Basic Concepts in Classical Test Theory:

Relating Variance Partitioning in Substantive Analyses

to the Same Process in Measurement Analyses

Thomas E. Dawson

Texas A&M University

ABSTRACT

The basic processes in univariate statistics involve partitioning the sum of squares into two components: explained and within. The present paper explains that the same partitioning occurs in measurement analyses, i.e., splitting the sum of squares into reliable and unreliable components. In addition, it will be shown how the three types of error inherent in all statistical analyses (i.e., sampling error, model specification error, and measurement error) impact any analysis the researcher attempts. Also emphasized will be that tests are not reliable, rather scores have varying degrees of reliability.

Basic Concepts in Classical Test Theory: Relating Variance

Partitioning in Substantive Analyses to the Same Process in Measurement Analyses.

Variance in the dependent variable is the crux of all statistical analyses, hence, it is the focus of all statistical analyses. As an analogy, the variance observed in the dependent variable can be viewed as making up one entire pie, e.g., as the variance increases, so would the size of the pie--usually the size of the pie is equal to the sum of squares (SOS) total. Further, there are three different ways to interpret variance in the dependent variable, or "knives" which can be used to cut the pie. One knife will discriminate between which ingredients were used, another will describe how much of which ingredient, and yet another that will give an indication of how accurate the baking process is (measuring cups or utensils used to make the pie actually are). In statistical analyses, these knives are the "who," "how," and "reliability" partitions of the observed variance, i.e., who accounts for the variance, how the variance is accounted for by other variables, and the reliability of the measurement. Which knife we use depends on what question we want to answer. Further, there is error directly associated with each knife, i.e., sampling error, model specification error, and measurement error respectively.

In a substantive context, variance is partitioned into "who" accounts for it and "how" it is accounted for (via which predictor variable). All substantive analyses are in effect, regression, i.e., they all produce a y-hat and an error score. The present paper will show that the same model that is utilized in substantive analyses for partitioning the observed variance into explained and unexplained components, is also used to partition the variance in a measurement context into reliable and unreliable components. Indeed, the substantive and measurement contexts even have similarities at the sore level.

In a measurement context, partitioning focuses on reliability. Measurement analyses asks about the *stability*, *equivalency*, or *consistency* of the dependent variable score(s). That is, with what amount of certainty can the researcher believe that the obtained value or

score will replicate in the future, or that the obtained value or score is "true". Reliability generally refers to the degree to which test scores are free from measurement error (Sax, 1989). Reliability always refers to the *scores* obtained on an instrument for a *particular group* of examinees on a *particular occasion*--and not the instrument itself (Eason, 1991; Rowley, 1976; Thompson, 1994). Reliability also impacts effect sizes in substantive research, as will be explained below.

Heuristic Examples of Substantive Analyses

Who accounts for the variance?

For answering the "who" question in substantive analyses, the variance observed is partitioned by who (which participant) accounts for, and the amount of variance they account for, in the total observed (or dependent variable) score. If Marjie, Tommy, and Diane all completed a hypothetical newly formed short version of the Behavior Assessment for Children (BASC) (Reynolds & Kamphaus, 1994) to identify future school performance, a "who" analysis could look something like Table 1 and Figure 1.

Table 1
Hypothetical data for a "who" analysis

Student	***Y***	***Y-Mean***	y^2
Marjie	7	1	1
Tommy	8	2	4
Diane	3	-3	9
SUMS	18	0	14
MEAN	6		

Figure 1. Venn diagram displaying partitioned variance reflecting a "who" analysis.

$$y_i = Y_i + YMean$$

M	T	T	T	D	D	D
T	D	D	D	D	D	D

Note. M = Marjie, T = Tommy,

D = Diane.

Thus, the sum of squares (SOS) total is partitioned into its component parts according to who accounts for how much of the total. As with all analyses, error is a factor. In the formula, $y_i = Y_i + YMean$, there is no error component. This is because the type of error that impacts the who analysis is sampling error. If the sample is not representative of the population, then Figure 1 will not reflect reality. For example, if the sample is drawn from the tails of the population distribution, then the SOS_{total} will be overestimated, leading to erroneous findings. However, the sampling error would not effect the SOS at the individual score level though.

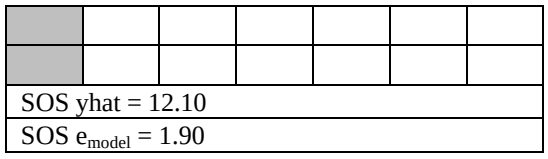
How the Variance is Accounted For

The other substantive "knife" cuts the variance into *how* it is made up. In a "how" analysis, a predictor variable is added to see how much variance that predictor variable accounts for, or does not account for, in the total observed variance. Assume that the small sample utilized in Table 1 reflects the total population. Assume also that this researcher wants to predict that an age difference as little as a few months will make a difference on the BASC=s identification of future school performance. If all three students above were 7 years old at the time of administration, but Marjie was 1 month past 7 years old, Tommy was 2 months past 7 years old, and Diane was 6 months past 7 years old, then number of months could be a predictor variable, and the "how" substantive analysis could look like Table 2 and Figure 2.

Table 2
"How" analysis with predictor "X" being months of age

Student	**Y**	**X**	**x**	**x^2**	**Y**	**X**	**Yhat**	**e_{mod}**	**yhat**	**$yhat^2$**	**e_{mod}**	**e^2_{mod}**
Majie	7	1	-2	4	7	1	7.86	-.86	1.86	3.46	-.86	.74
Tommy	8	2	-1	1	8	2	6.93	1.07	.93	.86	1.07	1.14
Diane	3	6	-3	9	3	6	3.21	-.21	-2.79	7.78	-.21	.04
SUMS	18	9	0	14	18	9	18	0		12.10		1.90
MEAN	6	3			6	3						
SD_x	2.65											
COV_{xy}	-6.5											
r_{xy}	-.93											
r^2	.86											

Figure 2. Venn diagram displaying partitioned variance reflecting a "how" analysis.



Error variance found in this model ($y_i = yhat_i + e_{model}$) would be due to choosing the wrong predictor variable(s); thus, model specification error--the predictor variable did not account for all the variance in the dependent variable, meaning something else does. In Table 2b, a regression model is utilized to partition the variance into explained and unexplained components to determine *how* months of age explains, or does not explain, the dependent variable variance.

Heuristic Examples of Measurement Analyses

As stated earlier, from a measurement perspective (the last of the three ways of partitioning variance, depending on which questions the study wishes to answer), reliability is the question addressed. Will these results replicate? This is important to know for many reasons. For example, if someone=s IQ fluctuated by 50 points each time they were tested, then those measurements on that IQ test give no dependable information and are unreliable. Using unreliable data such as that would be as inane as attempting to predict a person=s IQ from their shoe size: it=s not possible, i.e., not stable, not equivalent, not consistent, NOT RELIABLE.

Taking the substantive equation $y_i = yhat_i + e_i$, we substitute T (true score) for yhat, and e_{meas} (measurement error) for e. Thus, the equation becomes: $y_i = T_i + e_{meas}$. This equation is the true-score theory=s premise: that a person=s observed score is equal to that person=s true score + error. True score in this sense speaks to the "pure" indigenous trait the person holds--the true knowledge or ability (Sax, 1989). This value is a hypothetical value and is expected to yield consistent knowledge of individual differences. The true score is based on the premise that the person=s inherent ability is stable, and over repeated testing the mean of those scores would be the true value. Since infinite numbers of repeated testing are not feasible, the obtained value + measurement error is substituted for the true score. A measurement which contained no error, would in fact measure only true ability, so in a sense by measuring reliability, we are approximating true scores (Pedhazur & Schmelkin, 1991). Since true scores are not known, then the amount of measurement error cannot be known either. Still, it is possible to estimate the effect of measurement error in general (Sax, 1989). To the extent that error is eliminated, reliability will be high. When measurement error variance is high, there must be a corresponding decrease in reliability. Similarly, when error variance is reduced, true and obtained scores will more closely approximate each other, thereby increasing reliability (Sax, 1989). In classical test theory there are three ways to measure reliability: measurement error resulting through an error in test occasions (stability), or an error in test forms (equivalence), or an error in items (internal consistency) (Crocker & Algina, 1986). We will explore these methods in the order given.

Reliability as Stability

The test-retest method has been utilized to measure the stability of scores over a period of time. If individuals respond consistently from one test to another, the correlation between the test scores will be high. Some researchers point to the squared correlation coefficient as a coefficient of stability. The time difference between tests impacts the stability coefficients. If time intervals between tests are short, the stability coefficients are likely to be high. If the time period is longer, the stability coefficient is likely to be lower (Pedhazur & Schmelkin, 1991). This is one reason to speak about the reliability of measurements and not the reliability of tests--the test is the same one given at a different point in time, possibly yielding much different reliability coefficients as test intervals are varied! A shorter interval usually produces higher stability coefficients than a longer time interval. To demonstrate how a regression model can be employed in both a substantive and measurement context, the same heuristic data set will be used in this example as in the previous example. If the students reported on earlier were administered the same test at a different point in time, a test-retest measure of reliability might look something like Table 3 and Figure 3. Let Y_1 = the first administration and Y_2 = the second administration of the BASC.

Table 3a-b
Test - rest method hypothetical BSAC scores

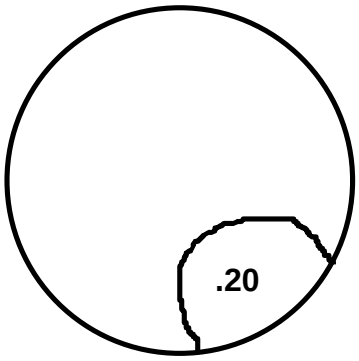
	a. Test-retest method on hypothetical BSAC scores						**b. Regression model for test-retest data substituting T for Y and e_{meas} for e**				
Student	**Y_1**	**y_1**	**y_1^2**	**Y_2**	**y_2**	**Y_2^2**	**T**	**t**	**t^2**	**e_{meas}**	**e^2_{meas}**
Majie	7	1	1	5	-.33	.11	5.52	-.49	.24	1.48	2.19
Tommy	8	2	4	7	1.67	2.79	8.5	2.49	6.2	-.5	.25
Diane	3	-3	9	4	-1.33	1.77	4.03	1.98	3.9	-1.0	1.06
SUMS	18			5.33		4.67	18.0	0	10.4		3.5
MEAN	6			1.53			6.0				
SD	2.65										
COV_{y1y2}	3.5										
R_{y1y2}	.86										
r^2	.74										

As can be seen in Table 3b, the derivation of the true score and measurement error in a measurement context (i.e., partitioning the variance into reliable versus unreliable components), is the same one employed in a substantive context utilizing yhats and error scores to partition variance into explained versus unexplained components. The mechanics of the partitioning is the same, only the purposes of the partitioning differ. Following the premise of true-score theory, any error inherent in this design would be due to the measurement process, and not changes in the individual themselves, because as stated earlier, the "pure" indigenous trait the person holds, is consistently present.

Reliability as Equivalence

These same methods can be used in the second measure of reliability - the equivalence, or parallel forms of a test. In this measure, two or more forms of a test are constructed and administered to the same person at approximately the same time. To eliminate practice or transfer effects, half of the participants take one form followed by the other, and the sequence is reversed for the other half of participants. The correlation between the scores on the forms is a measure of their equivalence, and is designated as a *reliability index.* When squared, this reliability index is the reliability coefficient of the measurement (Gronlund & Linn, 1990). All reliability coefficients are squared concepts. As stated earlier, this example elucidates the fact that a reliability coefficient places a ceiling on effect sizes (Rienhardt, in press). From a reliability standpoint, the e_{meas} is the part of the pie that cannot be eaten (explained). As can be seen in Figure 4, the e_{meas} = 20%. If we add a predictor variable that explained all the remaining 80% of the pie, an effect size could not exceed that 80%. In the worst case scenario, a dependent variable is measured such that scores are perfectly unreliable, hence, the effect size will be "0", and the results will not be statistically significant at any sample size, even an incredibly large one (Reinhardt, in press).

Figure 4 Example of a reliability coefficient between parallel forms of an instrument, placing a ceiling on effect size



r_{ylyw}= .80 r^2_{yx} # .80

Parallel forms are never perfectly correlated and the further from a correlation of 1 that they differ, the greater the amount of unreliability. However, because equivalence is determined by correlating scores on tests designed to be parallel, the unreliability must come from differences in item sampling and not, as in measures of stability, changes within the individuals themselves. The exact same methods for deriving T scores and measurement error utilized in the test-retest example are employed in the equivalence check on reliability.

Reliability as Consistency

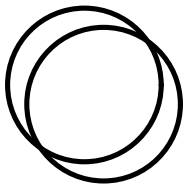
The final technique of estimating reliability, is the internal consistency method. Because of practicality, most teachers, psychologists and researchers will usually not administer the same test twice, or develop an alternate form of an instrument. In most cases researchers would like to estimate reliability from one administration of an instrument. This desire has led to measures of internal consistency, historically, the split-half method. In this method, one test or instrument is split in half and the 2 halves are treated as alternate forms of the other, thereby obviating the need to construct more than one instrument (Gronlund & Linn, 1990).

Many different ways of splitting a test are available, but the most important consideration is that the two halves be parallel. If we use the same example employed previously, and split the items by way of even (E) or odd numbers (O), the resulting scores would look like Table 4 and Figure 5.

Table 4
Split-half method of hypothetical BASC scores

Student	O	o	O^2	E	e	e^2
Majie	3	.7	.49	4	.33	.11
Tommy	3	.7	.49	5	1.3	1.8
Diane	1	-1.3	1.7	2	-1.7	2.8
SUMS	7		2.7	11		4.7
MEAN	2.3					3.7
SD	1.15					1.53
COV_{oe}	1.67					
r_{oe}	.94					

Figure 5 Venn diagram partitioning variance into reliable versus unreliable components by way of consistency



When a correlation coefficient is computed on a split-half reliability measure, the resulting correlation is a measure of the "agreeability" between one half of the instrument and the other. When squared, such correlations provide a measure of reliability for half an instrument, but not for the instrument as a whole. To estimate the reliability of the whole instrument from knowledge of the correlation between the halves, the Spearman-Brown formula must be employed (Thorndike, Cunningham, Thorndike, & Hagen, 1991), and is as follows:

$$\frac{2 \text{ X the correlations between the halves}}{1 + \text{the correlations between the halves}}$$

From the example in Table 4: $2(.94)/1+.94 = 1.88/1.94 = .97$, $r^2 = .94$

Thus the actual correlation between the two halves of the test is .97, and when squared, this is the reliability of the measurement in terms of consistency (.94).

Coefficient Alpha α (also named, "Cronbach=s alpha," (Cronbach, 1951)) is another measure of internal consistency that is a squared concept, even though there is no squared sign in the symbol designating α. Theoretically, coefficient alpha is an estimate of the squared correlation expected between two tests drawn at random from a pool of items similar to the items in the test under construction (Pedhazur & Scmelkin, 1991). Coefficient alpha can be used as an index of internal consistency conceptually exhibiting how item responses correlate with total test score, and employs the same concept as the split-half measure of internal consistency, except that coefficient alpha pairs each item on the instrument with all combinations of all other items. Coefficient alpha is superior to the use of split-half

measures, because as stated earlier, there are many different ways in which to split an instrument. Estimates associated with different splits for the same data may yield contradictory results (Sax, 1989). For example, a 4-item test has 3 splits, a 6-item test has 10 splits; and for a test with 10 items, there are 126 different ways to split the test (Reinhardt, in press)! So, as the number of items increase, so do the number of possible splits. The formula for coefficient alpha is as follows:

$$\alpha = k/k\text{-}1(1\text{-}3\sigma^2_i/\sigma^2_x)$$

k is the number of items

$3\sigma^2_I$ = the sum of the variances of the items

σ^2_x = the variance of the total score, or composite score

Where k is the number of items; $3\sigma^2_I$ = the sum of the variances of the items; and σ^2_x = the variance of the total score, or composite score (Pedhazur & Schmelkin, 1991). Using our data set of the children=s scores on the BASC, coefficient alpha would look something like Table 5.

Table 5

Student	**Items**								
	1	2	3	4	5	6	7	8	Score
Marjie	1	1	1	1	0	1	1	1	7
Tommy	1	1	1	1	1	1	1	1	8
Diane	0	0	1	0	0	1	0	1	3
P	.66	.66	100	.66	.33	100	.66	100	3 18
Q	.33	.33	0	.33	.66	0	.33	0	
σ^2 [Pxq]	.21	.21	0	.21	.21	0	.21	0	1.05

The P values in Table 5 are derived by finding the ratio of scores of 1 on an item, to a score of 0. In item one this is 2/3 = .66. The P value is also an index of item homogeneity, i.e., how alike the P values are gives an indication as to how varied the scores are. The item variance is found and summed (in this example, item variance = 1.05). The composite score is computed by finding the variance of the row totals: 7+8+3=18. In this case the composite variance = 7.02. The numbers are then plugged into the above formula giving a coefficient alpha: 8/7 (1-1.05/7.02) = .97. Therefore, we have an estimate of the reliability of the items and how they relate to each other, and to total test variance.

Upon examination of the formula for coefficient alpha, we find that the total item variance is the numerator, and total test or composite score, is the denominator. The alpha coefficient is 1 minus this ratio. Therefore, it behooves the test constructor to maximize total test variance, while item variance is minimized. As can be seen from the following hypothetical data sets, the alpha coefficient can even be negative (Reinhardt, in press). This usually happens when item variance is larger than total test variance (Arnold, 1996). Table 6 and Figure 6 are employed to help make these concepts concrete.

Table 6a, b, c
Probability Target Matrix Depicting Effects of Item and Composite Variance on Coefficient Alpha

a. Min Test Variance, Max Item Variance, Homogeneous p
Known Results for the population

	Item							
n_I	1	2	3	4	5	6	7	Total
1	1	0	1	0	1	0	1	4
2	0	1	0	1	0	1	0	3
3	1	0	1	0	1	0	1	4
4	0	1	0	1	0	1	0	3
5	1	0	1	0	1	0	1	4
6	0	1	0	1	0	1	0	3
7	1	0	1	0	1	0	1	4
8	0	1	0	1	0	1	0	3
9	1	0	1	0	1	0	1	4
10	0	1	0	1	0	1	0	3
p	.5	.5	.5	.5	.5	.5	.5	
var_k	.25	.25	.25	.25	.25	.25	.25	.25

α = 1.166667 x (1 - (1.75 / .25)) = -7

Note. Table adapted from "Factors Affecting Coefficient Alpha: A Mini Monte Carlo Study," by B. Reinhardt, (in press), in B. Thompson (Ed.), Advances in social science methodology (Vol. 4). Greenwich, CT: JAI Press. Adapted with permission.

b. Mod Test Var, Mod Item Var, Heterogeneous p
Known Results for the Population

	Item							
n_i	1	2	3	4	5	6	7	Total
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	1	1	1	1	4
7	0	0	0	0	1	1	1	3
8	0	0	0	1	1	1	1	4
9	0	0	0	0	1	1	1	3
10	0	0	0	1	1	1	1	4
p	0	0	0	.3	.5	.5	.5	
var_k	0	0	0	.21	.25	.25	.25	3.36

α = 1.16667 x (1 - (.9600001 / 3.36)) = .8333333

Note. Table adapted from "Factors Affecting Coefficient Alpha: A Mini Monte Carlo Study," by B. Reinhardt, (in press), in B. Thompson (Ed.), Advances in social science methodology (Vol. 4). Greenwich, CT: JAI Press. Adapted with permission.

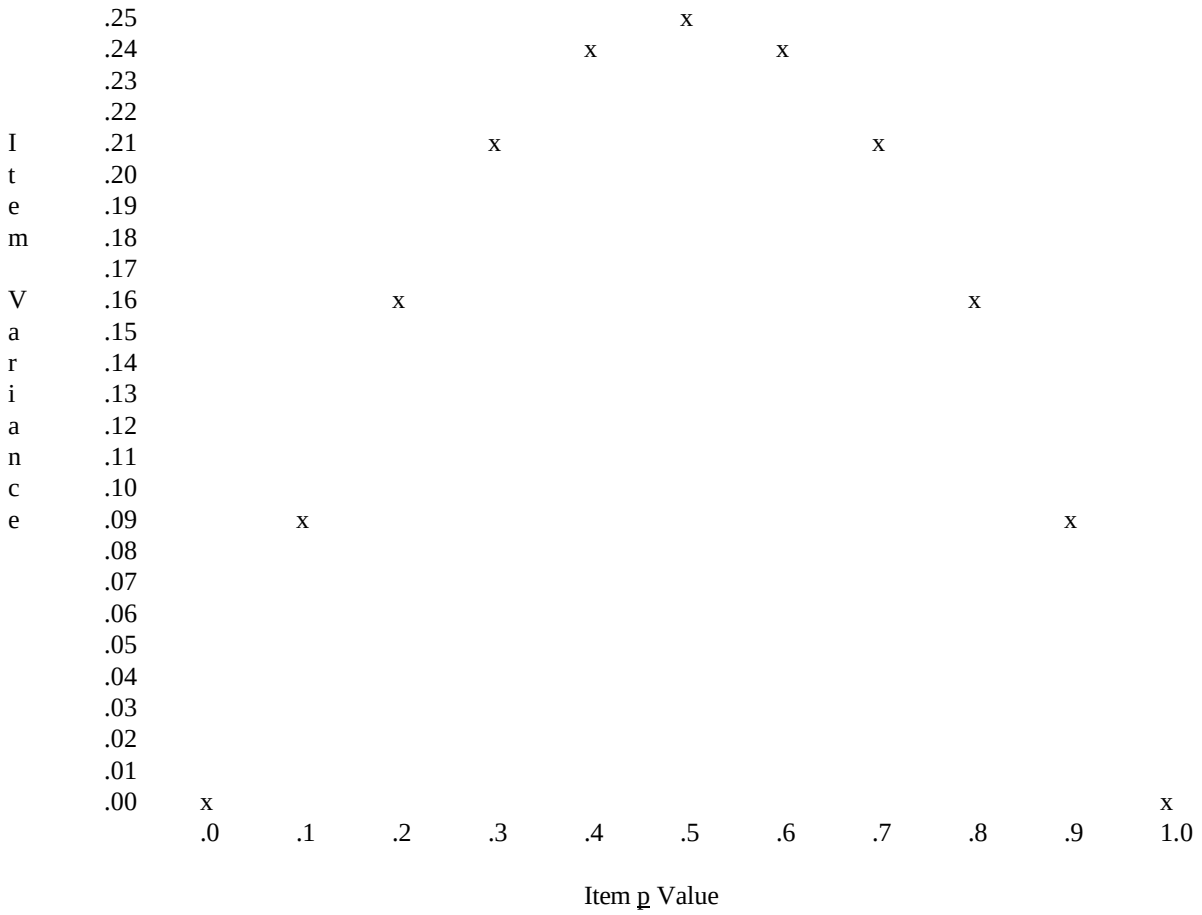
c. Max Test Var, Max Item Var, Homogeneous p

	Item							
n_i	1	2	3	4	5	6	7	Total
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	1	7
7	1	1	1	1	1	1	1	7
8	1	1	1	1	1	1	1	7
9	1	1	1	1	1	1	1	7
10	1	1	1	1	1	1	1	7
p	.5	.5	.5	.5	.5	.5	.5	
var_k	.25	.25	.25	.25	.25	.25	.25	12.25

$\alpha = 1.16667 \times (1 - (1.75 / 12.25)) = 1$

Note. Table adapted from "Factors Affecting Coefficient Alpha: A Mini Monte Carlo Study," by B. Reinhardt, (in press), in B. Thompson (Ed.), Advances in social science methodology (Vol. 4). Greenwich, CT: JAI Press. Adapted with

Figure 6.
Variance of Scores on Dichotomously-Scored Items With 10 Examinees



Note. With 10 examinees completing a given item, there are 11 possible p values, each with an associated item variance.

Note. Table adapted from "Factors Affecting Coefficient Alpha: A Mini Monte Carlo Study," by B. Reinhardt, (in press), in B. Thompson (Ed.), Advances in social science methodology (Vol. 4). Greenwich, CT: JAI Press. Adapted with permission.

From these results in this data set, it can be inferred that maximum total test variance is important to maximize coefficient alpha, and that total test variance has more impact on alpha than item variance.

Another measure of internal consistency for dichotomously-scored items is the KR-20 formula. The KR20 formula and the alpha coefficient formula are the same, except for the derivation of item variance, as can be seen below:

$$KR\text{-}20 = k/k\text{-}1(1\text{-}3\ pq/\sigma^2_{Total})$$

$$\alpha = \quad k/k\text{-}1\ (1\text{-}\ 3\sigma^2_I/\sigma^2_{Total})$$

But the formulas are algebraically equivalent, even though the ways for computing item variance seem different.

Summary

In sum, the present paper has explained the three types of statistical analyses and the corresponding error which accompanies each. Two of the analyses are substantive ("who" and "how"), and one involves a measurement perspective (reliability). Further, the same method to analyze the data (regression) has been utilized in both substantive and measurement analyses to partition explained versus unexplained variance, and reliable versus unreliable variance in the dependent variable, according to which question the researcher wishes to answer. The "classical" methods of estimating reliability have been explained with an emphasis on coefficient alpha.

References

Arnold, M. E. (1996, January). Influences on and limitations of classical test theory reliability estimates. Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 197-334.

Eason, S. (1991). Why generalizabilty theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 83-89). Greenwich, CT: JAI Press.

Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed). New York: Macmillan.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Reinhardt, B. (in press). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4). Greenwich, CT: JAI Press.

Reynolds, C. R., & Kamphaus, R. W. (1992). Behavior assessment for children (BASC): Manual. Circle Pines, MN: American Guidance Service.

Rowley, G. L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Sax, G. (1989). Principles of educational and psychological measurements and evaluation (3rd ed.). Belmont, CA: Wadsworth.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991).

Measurement and evaluation in psychology and education (5th ed.). New York:

Macmillan.