Eğitim ve Bilim
2009, Cilt 34, Sayı 152

Education and Science
2009, Vol. 34, No 152

Misuses of KR-20 and Cronbach's Alpha Reliability Coefficients

KR-20 ve Cronbach Alfa Katsayılarının Yanlış Kullanımları

Şeref TAN*
Uludağ Üniversitesi

Öz

Bu çalışmada, iç tutarlılık kestirimi olarak KR-20 ve Cronbach Alfa güvenirlik katsayılarının yanlış kullanımları gösterilmektedir. Tek boyutluluk sayıtlısı ihlal edildiğinde test puanlarının iç tutarlılık katsayılarındaki farklılaşmaları görmek için iki gerçek veri seti kullanılmıştır: "0-1" puanlama yönteminin kullanıldığı (KR-20) veri seti ve dereceleme yöntemiyle puanlamanın kullanıldığı (alfa) veri seti. Araştırma bulguları, iç tutarlılık katsayılarının hesaplanmasında, tek boyutluluk sağlanmadığı durumlarda bile çok yüksek ama araştırmacıları yanlış yönlendiren KR-20 ve Cronbach's alfa iç tutarlılık katsayılarının elde edilebileceği göstermiştir. Verilerin gerekli sayıltıları sağlamadığı durumlarda, güvenirliğin bir göstergesi olarak ölçmenin standart hatasının kullanımının da benzer problemlere sebep olduğu gösterilmiştir. Çalışmanın sonunda, bir ölçme aracının iç tutarlık katsayısının rapor edilmesindeki yanlış kullanımlar ve bu yanlış kullanımlardan kaçınma yollarına yönelik olarak öneriler sunulmuştur.

Anahtar Sözcükler: KR-20, Alfa güvenirlik katsayısı, teta katsayısı, iç-tutarlık kestirimlerinin yanlış kullanımı, tabakalanmış alfa katsayısı.

Abstract

In this study, misuses of KR-20 and Cronbach's alpha reliability coefficients, used as internal consistency estimates, are illustrated. Two real data sets were used, dichotomously scored, KR-20, data set and polytomously scored, alpha, data set, to see variations in internal consistency coefficients when the unidimensionality assumption is violated. It is shown that a very high, but misleading internal consistency coefficient, KR-20 or alpha, can be obtained even when the unidimensionality assumption is violated. It is also shown that using the standard error of measurement as an indicator of reliability results in similar problems. Finally, misuses and the ways of avoiding those misuses of reporting internal consistency of a scale are suggested and briefly illustrated.

Keywords: KR-20, alpha reliability coefficient, coefficient theta, misuses of internal consistency estimations, stratified alpha coefficient.

* Yrd. Doç. Dr. Şeref TAN, Uludağ Üniversitesi, Eğitim Fakültesi Eğitim Bilimleri Bölümü

Introduction

In educational and psychological research, it is common to use a measurement instrument to collect some data. This instrument could be an achievement test or an attitude scale or some other scales. When a measurement instrument is used in a research, some attributes of this scale must be calculated and reported, such as reliability and validity. All tests contain some error. This is true for tests in the physical sciences and for educational and psychological tests (Rudner & Schafer, 2001: 1). As Cronbach (2004: 2) pointed out: "where the accuracy of a measurement is important, whether for scientific or practical purposes, the investigator should evaluate how much random error effects the measurement." Reliability, along with validity, is central issues in all scientific measurement (Gaffney, 1997: 1). In research reports, reliability and standard error estimations are being reported for measurement instruments used in a study. In classical test theory, alpha or KR-20 reliability coefficient is used to estimate the reliability of obtained scores for a group and a standard error of measurement score estimates is given for a group, these estimates are changes for different groups. In item response theory, reliability estimates do not depend on groups, and measurement errors are estimated for every different score. In item response theory, information functions, and precision at each point are estimated. However, using classical test theory to report reliability is still very common and easier way.

Because of using internal consistency coefficients is very common, there are some studies that give criteria to interpret reliability and validity coefficients; a sample of this is given below to interpret a reliability coefficient as an internal consistency index:

"The range of reliability measures are rated as follows:

a) Less than 0.50, the reliability is low,

b) Between 0.50 and 0.80 the reliability is moderate and

c) Greater than 0.80, the reliability is high" (Salvucci, Walter, Conley, Fink, & Saba, 1997: 115).

As Wainer and Thissen (1996) made it clear that the answer of "how reliable a test be?" is as reliable as possible.

When an investigator gets an insufficient reliability or validity coefficient, s/he tries to get rid of this obstacle by eliminating some items in the measurement instrument. Also, most of the computer programs give suggestions to eliminate some items to get a higher internal consistency coefficient. When a researcher has no aim on test development, s/he still uses to eliminate some items to increase the reliability estimates of obtained scores. However, this may cause some more serious problems in terms of the scale reliability and validity. As it is known, there are some applications to improve the internal consistency of a measurement instrument and there are also some criticisms of these kinds of applications. Enders and Bandalos (1999) investigated the degree to which Cronbach's alpha coefficient is affected by the inclusion of items with different distribution shapes within a unidimensional scale. Their research showed that when interitem correlations were high and the number of response categories was large, reliability of a scale is least affected and they pointed out that the practice of deleting items from a scale on the basis of distributional nonnormality may adversely affect scale validity with no appreciable increase in reliability. Nunnally and Bernstein (1994) pointed out that dropping some items that have low item-total correlation or small standard deviation, and high skewness or kurtosis to improve the internal consistency of measurement instruments may reduce validity. Hence, excluding some items from the scale to improve the internal consistency may severely cause a decrease in content and construct validity and such information is missing in most of educational and psychological studies.

One way to compute the reliability of a scale is to use an internal consistency coefficient. Often, computer programs can be used in an inappropriate way to estimate the internal consistency of test scores. Erkuş (1999) indicated that there are some misuses of statistical package programs. One of these misuses is inaccurately calculating internal consistency coefficients (KR-20 and Cronbach's alpha) by using SPSS. Item-total correlation coefficients can be miscalculated by applying an inappropriate item-total score correlation technique by SPSS such as using Pearson correlation technique when Point biserial correlation technique is appropriate for measurement scale type of the data. As Erkuş (2003: 69) pointed out internal consistency coefficients give information about consistency among items of a test. Internal consistency focuses on the degree to which the individual items are correlated with each other and is thus often called homogeneity (Rudner & Schafer, 2001: 4). In classical test theory, KR-20 and alpha coefficients are used to calculate the internal consistency of a scale as a reliability index. KR-20 reliability formula is used to calculate the internal consistency of an achievement test when a test measures a unidimensional trait and if it is scored dichotomously. Cronbach's alpha reliability coefficient is used to calculate the internal consistency of an attitude scale or an achievement test when a test measures a unidimensional trait and if it is scored by ratings, such as "1 to 5" scoring procedure. Cronbach's alpha is generally considered the most appropriate type of reliability for attitude instruments and other scales that contain a range of possible answers for each item (McMillan, 1992). As Feldt and Qualls (1996: 277) pointed out, "a necessary and sufficient condition of coefficient alpha as an estimate of test reliability is that the part scores be essentially tau equivalent. This condition implies that the test is unidimensional in the factor analytic sense, and all parts must measure the same unitary trait or ability." Both in KR-20 and alpha coefficients "the unidimensionality" assumption of the scale must be met. However, even the unidimensionality assumption is violated; it is a very common situation to get a very high KR-20 or alpha reliability coefficient. Even when you measure completely different factors by a measurement instrument, still you may calculate a very high KR-20 or Cronbach's alpha reliability coefficient. In this case, only reporting the KR-20 or alpha coefficient as a consistency coefficient is not a valid measure and it is not right to do so. However, this situation is usually seen in some important amount of research reports. As Feldt and Qualls (1996) pointed out that, in the use of achievement and intelligence tests, the tau equivalence or a single factor assumption would be appear to be an assumption that could severely limit the application of coefficient alpha and it would appear that the tau equivalence or single-factor assumption is violated to some degree in such tests. In this study, it is shown that, a very high, but misleading internal consistency coefficient, KR-20 or alpha, can be obtained even when the unidimensionality assumption is violated by using real data sets.

As it is emphasized above, there are some misuses of internal consistency coefficients in some research reports. Some examples of those misuses are as follows, (a) lack of reporting the reliability of scores gathered on the actual study data, (b) not considering the factorial structure of the data, (c) reporting the standard error of measurement as a reliability indication when the unidimensionality assumption is violated, and (d) using a high KR-20 or Cronbach's alpha coefficient as an indicator for homogeneity. In this study, the ways of avoiding those misuses and reporting internal consistency of a scale more accurately are suggested and briefly illustrated.

Purposes of the Research

The main purpose of this study is to show some misuses of internal consistency coefficients by providing some suggestions about the ways of avoiding those misuses of reporting internal consistency coefficients. This purpose can be written in two related sub purposes, given below,

(a) to see the variations in internal consistency coefficients when unidimensionality assumption is violated. Two real data sets were used for this purpose, the dichotomously scored data set (by applying a KPSS test, consists of 7 subtests) and polytomously scored data set (a Likert-type attitude scale, and

(b) to determine misuses and the ways of avoiding those misuses of reporting internal consistency of a scale are investigated.

Method

Two real data sets were used to compute and interpret the KR-20 and Cronbach's alpha coefficients when the unidimensionality assumption is violated. As it is well known KR-20 and alpha coefficients are same coefficients; however, it calls KR-20 when it is used for dichotomously scored items and it calls alpha when it is used for polytomously scored items. "Alpha is a general version of the Kuder-Richardson coefficient of equivalence. It is a general version because the Kuder-Richardson coefficient applies only to dichotomous items, whereas alpha applies to any set of items regardless of the response scale" (Cortina, 1993: 99). In this study, both dichotomously and polytomously scoring procedures are used. The data sets were used to calculate the KR-20 and alpha reliability coefficients are given below.

KR-20 Data Set

A real data set was used in this study to calculate the reliability coefficients of dichotomously scored achievement test and its subtests, by using KR-20 as internal consistency estimates. The results of 2005 KPSS test, which is a nationwide test applied by a private publication, was used to collect data. The KPSS test is a personal selection test, and it is used to select teachers to government schools in Turkey. The KPSS test is divided into three subtests; these subtests are as follows:

The subtest 1: The general ability test: General ability test consists of two subtests;

a) the verbal ability test and

b) the mathematical ability test.

The subtest 2: The general culture test: This test includes questions about some cultural topics such as history, and geography.

The subtest 3: The educational sciences test: This test includes questions about;

a) the program development and teaching strategies,

b) the developmental and learning psychology,

c) the measurement and assessment in education and

d) the educational guidance.

The subtests and their number of questions are summarized in Table 1, below:

Table 1.
Tests and Their Subtests and Number of Questions in a KPSS Test

Test or Subtest	Number of question
1. The general ability test	60
a) The verbal ability tests	30
b) The mathematical ability tests	30
2. The general culture test	60
3. The educational sciences test	120
a) The curriculum development and teaching strategies test	42
b) The developmental and learning psychology test	42
c) The measurement and assessment in education test	18
d) The educational guidance test	18

A practice test of KPSS was applied to the 6383 teacher candidates in 2005. The scores of this test were used to calculate KR-20 reliability coefficients as internal consistency measures, for dichotomously scored items.

Alpha Data Set

A real data set was used to calculate the reliability coefficients of polytomously scored attitude test by using Cronbach's alpha reliability coefficient as internal consistency estimate. The data was collected by applying a 37-item Likert-type scale. The Likert-type scale was applied to collect the opinions of the people of Evenston, USA about the quality of the education given by a private school called Evenston High School. This scale was a 5-point Likert-type attitude scale and it was designed to measure one trait, attitude towards the quality of the education given by Evenston High School. The scale was applied to 1274 people by a research center. This Likert-type test scores were used to calculate Cronbach's alpha reliability coefficient as an internal consistency measure for polytomously scored items.

Results

In addressing the purposes of this study, several statistical procedures were used to analyze the data sets. In the analysis, the reliability coefficients of test and subtests and their factorial structures were computed. KR-20 formula was used to calculate the reliability coefficients for the KPSS test data. Cronbach's alpha reliability coefficient was used to calculate the reliability coefficient for the Likert-type scale data. To see dimensionality of the tests exploratory factor analysis were performed by using SPSS (16. version). In factor analysis, "the principle axis method" was utilized as a factor extraction method.

The Factorial Structure and Internal Consistency Coefficients of the KPSS Test and Its Subtests

All possible internal consistency coefficients were calculated for the KR-20 data set by using the SPSS (16. version). In addition to the calculation of KR-20, factor analyses were performed for the related test or subtests. The tests, their factorial structures, the standard error of measurements on K and 100 score scale and KR-20 reliability coefficients are given in Table 2, below:

Table 2.
Tests and Their KR-20 Coefficients, Standard Error of Measurements on K and 100 Score Scale and Related Factorial Structure of the KPSS Tests (N=6383)

Test and Subtest	Number of Question	Number of Factor by Using Kaiser criterion	Accounted Variance by the Factors	KR-20	Se	Se%
KPSS	237	71	48.36%	0.92	6.43	2.71
1. The General ability test	60	14	45.29%	0.88	2.97	4.95
a. The verbal ability tests	30	10	40.18%	0.55	1.80	5.99
b. The mathematical ability tests	30	4	49.75%	0.94	1.98	6.60
2. The general culture test	58*	11	34.81%	0.87	3.24	5.59
3. The educational sciences test	119*	36	44.14%	0.89	4.43	3.72
a) The curriculum development and teaching strategies test	41	13	40.78%	0.66	2.66	6.49
b) The developmental and learning psychology test	42	9	35.59%	0.79	2.59	6.16
c) The measurement and assessment in education test	18	4	35.72%	0.71	1.66	9.21
d) The educational guidance test	18	4	32.61%	0.58	1.72	9.54

*: Two questions of the general culture test and 1 question of the educational sciences test were cancelled by the testing company.

As can be seen in Table 2, the reliability of KPSS test is 0.92 and it includes 7 subtests. Even though calculating an internal consistency coefficient(KR-20) for the entire test(KPSS), which includes 7 subtests, is not right thing to do, it is calculated purposefully to get attention that a very high, but misleading internal consistency coefficient, KR-20 or alpha, can be obtained even when the unidimensional assumption is violated. KR-20 reliability coefficients for the subtests varied from 0.55 to 0.94. The number of factors (when Kaiser Criterion, eigenvalues greater than unity, is used) varied from 4 to 13 for the subtests. "The mathematical ability test", "the measurement and assessment in education test" and "the educational guidance test" yielded 4 factors while "the curriculum development and teaching strategies test" yielded 13 factors. So, all of the subtests consists of multiple factors.

The KR-20 reliability coefficient for the entire scale (KPSS test itself) was 0.92, suggesting that the items are internally consistent based on this data set and a single composite score is reasonably reliable. However, this KPSS test consists of 71 factors with 237 questions and 7 subtests. The KR-20 reliability coefficient for "the general culture test" was 0.87, and it consists of 11 factors. As it is seen in Table 2, the four subtests of "the educational sciences test" have lower reliability coefficients (0.58, 0.66, 0.71, and 0.79) than the reliability of entire "educational sciences test (0.89)". When the data consists of multiple factors or subtests, getting composite scores using these subtests should give lower reliability estimates. This situation can not be explained due to increase in number of items, as it is well known when the number of item is increases the estimated reliability increases too. In this situation when dimensionality or heterogeneity increases, the KR-20 coefficient increases too. This result, conflicts with the nature of internal consistency estimates.

There is a different situation for the general ability test. "The general ability test", which contains "the verbal ability" and "the mathematical ability" subtests, has lower reliability coefficient than "the mathematical ability" subtest and higher reliability coefficient than "the

verbal ability" subtest. It is very well known that verbal ability and mathematical ability are different traits. Getting composite scores of these two subtests should give lower reliability estimates. This situation is partially met here.

As it is seen in Table 2, none of the tests or subtests meets the unidimensionality assumption and least these tests are not homogeneous. The number of factors of the tests or subtests varied from 4 to 71. The accounted variances by the related factors varied from 32.61% to 48.36%. However, very high KR-20 reliability coefficients, varied from 0.58 to 0.94, have been found.

Last two columns in Table 2 shows the standard error of measurements; (Se) on "0 point to K points" scale, where K is the number of items for related tests, and (Se%) "0 point to 100 points" score scales. The standard error of measurement values of scales are interpreted considering the maximum score, which is the number of questions, for the alpha data set. The item numbers or the maximum scores of the tests are different, which makes standard error of measurement values incomparable. Thus, the values of the standard error of measurement are altered to "0 to 100 points" scale to make comparisons possible. As an example; the standard error of measurement for the verbal ability test is 1.80 on "0 to 30 points" scale score (number of items, K is 30, in this test) and the standard error of measurement is 5.99 on "0 to 100 points" scale score. The standard errors of measurements on same scale for the tests are presented in the last column (Se%) in Table 2.

The Internal Consistency Coefficient of the Likert-type Scale

The internal consistency coefficient of the Likert-type scale was calculated by using SPSS (16. version). In addition to the calculation of Cronbach's alpha reliability coefficient, a factor analysis was performed for the attitude scale. Because the attitude scale is designed as unidimensional only one factor analysis performed to see factorial structure of the polytomously scored data. The number of questions, the number of factors, accounted variance by the factors and the Cronbach's alpha reliability coefficients of the attitude scale were calculated. The results are summarized in Table 3, below:

Table 3.
The Alpha Reliability Coefficient and the Factorial Structure of the Attitude Scale (N=1274)

Scale	Number of Question (item)	Number of Factor by Using Kaiser criterion	Accounted Variance by the Factors	Alpha Reliability
Likert-type Scale	37	7	55.3%	0.93

As can be seen in Table 3, alpha data set consists of 7 factors, by using Kaiser Criterion with 55.3% accounted variance and a 0.93 Cronbach's alpha reliability coefficient. Same findings as in KR-20 data set have been found for the Cronbach's alpha reliability coefficient too. When you get composite scores by summing up 7 subtests' scores, you can have very high reliability estimates. Hence, those findings are identical with the finding of alpha data set.

Discussion and Conclusion

"Psychological research involving scale construction has been hindered considerably by a widespread lack of understanding of coefficient alpha and reliability theory in general (Cortina, 1993: 98)." The misuses of KR-20 and Cronbach's alpha coefficients, discussions, conclusions, and suggestions for the potential solutions to avoid those misuses are presented in this section.

Misuse 1 : Lack of reporting the reliability of scores gathered on the actual study

When a scale is used in a study, the reliability of the scores should be reported. Reliability of the scores may have different values for different groups for the same instrument. Any measurement device that is reliable in one setting or for one purpose may be unreliable in another setting or for a different purpose (Vockell & Asher, 1995). As Rudner and Schafer (2001: 2) pointed out, reliability is a joint characteristic of a test and examinee group, not just a characteristic of a test. Indeed, reliability of any test varies from group to group. Therefore, the better research studies will report the reliability for their sample as well as the reliability for norming groups as presented by the test publisher.

When researchers use very reliable and valid scale, scores gathered from this scale for the group used in the study may not be reliable because of error due to application and scoring conditions. For this reason, whenever a measurement instrument is used in a group, the reliability of scores should be calculated and reported for this group as well. Without doing that, only reporting the reliability of norming groups as presented by the test publisher is an example for misuse of the reliability coefficients.

Misuse 2 : Not considering the factorial structure of the data

If a scale has enough items (i.e., more than 20), then it can have an alpha of greater than 0.70 even when the correlation among items is very small (Cortina, 1993: 102). Cortina's study showed that if a scale has more than 14 items, then it will have an alpha of 0.70 or better even if it consists of two orthogonal dimensions with modest (i.e., 0.30) item intercorrelations (Cortina, 1993: 102). Another finding that Cortina's study showed, given a sufficient number of items, a scale can have a reasonable alpha even if it contains three orthogonal dimensions. If the three dimensions themselves have average item intercorrelations of 0.70 or better and are somewhat correlated with each other, then the overall scale has an alpha of 0.80 or better(Cortina, 1993: 103).

Both data sets showed that when a scale has multiple factors (even 71 factors), still a high reliability coefficient can be calculated by using a KR-20 or Cronbach's alpha coefficient as an internal consistency coefficient. In this situation, items can be interpreted as internally consistent, but it is not exactly true. Only reporting a KR-20 or Cronbach's alpha coefficient as a consistency coefficient may mislead researchers, and this is an example of misuse of KR-20 and Cronbach's alpha coefficients. So, when an internal consistency coefficient is needed to be reported, presenting only a KR-20 or Cronbach's alpha coefficient is not enough. Schmitt emphasized the importance of this subject too. "Both intercorrelations and alpha must be reported if the reader is to be adequately informed about the obtained results" (Schmitt, 1996: 353)

When a KR-20 or Cronbach's alpha coefficient is reported, the factorial structure of the data should be reported along with it. The appropriateness of a KR-20 or Cronbach's alpha coefficient must be discussed not only by the size of it but also the appropriateness of the factorial structure of the data. As Erkuş (2003: 69) emphasized that when a test consists of some homogeneous subtests, internal consistency coefficients should be calculated for each subtest.

One way of reporting the alpha reliability estimate when scale has multiple factors is using stratified alpha coefficient proposed by Cronbach, Schonemann, and Brennan (1965). Feldt and Qualls (1996: 278-9) introduced the formula of stratified alpha is as follows:

Strat$\alpha\ \rho_{xx'} = 1 - \frac{\sum^{C} \sigma^2_{xj}(1-\alpha\rho_{xjxj'})}{\sigma^2_{Xtot}}$ where X_j is a cluster score, $\alpha\rho_{xjxj'}$ is the alpha coefficient for cluster j, C is the number of clusters, and X_{tot} is the total score obtained by $X_1 + ... + X_C$. Using stratified alpha coefficient is a better way to report reliability estimates when test measures multiple factors as it is in KR-20 data set. Treating the KR-20 data set as 7 factors (each subtest of KPSS can be treated as a factor, at least theoretically) then using stratified alpha coefficient is a better way to report reliability of the KPSS scores. Considering Likert-type attitude scale data, alpha data set, stratified alpha coefficient can be used if factors are theoretically meaningful and clearly defined; however, this is not the situation for this data set. As Güloğlu and Aydın (2001) indicated that factor analyses may yield multiple factors, by using Kaiser Criterion; however, not all of them may be theoretically meaningful. Feldt and Qualls (1996) concluded that usual formula for coefficient alpha is quite robust with respect to alpha's cardinal assumption. Where bias is substantial, the stratified version of alpha can be accommodate to multidimensionality. For alpha data set, using split half method as an alternative way to estimate internal consistency is a better thing to do.

Using coefficient theta (θ) is an alternative way to report internal consistency estimates for both data sets. Coefficient theta, as an internal consistency estimates, is more appropriate than KR-20 or Cronbach's alpha coefficient when assumptions of KR-20 or Cronbach's alpha coefficients are violated, which is the situation for both data sets of this study. As Carmines and Zeller (1979: 61) pointed out coefficient theta(θ) is based on a principal component analysis of the test items and it differentially weights items that correlate more with each other to make alpha maximum. Kline (2005: 175) introduced coefficient theta formula as it is as follows: Coefficient theta $(\theta) = \frac{N}{N-1}\left(1 - \frac{1}{\lambda}\right)$ where N= the number of items in the test and λ = the eigenvalue of the first principal component.

Another way to report the reliability of a scale is using item response theory if data is appropriate for using an item response theory model. Raykov (1997) recommended a latent variable model for estimating scale variability in case of violation of the assumption of essential tau-equivalence (for component measures) to use alpha coefficient.

Misuse 3: Reporting, the standard error of measurement as a reliability indication when the unidimensionality assumption is violated.

In addition to reliability, the American Psychological Association requires test publishers to report standard error of measurement information for tests offered for public use (Thorndike, 1997). The standard error of measurement and the reliability coefficient are alternative ways of expressing test reliability (Anastasi, 1988). The standard error of measurement tells the probable score range within which an individual's true score may fall, and the range variability of an individual's score is a function of an instrument's reliability (Drummond, 1996). Another reporting way of a KR-20 or Cronbach's alpha coefficient is reporting the standard error of measurement along with them. Reporting standard error of measurement with KR-20 or Cronbach's alpha coefficient gives an opportunity to see the errors of the test scores in term of the score scale. If unidimensionality assumption is not meet as it is in the data sets used in this study, using standard error of measurement may mislead researchers too.

According to Table 2, the smallest standard error of measurement on "0 to 100 point" score scale belongs to the KPSS test which consists of 237 questions and 71 factors which makes standard error of measurement meaningless as it was in their internal consistency estimates, KR-20. As it is very clearly seen in Table 2, when unidimensionality assumption is violated, using standard error of measurement to get idea about the error score is an example of misuse

index too. The value of a standard error of measurement primarily depends on the value of the reliability estimate of a scale such as a KR-20 or an alpha coefficient. If a KR-20 or Cronbach's alpha is not an appropriate measure for a scale, the standard error of measurement is not an appropriate measure for interpreting the reliability or random errors of the scale scores too.

Misuse 4: Using a high KR-20 or Cronbach's alpha coefficient as an indicator for homogeneity.

Gullikson (1962) pointed out that "one approach to the problem of item homogeneity is to make a factor analysis of the inter-item correlations for a test. If there is only one factor, the items are homogeneous. If the analysis reveals more than one common factor, it might be desirable to consider dividing the test into parts, each of which represented a single common factor." As Zimmerman, Zumbo and Lalonde (1993) pointed out that attempts have been made over the years to interpret coefficient alpha in terms of other test properties including item homogeneity and unidimensionality. Zimmerman, Zumbo and Lalonde (1993) also, indicated that reported values of coefficients alpha in test manuals, research studies, and so on, sometimes may be far above and sometimes far below those of the population reliability coefficients due to violation of the assumptions.

As Gullikson(1962) pointed out, reliability coefficients as KR-20 or alpha equations based on only one assumption, that the average covariance between non-parallel items is equal to the average covariance between parallel items. Thorndike (1961) pointed that "the essential assumption in KR-20 coefficient is that the items within one form of a test have as much in common with one another as do the items in that one form with the corresponding items in a parallel or equivalent form. This means that the items in a test are homogeneous in the sense that every item measures the same general factors of ability or personality as do the other." So, what is the connection between alpha and homogeneity? "Alpha is a function of the extent to which items in a test have high communalities and thus low uniquenesses. It is also a function of interrelatedness, although one must remember that this does not imply unidimensionality or homogeneity" (Cortina, 1993: 100).

The magnitude of KR-20 or alpha coefficient is a function of inter-item covariances. The average inter-item covariances determines the value of the KR-20 or alpha coefficient. It is possible, having more inter-item covariances when different subtests are combined. In this situation, you can get a higher reliability estimates for entire test than its subtests. When subtests are correlated or some other factors, not considered to be measured, such as; bais, testing situations, appropriateness of test language, and unclearity of items may cause a common factor in test scores. In this case mean inter-item covariance may increase, then it results to have a higher internal consistency estimates. So, the important question for this particular problem is "getting composite scores by combining the subscales' scores is meaningful in term of theoretical base?" If the answer of this question is "yes" than do not worry about the dimensionality or homogeneity and use KR-20 or alpha coefficient as a reliability estimates, as Feldt and Qualls (1996) recommended. However, if the answer to this question is "no" than using KR-20 or alpha coefficient may be misleading. Feldt and Qualls' explanations are meaningful when the answer of this question is "yes". Their conclusions are as follows: "Because alpha and KR-20 coefficients are commonly called internal consistency coefficients, many researchers have investigated their sensitivity as a measure of item homogeneity, that is, test dimensionality. In general, the results suggested that alpha was not a good index of this characteristic of a test and was subsequently abandoned (Green, Lissitz, & Mulaik, 1977; Hattie, 1985; Raju, 1982; Terwilliger & Lele, 1979). Although coefficient alpha was clearly a poor index of dimensionality, these investigations did suggest that coefficient alpha provided a robust estimation of reliability in the presence of dimensionality violations. Subsequent findings by Hsu and Fan (1995) and Woodruff (1993) also indicate that the

reliability of scores from a factorially complex instrument can be estimated using coefficient alpha (Feldt and Qualls, 1996: 279).

Internal consistency is certainly necessary for homogeneity, but it is not sufficient (Schmitt, 1996: 350). According to Schmitt, "Cronbach (1951) viewed reliability, including internal consistency measures, as the proportion of test variance that was attributable to group and general factors. Specific item variance, or uniqueness, was considered error. Clearly, Cronbach would not treat alpha as a measure of unidimensionality" (Schmitt, 1996: 350). Schmitt also concluded that if alpha is used as "proof" that a set of items have an unambiguous or unidimensional interpretation, the conclusions drawn may or may not be correct (Schmitt, 1996: 350). As Gorsuch made it clear that "most scales are designed to measure just one construct, to be homogeneous. It is often incorrectly assumed that a measure of internal consistency (e.g..coefficient alpha) provides a means to address this question. However, it is easy to design a set of items that measure more than one independent construct and yet produce a coefficient alpha of 0.90 or better when scored as if they measured only one construct. Factor analysis provides a better cleans of examining scale homogeneity" (Gorsuch, 1997: 535).

It is very clear that using a high KR-20 or Cronbach's alpha coefficient as an adequate index of homogeneity is an example to misuse of these reliability coefficients. Leedy (1997) pointed out that the higher the score on Cronbach's alpha then the better the evidence that items on the instrument are measuring the same trait. However, this conclusion may mislead the researchers if researchers just only considered to the magnitude of Cronbach's alpha coefficient. Schmitt (1996: 351) recommended that "when assessing the degree to which a measure is actually unidimensional, an increasingly popular approach in determining the extent of unidimensionality is to test whether the interitem correlation matrix fits a single-factor model" (Joreskog & Sorbom, 1979).

References

Anastasi, A. (1988). *Psychological Testing*. New York: Macmillan.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1995). *Standards for educational and psychological testing*. Washington, DC: Authors.

Carmines, E. G., & Zeller R. A. (1979). *Reliability and Validity Assessment*. Beverly Hills, CA: Sage.

Cortina, J., M. (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology 78*(1), 98-104.

Crocer, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, Inc.

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

Cronbach, L. J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-paralel tests. *Educational and Psychological Measurement*, *25*, 291-312.

Cronbach, L. J. (2004). My current thoughts on Coefficient Alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391-418.

Drummond, R. J. (1996). *Appraised procedures for counselors and helping professionals*. Englewood Cliffs. NJ: Prentice-Hall.

Enders, C. K., & Bandalos, D. L. (1999). The effects of heterogeneous item distributions on reliability. *Applied Measurement in Education*, *12*(2), 133-150.

Erkuş, A. (1999). İstatistik paket programlarını doğru kullanabiliyor muyuz?: Birkaç uyarı. *Türk Psikoloji Bülteni, 5*(12), 14-16.

Erkuş, A. (2003). *Psikometri üzerine yazılar: Ölçme ve psikolojinin tarihsel kökenleri, güvenirlik, geçerlik,madde analizi, tutumlar: Bileşenleri ve ölçülmesi.* Ankara: Türk Psikologlar Derneği Yayınları.

Feldt, L. S., & Qualls, A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education*, *9*(3), 277-286.

Gaffney, P. V. (1997). A test reliability analysis of an abbreviated version of the pupil control ideology form. *Reports-Evaluative*. ED 407 422.

Gorsuch, Richard L. (1997) 'Exploratory Factor Analysis: Its Role in Item Analysis', Journal of *Personality Assessment*, *68*(3), 532–560.

Gullikson H. (1962). *Theory of mental tests* (4th. ed.) USA: John Wiler Sons, Inc.

Güloğlu, B., & Aydın, G. (2001). Coopersmith özsaygı envanteri'nin faktör yapısı. *Eğitim ve Bilim*, *26*, 66-71.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, *19*, 49-78.

Joreskog, K. G., & Sorbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Kline, T. J. B. (2005). *Psychological testing: A practical approach to design and evaluation.* CA: Sage.

Leedy, P. D. (1997). *Practical research: Planning and design*. Upper Saddle River, NJ: Prentice-Hall.

Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph Supplement, No:7*.

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

McMillan, J. H. (1992). *Educational research: fundamentals for the consumers*. New York: Harper Collins.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd Ed.). New York: McGraw-Hill.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*(4), 329-353.

Rudner, L. M. Schafer, W. D. (2001). Reliability, *ERIC Digest*. ED 458213.

Salvia, J., & Ysseldyke, J. E. (1995). *Assessment*. USA: Haughton Mifflin Company.

Salvucci, S., Walter, E., Conley, V., Fink, S., & Saba, M. (1997). *Measurement error studies at the National Center for Education Statistics (NCES).* Washington D. C.: U. S. Department of Education.

Schmitt, N. (1996). Uses and Abuses of Coefficient Alpha. *Psychological Measurement*. *8*(4), 350-353

Thorndike, R. L., & Hagen E. (1961). *Measurement and evaluation in psychology and education*. (2nd. ed) Upper Saddle River, NJ: Prentice-Hall.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education*. Upper Saddle River, NJ: Prentice-Hall.

Vockell, E. L., & Asher J. W. (1995). *Educational research*. Englewood Cliffs. NJ: Prentice-Hall.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice* 22-29.

Zimmerman, D. W., Zumbo B. D., & Lalonde C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*. *53*, 33-49.