June 1992

Recent Developments in Reliability Analysis

Klaus Krippendorff
University of Pennsylvania, kkrippendorff@asc.upenn.edu

Recent Developments in Reliability Analysis

Abstract

For many researchers, the literature of reliability coefficients seems bewildering although the methodological problem in which they are embedded is reasonably clear: Since we can never know what it is that we claim to see independent of our seeing it, or, translated into the language of science, since we can not test hypotheses about reality without first generating the observations or data to talk about, the accuracy by which primary data "represent" an unobserved nature remains unascertainable in principle (Krippendorff, 1991). Yet, to assure that the data that go into scientific inquiries are not accidental, it is important to demonstrate that the data-generating procedures are reproducible under varying circumstances and by several observers. All reliability measures are intended to express the degree to which several observers, several measuring instruments, or several interrogations of the same units of analysis yield the same descriptive accounts, category assignments, quantitative measures or data for short.

Comments

Postprint version.

Paper presented at the meeting of the
International Communication Association
in Miami Florida, May 21-25, 1992, revised June 3, 1992.

Recent Developments in Reliability Analysis

Klaus Krippendorff
The Annenberg School for Communication
University of Pennsylvania, Philadelphia
kkrippendorff@asc.upenn.edu

Introduction

For many researchers, the literature of reliability coefficients seems bewildering although the methodological problem in which they are embedded is reasonably clear:

Since we can never know what it is that we claim to see independent of our seeing it, or, translated into the language of science, since we can not test hypotheses about reality without first generating the observations or data to talk about, the accuracy by which primary data "represent" an unobserved nature remains unascertainable in principle (Krippendorff, 1991). Yet, to assure that the data that go into scientific inquiries are not accidental, it is important to demonstrate that the data-generating procedures are reproducible under varying circumstances and by several observers. All reliability measures are intended to express the degree to which several observers, several measuring instruments, or several interrogations of the same units of analysis yield the same descriptive accounts, category assignments, quantitative measures or data for short.

But,

- Reliability coefficients are often specialized to different metrics (levels of measurement). There are nominal scale coefficients, Scott's (1955) pi, for example, and interval scale coefficients, Kuder and Richardson's (1937) Formula #20, for example, that differ in the metric to which they claim applicability but moreover stem from incompatible analytical traditions.
- Reliability coefficients have built-in assumptions that do not easily reveal themselves to their users, and the often yield vastly different results. For nominal scales alone, there is % (percent) agreement, Bennett, Alpert and Goldstein's (1954) S, Goodman and Kruskal's (1954) family of lambda coefficients, Scott's (1955) pi, Cohen's (1960) kappa and Fleiss (1971) kappa, which are different, Perreault and Leigh's (1989) I_r and many more, not to forget my own alpha (Krippendorff, 1980). Researchers encounter difficulties in choosing among them without detailed examination of their assumptions. Often this is not obvious. For example, Cohen's (1960) frequently used kappa turns out to be a hybrid that behaves like an agreement coefficient near its largest value of plus one, and like an association or correlation coefficient near its zero-value (Krippendorff, 1978). Where its values would matter most, kappa is not consistently interpretable.

- Reliability coefficients often vary in their ranges of values. Some range from zero to one, the notorious % agreement, for example, but also Kuder and Richardson's (1937; Cronbach, 1951) proportion of systematic to total variance. Some range from minus to plus one (Scott, 1955; Cohen, 1960). Variations in their ranges make it virtually impossible to assign uniform meanings to the numbers they produce.

While one can always find reasons for preferring one coefficient over another, when it is desirable to set data reliability standards for a class of scientific inquiries, or when one needs to compare and select among many different kinds of data whose reliabilities are crucial to a particular research undertaking, one needs a single coefficient that is adaptable to all or most situations of interest.

In pursuit of this aim, I have over the years developed the agreement coefficient alpha (Krippendorff, 1967, 1970, 1978, 1980), which takes this general form

$$\alpha = 1 - \frac{D_o}{D_e} \quad . \qquad (1)$$

Alpha is zero when the observed disagreement D_o equals the disagreement D_e, which would be expected under conditions of chance, one when observed disagreement D_o is absent, indicating the absence of reliability, and becomes negative when the observed exceeds the expected disagreement, which can arise only under conditions of consensual disagreement. While the plus one and zero values of this coefficient make alpha easily interpretable, the general form of (1) is common to several other coefficients as well and not yet specific about the assumptions that go into the definitions of the observed and the expected disagreements.

As acknowledged above, we cannot state anything about reality until after data have been created. Without a standard to compare the data to, this leaves us with reliability or reproducibility as the only measurable criterion. Reproducibility becomes evident in substantial agreement among the results of applying a battery of the same observational, accounting or measuring procedures to the same set of units of analysis. Under these conditions – and without privileging any one observer over another – the only defensible statement one can make about the "true nature" of the data depends on what all observers concur they see, or on what all measuring devices agree. From its beginning, this epistemological fact was built into alpha. This is manifest in both, in how the observational accounts of the individual units are evaluated, and in how the statistical distribution of the data is characterized, to which all observers or measuring instruments jointly contribute. The former leads to the observed disagreement D_o, and the latter to the expected disagreement D_e. Measures of agreement may estimate the nature of what is observed but must acknowledge its unknowability. In this respect agreement measures that are suitable for reliability interpretations differ from measures of correlation or association, which make very different assumptions (Krippendorff, 1978).

Let me jump a bit ahead of the developments of alpha and start with the canonical form of reliability data, an r-by-m matrix of up to rm single values, each generically denoted by b or c:

	Units:	1	2	…	u	…	r
	1				.		
	2				.		
	.				b_u		
Observers:	.				c_u		
	.				.		
	m				.		
Number of values in u:					m_u		

Reliability data must provide the basis for comparing the values that observers assigned to units u. As there may be missing data, the actual number of values in the reliability data matrix is less important than that they are comparable within units. Let n be the number of values that contribute to pair comparisons within units.

$$n = \sum_u m_u \mid m_u > 1 \tag{2}$$

n excludes all units with "lone values," $m_u \leq 1$, which are the values that cannot be compared within units. $n \leq rm$.

An important early decision was to correct alpha for small sample sizes (small numbers of either units of analysis or observers/instruments or both). Many of the coefficients used in content analysis, Scott's (1955) pi, for example, did not provide for this correction and systematically underestimated reliability when samples were small.

In the above terms, the expected disagreement D_e, mentioned in (1), can be expressed as

$$D_e = \overline{D} = \frac{1}{n(n-1)} \sum_b \sum_c d_{bc}^2 \tag{3}$$

where d_{bc} is a difference between any two values, observations, or data points, b and c. The nature of this difference will be addressed below. By analogy to (3), the disagreement within any one unit u is

$$\overline{D}_u = \frac{1}{m_u (m_u - 1)} \sum_{b_u} \sum_{c_u} d_{b_u c_u}^2 \quad . \tag{4}$$

The observed disagreement D_o, also mentioned in (1), is defined as the average disagreement observed within units u.

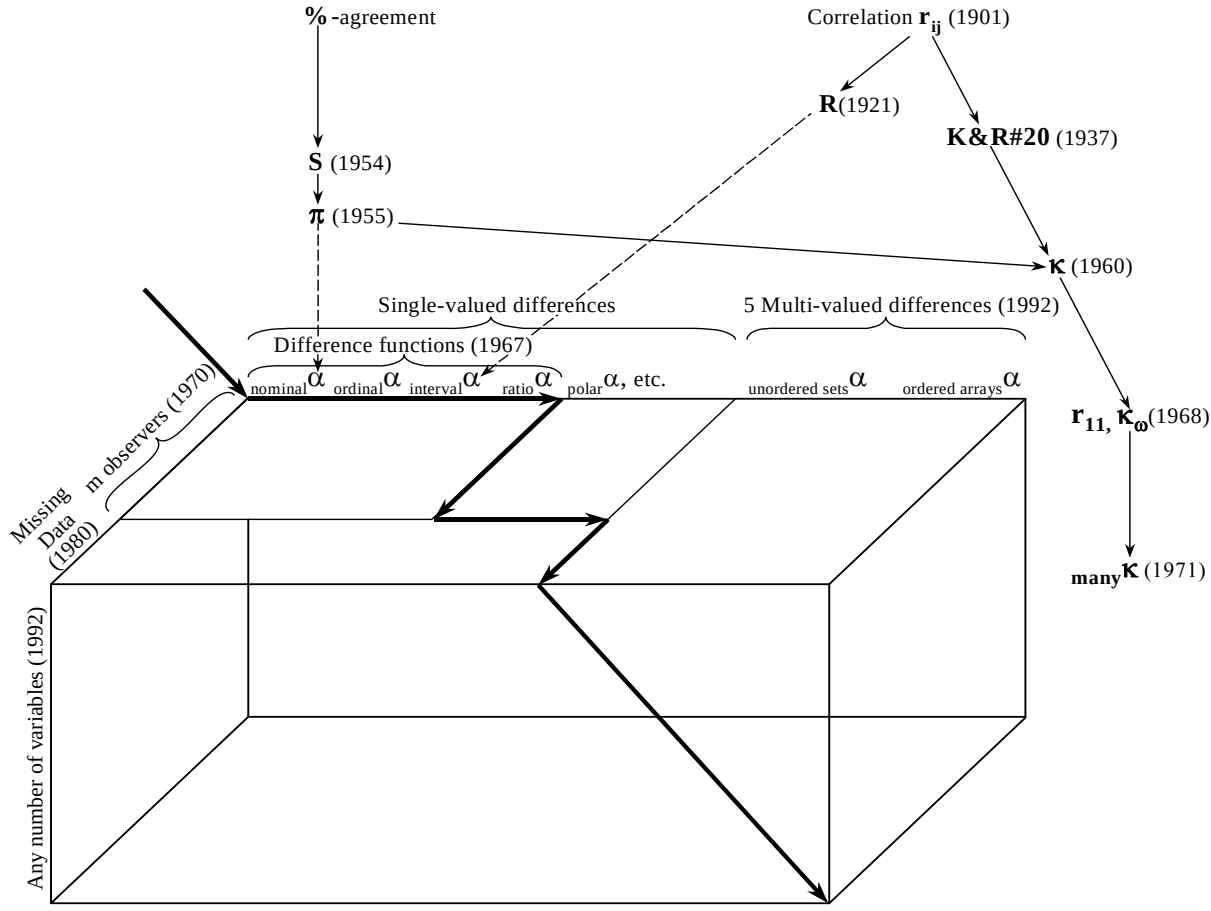
$$D_o = 1 - \sum_{u=1}^{r} \frac{m_u}{n} \overline{D}_u \tag{5}$$

In these terms, (1) becomes:

$$\alpha = 1 - (n-1) \frac{\sum_{u=1}^{r} \frac{1}{m_u - 1} \sum_{b_u} \sum_{c_u} d_{b_u c_u}^2}{\sum_b \sum_c d_{bc}^2} \tag{6}$$

(3) and (4) reveal measures of disagreement to be average differences. The differences between all possible pairs of values within the whole reliability data matrix and within each unit respectively are enumerated, and divided by the number of possible differences. (6) reveals alpha as one minus an error, the proportion of disagreement within units and the total disagreement.

Before going into various forms of alpha, let me introduce with Figure 1 a kind of travel plan that shows in bold arrows how my thinking developed and how the space within which alpha is applicable came to be expanded. This arrows indicate acknowledged sources and broken arrows reconstructed relationships. The following describes some of the steps – in bold arrows – that I took.



Steps Taken to Create a Larger Space for the Agreement Coefficient Alpha
Figure 1

The generalization to any metric was accomplished in 1967 when I wrote a computer program for a large content analysis project (Brouwer, et al., 1969) using the first four of the following difference functions. These four are shown also in Table 1, each associated with one metric or scale of measurement.

$$\text{nominal } d^2_{bc} = \begin{cases} 1 \text{ iff } b \neq c \\ 0 \text{ iff } b = c \end{cases} \tag{7}$$

$$\text{ordinal } d^2_{bc} = \left(\frac{n_b}{2} + \sum_{k>b}^{k<c} n_k + \frac{n_c}{2} \right)^2 \tag{8}$$

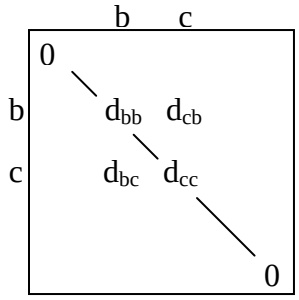
whereby n_b, n_k, and n_c are the frequencies of values b, k, and c in all reliability data for that variable.

$$\text{interval } d^2_{bc} = (b-c)^2 \tag{9}$$

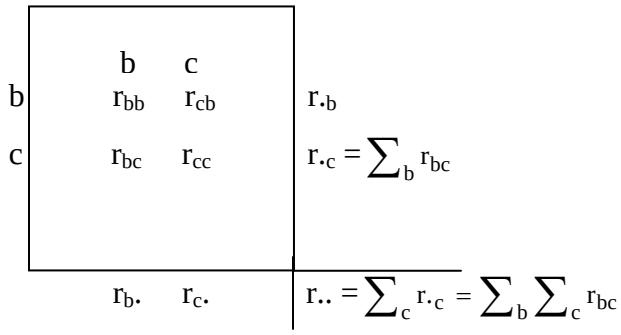
$$\text{ratio } d^2_{bc} = \left(\frac{b-c}{b+c} \right)^2 \tag{10}$$

$$\text{polar } d^2_{bc} = \frac{(b-c)^2}{(b+c-2k_{min})(2k_{max}-b-c)} \tag{11}$$

One can visualize the values of d_{bc} by entering them into a difference matrix (Krippendorff, 1980), which is square, its rows and columns are defined by the values occurring in the data, and its diagonal entries are zero.



In the literature, it is customary to conceptualize correlations and agreements largely between two variables or two observers, coders or measuring instruments and tabulate data in terms of <u>contingency</u> matrices, which contain one pair of values for each unit of analysis u, r in total number.



For two observers and in the above contingency matrix notations, (1) or (6) can be restated as:

$$\alpha = 1 - \frac{\sum_b \sum_c r_{bc} d^2_{bc}}{\sum_b \sum_c e_{bc} d^2_{bc}} \qquad (12)$$

wherein the expected frequencies e_{bc} are obtained by drawing pairs of values at random and without replacement from the n values available for comparisons

$$e_{bc} = \frac{(r_{b.} + r_{.b})(r_{c.} + r_{.c} - \vartheta_{bc})}{2r..(2r.. - 1)} \qquad (13)$$

wherein
$$\vartheta_{bc} = \begin{cases} 1 \text{ iff } b = c \\ 0 \text{ iff } b \neq c \end{cases} = 1 - {}_{\text{nominal}}d_{bc} \qquad (14)$$

All versions of alpha can be obtained by inserting appropriate difference functions into (6) or (12). They can be seen to serve as weights of the frequencies of pairs of observations. Although alpha did not derive from any agreement coefficients that I knew at that time, the two-observer nominal scale version of alpha, with the difference ${}_{\text{nominal}}d_{bc}$ taking the place of d_{bc} in (12), turned out to be Scott's (1955) pi, but corrected for small sample sizes, the relation between the alpha and pi being

$$\alpha = \frac{1 - \pi}{2r..} + \pi \, . \qquad (15)$$

When the sample size 2r.. (number of values generated by two observers) becomes large, the proportion $(1-\pi)/2r..$ converges to zero and alpha and pi then become indistinguishable. The interval alpha, with ${}_{\text{interval}}d_{bc}$ inserted into (12), turned out to be Pearson's (1901; Tildesley, 1921) intra-class correlation coefficient R. For dichotomous decisions, i.e., for 2-by-2 contingency tables, all pairs of values are either same or different and all difference functions (7) through (11) produce the same alpha, as they should. Thus, a comparison of the coefficients computed with unlike difference functions can reveal the information that a metric contributes to reliability and which metric most likely underlies the observers' handling of the data. The computer program used since 1967 produced alphas for the four standard metrics: nominal, ordinal, interval, and ratio.

Generalization to m observers or measuring instruments. This required a measure of agreement applicable to patterns of disagreements that are more complex than can be observed between two observers. I opted for a disagreement functions that accounted for all pairwise differences within a set of values contributed by up to m observers, see (3) and (4). This is by no means the only function possible. Entropy measures would do much the same, at least for nominal data. I tried them out (Krippendorff, 1971) but my preference was to conform to the conventions of the most common statistical techniques, particularly in the tradition of correlational statistics and analysis of variance, in which data are likely analyzed once reliability is established. Indeed, one can argue that reliability should ideally reflect the disagreements that matter in subsequent analyses and these can often be reduced to pairwise differences.

In m-dimensional contingency matrices, the computation of chance agreement proved difficult. I am suggesting that the customary representation of data in

contingency matrices, taken by both Scott's (1955) pi and Cohen's (1960) kappa, and leading to their common form

$$\text{Pi or Kappa} = \frac{\text{P(observed agreement)} - \text{P(expected agreement)}}{1 - \text{P(expected agreement)}} \qquad (16)$$

was a major conceptual obstacle for generalizations to more than two observers. The way I solved this problem was by abandoning contingency matrix representations of reliability data altogether in favor of what I called coincidence matrix representations (Krippendorff, 1980) and by no longer counting matches or agreements in favor of enumerating differences as in (3) and (4) or disagreements as in (1), (6), or (12). Coincidence matrices do not tabulate units of observation but all pairable values that observers associate with these units, and they do not distinguish among the individual observer's contributions to these data. In fact, they take observers as interchangeable, as is required when an agreement measure is to be interpreted as reproducibility.

	b	c	
b	n_{bb}	n_{cb}	$n_{\cdot b}$
c	n_{bc}	n_{cc}	$n_{\cdot c} = \sum_b n_{bc}$
	$n_{b\cdot}$	$n_{c\cdot}$	$n_{\cdot\cdot} = \sum_c n_{\cdot c} = \sum_b \sum_c n_{bc}$

When reliability data contain exactly rm values, which means that no data are missing from the reliability data matrix, the entries in a coincidence matrix are

$$n_{bc} = \frac{1}{m-1} \sum_{u=1}^{r} n_{b_u} (n_{c_u} - \vartheta_{bc}) \qquad (17)$$

where n_{b_u} is the number of values b in unit u and ϑ_{bc} is as in (14). In coincidence matrix terms, alpha for single-valued data becomes

$$\alpha = 1 - (n_{\cdot\cdot} - 1) \frac{\sum_b \sum_c n_{bc} d^2_{bc}}{\sum_b n_{b\cdot} \sum_c n_{\cdot c} d^2_{bc}} \qquad (18)$$

I should mention that the developments presented up to now were incorporated in the above mentioned computer program for alpha. Later, Cohen (1968) suggested a weighted kappa in which the frequencies in contingency tables were weighted in ways similar to how my difference functions weighted the frequencies in coincidence matrices. Cohen's weights had different purposes, however. Fleiss (1971) sought to generalize kappa to many "raters," but as this proved difficult, he generalized Scott's pi instead, maybe without recognizing it, in any case, without even citing Scott's approach.

Generalization to missing data. This turned out to be a natural extension of the generalization to m observers. Its key was the recognition that missing data prevented constructing coincidence matrices by (17). Since missing data meant $n.. \leq rm$, the generalization had to acknowledge that units could have been described by a variable number m_u of observers. If m_u is the number of pairable values in unit u, then each unit u contributes $m_u(m_u\text{-}1)$ differences. In order to preserve the definition (18) of alpha, coincidence matrices have now to be constructed by

$$n_{bc} = \sum_{u=1}^{r} \frac{n_{b_u}(n_{c_u} - \vartheta_{bc})}{m_u - 1} \qquad (19)$$

where ϑ_{bc} is as in (14). This was the whole adjustment needed to accommodate missing data.

Generalization to multiple values. Commonly, each unit of analysis is assigned exactly one value by each observer. When this is the case, disagreements (3) and (4) are simple averages of the difference d_{bc} between any pair of single values b and c. However, it may happen that observers are asked to represent each unit by an appropriate set of descriptors (keywords of articles, lists of relevant attributes, alternative descriptions, multiple categories). Under these conditions, the differences within any one set of values that describes one unit must not contribute to unreliability. What then matters are the differences between any two sets of values. The problem therefore was to define one or more difference functions between two sets of values where differences within either set are ignored while differences across these sets are aggregated into one numerical difference between the two sets.

Multi-valued descriptions of units to be compared are of two kinds, two unordered sets B and C of potentially unequal numbers g or s of values

$$B = \{b_1, b_2, \ldots, b_t, \ldots b_g\}$$

$$C = \{c_1, c_2, \ldots, c_t, \ldots c_s\}$$

and ordered arrays <b> and <c> of values with the same number z of values

$$<b> = <b_1, b_2, \ldots, b_t, \ldots b_z>$$

$$<c> = <c_1, c_2, \ldots, c_t, \ldots c_z>$$

Concerning the multi-valued differences between B and C, I have come to distinguish between two kinds. The core difference is the single-valued difference between the most representative elements of each set, acknowledging the metric of their values. The core differences are defined in (20) through (23) in Table 1.

- For nominal data, the core is the mode, the most frequent element in the set. Since there may be more than one value with the largest number of occurrences, the mode is the subset $\ddot{b}$ of values in B and $\ddot{c}$ in C with the same and highest number of occurrences in these sets.
- For ordinal data, the core is the median rank $\breve{b}$ and $\breve{c}$, the rank that occupies the midpoint when all values in either set are ranked. Should that midpoint fall

between two different values, the median is the arithmetic mean between the two ranks. The difference between the two core values is expressed relative to all available values in the variable, not just the two sets.

- For interval data, the core is the arithmetic mean $\bar{b}$ and $\bar{c}$ of the values in the sets.
- For ratio data, the core is the geometric mean $\hat{b}$ and $\hat{c}$ of the values in the sets.

Core differences ignore the variance with each set. The second multi-valued difference between two sets is to account for how much the two sets have in common. The obvious candidate for this difference was the set theoretical one, the number of values that the two sets do not share, numerically, $(\#B + \#C)/2 - \#(B \cap C)$. This form, however, would apply only to nominal data and ignore the shades of differences typical for data with ordinal, interval and ratio characteristics. The difference function that I sought defied operationalization for a long time. Finally, I succeeded in developing (25), which, as may not be obvious in Table 1, enumerates all single-valued differences between the values from the two sets, acknowledge their metric, and expresses the number of differences relative to the number of possible comparisons between them. (25) is not only intuitively correct, when applied to nominal data, it also reduced to (24), which resembles the set theoretical difference, and when applied to single-valued data it reduced to the single-valued difference d_{bc} chosen. As (24) and (25) express the lack of overlap between the two sets relative how large that overlap could be, I call it the average multi-valued difference function.

For ordered arrays <b> and <c> of values, I found three multi-valued differences particularly useful. Multi-valued arrays can be conceptualized as points in a multi-dimensional space. One attractive difference function is the hyper geometric difference between any two points in such a space. I used the Mahalanobis (1936) distance as a starting point for this difference function as it corrects for unequal magnitudes of variation in the dimensions of the space. In (26), this is accomplished by standardizing each of the z single-valued difference functions by the expected disagreement within the corresponding dimension (component of the array, or variable). In effect, (26) allots each variable or component of the arrays the same weight. But it also allows analysts to override this equality by using a weight ω_t that opens the possibility of considering potentially unequal contributions of variables or components to subsequent analyses and hence to reliability. Finally, Mahalanobis' conception of a multivariate distance made it possible to each dimension, variable or component to have its own metric. For this attractive feature, I called (26) the multi-metric difference function.

The second multi-valued difference function for ordered arrays is based on the Hamming distance between the two arrays. This distance simply enumerates the number of positions in the two arrays whose values differ. Whereas the multi-metric difference acknowledges that values in their respective positions may have different metrics, the Hamming difference treats them as nominal data. It is defined in (27) of Table 1.

Finally, I defined the absolute difference as any difference between two arrays, regardless of magnitude. (28) essentially ignores the complexities of the available arrays and treats them as nominal differences.

Metric:	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Single-valued differences $\text{metric}\, d^2_{bc}$	$= \begin{cases} 0 \text{ iff } b = c \\ 1 \text{ iff } b \neq c \end{cases}$ (7)	$= \left(\frac{n_b}{2} + \sum_{k>b}^{k<c} n_k + \frac{n_c}{2} \right)^2$ (8)	$= (b-c)^2$ (9)	$= \left(\frac{b-c}{b+c} \right)^2$ (10)
Multi-valued differences of unordered sets $\text{core}\, d^2_{BC}$	$\dddot{b}$ = the mode of B $\ddot{c}$ = the mode of C $= 1 - 2\frac{\#\,\dddot{b} \cap \ddot{c}}{\#\,\dddot{b} + \#\,\ddot{c}}$ (20)	$\breve{b}$ = the median of B $\breve{c}$ = the median of C $= \left(\frac{n_{\breve{b}}}{2} + \sum_{k>\breve{b}}^{k<\breve{c}} n_k + \frac{n_{\breve{c}}}{2} \right)^2$ (21)	$\bar{b} = \frac{1}{g}\sum_{t=1}^{g} b_t$ $\bar{c} = \frac{1}{s}\sum_{t=1}^{s} c_t$ $= (\bar{b} - \bar{c})^2$ (22)	$\hat{b} = \sqrt{\frac{1}{g}\sum_{t=1}^{g} b_t^2}$ $\hat{c} = \sqrt{\frac{1}{s}\sum_{t=1}^{s} c_t^2}$ $= \left(\frac{\hat{b} - \hat{c}}{\hat{b} + \hat{c}} \right)^2$ (23)
$\text{average}\, d^2_{BC}$	With #B = the number of values in B; #B∩$\overline{C}$ = the number of values in B and not in C; etc. $= 1 - 2\frac{\#B \cap C}{\#B + \#C}$ (24)	$= \frac{\frac{1}{\#B}\sum_{b\in B}\sum_{c\in C\cap\overline{B}} \text{metric}\, d^2_{bc} + \frac{1}{\#C}\sum_{b\in B\cap\overline{C}}\sum_{c\in C} \text{metric}\, d^2_{bc}}{\#B + \#C}$ (25)		
of ordered arrays $\text{multi-metric}\, d^2_{<b><c>}$	With ω_t = the weight associated with component t, normally set to 1 $\overline{D}_t$ = the expected disagreement (variance) of values in t, $\text{metric}\,\overline{D}_t = \frac{1}{n(n-1)}\sum_{b_t}\sum_{c_t} \text{metric}\, d^2_{b_t c_t}$ $= \sum_{t=1}^{z} \omega_t \frac{\text{metric}\, d^2_{b_t c_t}}{\text{metric}\,\overline{D}_t}$ By definition: $\frac{0}{0} = 0$ (26)			
$\text{Hamming}\, d^2_{<b><c>}$	$= \sum_{t=1}^{z} \text{nominal}\, d^2_{b_t c_t}$ (27)			
$\text{absolute}\, d^2_{<b><c>}$	$= \text{nominal}\ d^2_{<b><c>}$ (28)			

A Comparison of Difference Functions

Table 1

When applied to single-valued data, the first three multi-valued difference functions, (20) through (26) in Table 1, reduce to the single-valued differences of the chosen metric. Under the same conditions, the Hamming and absolute differences reduce to ${}_{\text{nominal}}d_{bc}$. The multi-metric difference, being standardized, yields single-valued differences that differ from the non-standardized ones, but standardization has no effect on the resulting alpha.

The observed disagreements D_o for multi-valued data do not differ from those for single-valued data, except for the difference functions entered. In coincidence matrix notations

$$D_o = \sum_b \sum_c \frac{n_{bc}}{n..} d^2_{bc}\ ; \qquad D_o = \sum_B \sum_C \frac{n_{BC}}{n..} d^2_{BC}\ ; \qquad D_o = \sum_{<b><c>} \sum \frac{n_{<b><c>}}{n..} d^2_{<b><c>} \tag{29}$$

This seamless continuity is not true, however, for obtaining the corresponding expected disagreements D_e.

The expected disagreements for multi-valued data must acknowledge how the values that do occur in unordered sets or in ordered arrays can be combined by chance. For single-valued data, D_e is as in (3) but now in coincidence matrix notations

$$D_e = \sum_b \frac{n_b \cdot}{n..} \sum_c \frac{n._c}{n.. - 1} d^2_{bc}\ . \tag{30}$$

In the case of unordered sets, consideration has to be given to the observed proportion $P_{(q)}$ of pairable q-valued sets of values in the reliability data, to the observed numbers n_b of values b available for forming sets of size q, without duplications, and to the common metric of the values in unordered sets, which is reflected in the choice of the single-valued difference ${}_{\text{metric}}d_{bc}$. With q_B as the number of values in the set B, $q_B = 0, 1, 2, \ldots,$ $B_{(q)}$ as a q-valued set B, and $\sum_{B_{(q)}}$ as enumerating all sets B that contain exactly q values, for unordered sets, the expected disagreement D_e is:

$$D_e = \sum_{q_B} P_{(q_B)} \sum_{q_C} P_{(q_C)} \sum_{B_{(q_B)}} \sum_{C_{(q_C)}} \frac{\prod_{b \in B} n_b \prod_{c \in C \cap \bar{B}} n_c \prod_{c \in C \cap B} (n_c - 1)}{\sum_{B_{(q_B)}} \sum_{C_{(q_C)}} \prod_{b \in B} n_b \prod_{c \in C \cap \bar{B}} n_c \prod_{c \in C \cap B} (n_c - 1)} {}_{\text{metric}} d^2_{BC} \tag{31}$$

The products essentially enumerate each q-valued set. If there are w values to choose from, there are w single-valued sets, w(w-1)/2 two-valued sets, w(w-1)(w-2)/6 three-valued sets, and w!/[(w-q)!q!] q-valued sets, for each of which, (31) computes the probability of being formed by chance times the appropriate difference function.

In the case of ordered arrays of values, each value in any one position may cooccur with each value in any other position. The expected disagreement D_e then is the average multi-valued difference of all arrays that are possible, given the numbers n_{bt} of bs in each component t.

$$D_e = \frac{1}{(n(n-1))^z} \sum_{b_1}\sum_{c_1} \cdot \sum_{b_2}\sum_{c_2} \cdots \sum_{b_t}\sum_{c_t} \cdots \sum_{b_z}\sum_{c_z} \prod_{t=1}^{z} (n_{b_t}(n_{c_t} - \vartheta_{b_t c_t}))d^2_{<b><c>} \qquad (32)$$

where ϑ_{bc} is as in (14).

Generalizations to many variables. Several variables may be aggregated to form a single one. Aggregation amounts to adding the cell contents of the reliability data matrices from each variable to be aggregated, always yielding multi-valued data. Aggregated variables are approached with the multi-valued difference functions (20) through (28), as discussed above. As far as the computation of alphas is concerned, no additional provision needs to be made.

A common reliability measure for aggregated variables is desirable when several variables subsequently are analyzed together, specifically, when the conclusion from one is dependent on the conclusion from another. This is the case when a number of variables go into the definition of an index, when two or more variables enter the test for a hypothesis, or when computing regression equations and similar relations between variables. Under these conditions, variables depend on each other and may not be able to afford an unreliable variable among them. Without aggregate alpha-measures, it is recommended to take the lowest reliability among that set of jointly analyzed variables as the joint reliability of these variables. This is consistent with the practice of dropping variables that do not achieve desirable levels of reliability. The choice of the metric for the available multi-valued difference functions is determined by the nature of the data, whereby it is noteworthy, considering the power of these metrics – nominal < ordinal < interval < ratio – that difference functions must be of a power equal to or lower than that of the data. The choice of an appropriate multi-valued difference functions depends on the kind of reliability needed, which is a function of how data are analyzed once they passed the reliability tests. Here are some guidelines.

- When, in subsequent analyses of the data, the values are averaged within variables, the variance within multi-valued descriptions of units might well be irrelevant to the conclusions drawn from such an analysis and a measure of reliability that does not discount this variance would overstate the unreliability in the data. In this situation, the use of core differences is suggested. As stated above, core differences are single-valued differences between the most representative values of each set of values. They, like averages, ignore the variance within multiple descriptions of the units of analysis.
- When values are analyzed as unordered sets, then the average difference functions are suggested as an appropriate form for expressing the reliability of aggregated variables. This family of difference functions looks for agreements across different sets and captures the variance that the core difference functions ignore.
- When aggregated variables have no missing values and form, hence, arrays of the same number of components, multi-metric difference functions are recommended. By standardizing each single-valued difference with the expected disagreement within each variable, this function assures that each variable makes the same contribution to reliability. This function also allows the researcher to weigh the

constituent variables differently, accommodating any unequal impact of variables on the conclusion drawn from the data.

Multi-metric difference functions, as their name suggests, allow each variable to have their own (nominal, ordinal, interval, or ratio) metric. By contrast, core and average differences require all values to have the same metric.

Although it is tempting to use this form of aggregation to evaluate the reliability of a whole multi-variate measuring instrument, the results may well be deceptive as highly reliable variables can overshadow unreliable variables and may lead to global acceptance of locally unreliable data.

- When reliable variables must not be allowed to compensate for unreliable variables in the data, the absolute difference function provides the most appropriate form of aggregation. It provides for the toughest reliability test. It counts any mismatch, large or small, as disagreement. Just as the nominal metric difference ignores all shades of agreement when data are single-valued, the absolute difference ignores all shades of agreement when data are multi-valued.

The work reported here is still in progress. Computer implementation and testing is planned. The hope is to create an extremely versatile analytical device for the analysis of the reliabilities in content analysis, survey research, and a variety of other data generating procedures.

References

Bennett, E.M., R. Alpert, and A. C. Goldstein (1954). Communication Through Limited Response Questioning. Public Opinion Quarterly, 18: 303-308.

Brouwer, M., C. C. Clark, G. Gerbner, and K. Krippendorff (1969). The Television World of Violence. Pages 311-339, 519-591 in R. K. Baker and S. J. Ball (Eds.). Mass Media and Violence, Vol. IX. A Report to the National Commission on the causes and prevention of violence. Washington DC: U.S. Government Printing Office.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Bulletin, 76: 378-382.

Cohen, J. (1968). Weighted Kappa. Psychological Bulletin 70,4: 213-220.

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

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, 76,5: 378-382.

Goodman, L. A., and W. H. Kruskal (1954). Measures of Association for Cross Classification. Journal of the American Statistical Association, 49: 733-764.

Krippendorff, K. (1967). A Computer Program for Analyzing Multivariate Agreements, User's Manual. Philadelphia: The Annenberg School for Communication, University of Pennsylvania, Mimeo [Version 2, 1970; Version 3, 1972].

Krippendorff, K. (1970). Bivariate Agreement Coefficients for Reliability of Data. Pages 139-150 in E. R. Borgotta and G. W. Bohrnstedt (eds.). Sociological Methodology 1970, Vol. 2. San Francisco, CA: Jossey-Bass, Inc.

Krippendorff, K. (1971). Reliability of Recording Instructions: Multivariate Agreement for Nominal Data. Behavior Science, 16: 222-235.

Krippendorff, K. (1978). Reliability of Binary Attribute Data. Biometrica, 34: 142-144.

Krippendorff, K. (1980). Content Analysis; An Introduction to its Methodology. Beverly Hills, CA: Sage Publications.

Krippendorff, K. (1987). Association, Agreement and Equity. Quality and Quantity, 21: 109-123.

Krippendorff, K. (1991). Reconstructing (some) Communication Research Methods. Pages 113-142 in F. Steier (Ed.). Research and Reflexivity. London: Sage Publications.

Kuder, G. F., and M. W. Richardson (1937). The Theory and Estimation of Test Reliability. Psychometrica, 2: 151-160.

Mahalanobis, P. C. (1936). On the Generalized Distance in Statistics. Proceedings of the National Institute of Science (India), 12: 49-55.

Maxwell, A. E. and A. E. G. Pilliner (1968). Deriving Coefficients of Reliability and Agreement for Ratings. British Journal of Mathematical and Statistical Psychology, 21: 105-116.

Pearson, K. (1901). Mathematical Contributions to the Theory of Evolution. IX: On the Principle of Homotyposis and its Relation to Heredity, to Variability of the Individual, and to that of Race. Part I: Homotyposis in the Vegetable Kingdom. Philosophical Transactions of the Royal Society (London). Series A, 193, 358-479.

Perreault, W. D., and L. E. Leigh (1989). Reliability of Nominal Data based on Qualitative Judgements. Journal of Marketing Research, 26: 135-148.

Scott, W. A. (1955). Reliability of Content Analysis: The Case of Nominal Scale Coding. Public Opinion Quarterly, 19: 321-325.

Tildesley, M. L. (1921). A First Study of the Burmese Skull. Biometrica, 13: 176-267.