes

$$Var\left(X_{1++k}\right) = k^2 \sigma_T^2 + \sum_{j=1}^{k} \sigma_{Ej}^2 \ .$$

Thus the reliability of a composite test made up of essentially tau equivalent components is given by

$$Rel\left(X_{1++k}\right)=\frac{Var\left(T_{1++k}\right)}{Var\left(X_{1++k}\right)}=\frac{k^2\sigma_T^2}{k^2\sigma_T^2+\sum_{j=1}^{k}\sigma_{Ej}^2}.$$

Now let's evaluate the right hand side of the inequality given in the previous section for the case of essentially tau equivalent components:

$$1-\frac{\sum_{j=1}^{k}\sigma_{Xj}^2}{Var\left(X_{1++k}\right)}=1-\frac{k\sigma_T^2+\sum_{j=1}^{k}\sigma_{Ej}^2}{k^2\sigma_T^2+\sum_{j=1}^{k}\sigma_{Ej}^2}=\frac{\left(k^2-k\right)\sigma_T^2}{k^2\sigma_T^2+\sum_{j=1}^{k}\sigma_{Ej}^2}=\left(\frac{k^2-k}{k^2}\right)Rel\left(X_{1++k}\right).$$

This allows us to write the following equality:

$$Rel\left(X_{1++k}\right)=\left(\frac{k}{k-1}\right)\left[1-\frac{\sum_{j=1}^{k}\sigma_{Xj}^2}{Var\left(X_{1++k}\right)}\right].$$

The expression on the right hand side of this equation is commonly referred to as *coefficient alpha* (Cronbach, 1951). It is equal to the reliability of a composite test if and only if the component tests are essentially tau equivalent.

However, it has an additional, general property. For any set of components for which the assumptions of multivariate classical test theory hold, we may write the following inequality for the reliability of the composite:

$$Rel\left(X_{1++k}\right) \geq \left(\frac{k}{k-1}\right)\left[1-\frac{\sum_{j=1}^{k}\sigma_{Xj}^{2}}{Var\left(X_{1++k}\right)}\right].$$

In other words, coefficient alpha gives a *lower bound* for the reliability of any composite test, and *equals* the reliability if and only if the components are essentially tau equivalent.

Thus it may be useful to have a notation for alpha that would apply for any composite test, not just ones with essentially tau equivalent components:

$$Alpha\left(X_{1++k}\right) = \left(\frac{k}{k-1}\right)\left[1-\frac{\sum_{j=1}^{k}\sigma_{Xj}^{2}}{Var\left(X_{1++k}\right)}\right].$$

While it is a better lower bound than the one given in the previous section, since $k/(k-1) > 1$, coefficient alpha is not, in general, a greatest lower bound for the reliability of a composite test. Nonetheless, this "conservative" property is an attractive one. (Note that there is lots of research on lower bounds for reliability.)

Estimation

Everything so far in this introduction has referred only to population quantities. The question naturally arises: How do we estimate the various quantities introduced here based on samples of test scores? The simple answer is that we typically use sample quantities that are unbiased estimates of the population quantities of interest. In the context of the multivariate Normal model that we have been working with, the first and second moments (means, variances and covariances) are the only population quantities needed to specify the model. Thus we may use

sample means of observed scores to estimate the corresponding population means, and we may use sample variances and covariances for the observed scores to estimate the corresponding population quantities.

Thus, if we have an independent random sample of N vectors of k observed scores given by $\mathbf{x}_i' = (x_{i1}, ..., x_{ik})$ for $i = 1, ..., N$, we may estimate the mean observed score for X_j using

$$\hat{\mu}_j = \left(\frac{1}{N}\right)\sum_{i=1}^{N} x_{ij} .$$

Similarly, we may estimate the observed score variance for X_j using

$$\hat{\sigma}_{Xj}^2 = \left(\frac{1}{N-1}\right)\sum_{i=1}^{N}\left(x_{ij} - \hat{\mu}_j\right)^2 .$$

The composite observed score for test taker i in our sample is given by

$$x_{i,1++k} = \sum_{j=1}^{k} x_{ij} .$$

The estimated mean for the composite observed score is

$$Est\left[Ave\left(X_{1++k}\right)\right] = \left(\frac{1}{N}\right)\sum_{i=1}^{N} x_{i,1++k} \quad .$$

Similarly, its estimated variance is

$$Est\left[Var\left(X_{1++k}\right)\right] = \left(\frac{1}{N-1}\right)\sum_{i=1}^{N}\left\{x_{i,1++k} - Est\left[Ave\left(X_{1++k}\right)\right]\right\}^2 .$$

We may put these estimates together, substituting estimates for the corresponding population quantities to give, for instance,

$$Est\left[Alpha\left(X_{1++k}\right)\right]=\left(\frac{k}{k-1}\right)\left[1-\frac{\sum_{j=1}^{k}\hat{\sigma}_{Xj}^{2}}{Est\left[Var\left(X_{1++k}\right)\right]}\right].$$

This is the formula customarily used to estimate coefficient alpha. It should be noted that, although the numerator and denominator of the ratio in this estimator are both unbiased estimators for the corresponding population quantities, the sample ratio will not be unbiased as an estimator for the population ratio. Moreover, although the population formula for alpha gives a general lower bound for the reliability of X_{1++k}, the sample estimator given above does not give a lower bound for the population reliability. Nonetheless, the estimator is widely used and may be useful in a variety of settings.

Applications

There are several practical applications for the reliability of a test beyond the basic one of describing the strength of association between observed and true scores for the test. When test scores are reported, it is often of interest to provide a corresponding *standard error of measurement*, σ_E in our notation. We may use the expression given at the beginning of this introduction, namely

$$Rel\left(X\right)=1-\frac{\sigma_E^2}{\sigma_X^2}.$$

Solving for σ_E, we obtain

$$\sigma_E=\sigma_X\sqrt{1-Rel\left(X\right)}\ .$$

There is (at least) one caution that should be given regarding this use of reliability. The reliability and the observed score standard deviation must both be associated with the same population for this formula to be valid. It is not appropriate to take a value for the reliability of a test based on one population and apply it to another population with a different standard deviation for the observed scores. To phrase this another way, the reliability of a test is specific to the population on which it is based. It cannot be evaluated for one population and applied to another population.

Suppose we want to estimate true scores based on observed scores. In this case, we want to consider the conditional distribution of *T* given *X*:

$$T|X \sim N\left\{\left[Rel(X)\right]X + \left[1-Rel(X)\right]\mu_X, \sigma_X^2 Rel(X)\left[1-Rel(X)\right]\right\} .$$

From a Bayesian perspective, this is the posterior distribution for *T* given *X*. The following formula for the posterior mean is due to Kelley:

$$\hat{T}(X) = \left[Rel(X)\right]X + \left[1-Rel(X)\right]\mu_X .$$

The corresponding posterior standard deviation is called the *standard error of estimation*:

$$\sqrt{Var(T|X)} = \sigma_X \sqrt{Rel(X)\left[1-Rel(X)\right]} .$$

Finally, suppose we have two tests, X_1 and X_2, with a correlation given by

$$Corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sigma_{X1}\sigma_{X2}} .$$

Suppose we would like to have an expression for the correlation between the true scores for the two tests. Recalling that the observed score and true score covariances are equal, we may write

$$Corr\left(T_1,T_2\right)=\frac{Cov\left(T_1,T_2\right)}{\sigma_{T1}\sigma_{T2}}=\frac{Cov\left(X_1,X_2\right)}{\sigma_{X1}\sigma_{X2}}\left(\frac{\sigma_{X1}\sigma_{X2}}{\sigma_{T1}\sigma_{T2}}\right)=\frac{Corr\left(X_1,X_2\right)}{\sqrt{Rel\left(X_1\right)Rel\left(X_2\right)}}.$$

This is sometimes called a *disattenuation formula*. We could say that the formula "corrects" the correlation between the observed scores of two tests for the attenuation due to their errors of measurement.

References

Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577-601.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.

Rulon (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, *9*, 99-103.