Running Head: COEFFICIENT OF CO-TERMINATION

Assessing Co-termination Between Coders in Unitizing Textual Data:

A Multi-response Randomized Blocks Permutation Approach

Li Cai

School of Journalism and Communication

The Ohio State University

Student Paper

Li Cai recently graduated from the master's program in Journalism and Communication at the Ohio State University, and will soon join the Ph.D. program in Quantitative Psychology at the University of North Carolina – Chapel Hill. Address: L. L. Thurstone Psychometric Lab., Department of Psychology, UNC – Chapel Hill, CB #3270 Davie Hall, Chapel Hill, NC 27599-3270. Phone: 614-596-9197. Email: lcai@email.unc.edu.

Abstract

The assessment of intercoder agreement in the unitizing phase of content analyses has long been overlooked. In particular, little attention has been paid to the issue of co-termination. Although multiple-coder kappa can be used for the purpose of summarizing the agreement of co-termination, its conservativeness often results in gross underestimates. A new family of coefficients based on Multi-response Randomized Blocks Permutation procedure is presented here and numerical examples are given.

Introduction

Content analysis is a quantitative research methodology widely employed in the field of communication. Berelson's (1952) often cited definition of content analysis as an objective, systematic, and quantitative endeavor to describe the content of communication messages clearly endorsed the usefulness of content analysis to communication scholars. In a recent "content analysis of content analyses" of articles published in *Journalism and Mass Communication Quarterly* between 1971 and 1995, Riffe and Freitag (1998) located 486 full-length reports using content analysis, comprising of roughly one fourth of the total number of articles published. An earlier study by Wilhoit (1984) suggested that more than 20 per cent of theses and dissertations listed in *Journalism Abstracts* used content analysis. Moffett and Dominick (1987) reported a similar result that 21 per cent of the articles published in *Journal of Broadcasting* between 1970 and 1985 employed content analysis.

As Krippendorff (1980) succinctly remarked, making "inferences from essentially verbal, symbolic, or communicative data" has always been at the heart of content analysis (p. 20). In order for scientific inferences to be valid, one must first ascertain the reliability of the research instrument. Just as chemists could ill-afford an uncalibrated balance in a chemical experiment, one could hardly imagine living a life in the communication scholarship without assessing the reliability of content analysis. Of course, if a stream of messages is to be analyzed by a well-designed computer program, one could probably worry less about reliability, but in most instances, if not all, content analyses still require much human labor, and thus the probable errors of human analysts become almost inevitable.

The standard process of content analysis, as described in most introductory communication research methods textbooks (e.g. Frey, Botan, & Kreps, 2000; Stewart, 2002; Wimmer & Dominick, 1994) essentially involves a *coding* process. Guetzkow (1950) observed that any transformation of qualitative data into a form “susceptible to quantitative treatment constitutes *coding*” (p. 47). He further emphasized that the coding process could be broken down into two related phases, that of separating the qualitative material into units, and that of classifying the unitized data into established categories. The former is often termed *unitizing*, and the latter *categorizing*. The two processes are integral elements of content analysis yet require different strategies of reliability assessment. One term that needs a little clarification is *categorizing*. In many practical situations the coded units are indeed later classified into categorical sets, but this is not necessarily true. Coded units may be rated on an ordinal/interval or ratio scale in subsequent analyses. The term *categorizing* will remain in use in the paragraphs to follow, but without any implication of merely categorizing the coded units into qualitative (nominal) sets.

In the coding process, usually a set of human *coders* or *judges* are involved. The assessment of reliability of the content analysis thus becomes an assessment of the reliability of the coders, even though this is not a sufficient condition for the entire content analysis study to be reliable, the coding process is of such importance that low intercoder reliability would render all subsequent analyses meaningless, because low intercoder reliability would suggest that the obtained results were largely not replicable (Krippendorff, 1980, p. 131). Ideally, the coders should be trained to rate or judge the content independently and yet to arrive at the same ratings in precisely in the same manner as intended by the coding scheme. Intercoder reliability is established when the same pieces (possibly a very large number) of content yield same ratings from independent coders using a common data language (Krippendorff, 1980, p. 133). Formally, the term intercoder reliability should be more appropriately termed *intercoder agreement* (see Lombard, Snyder-Duch, & Bracken, 2002), but the two terms will nevertheless be used interchangeably in the pages to follow since given the present context the means of two terms are not much different.

Another distinction that should be made is the "depth" at which the messages are to be coded. Berelson (1952) clearly intended content analysts to deal only with the manifest content, i.e. the information "as is," without invoking additional mental efforts of the coders to discover the latent content or the implied meaning. However, unless the research question can be easily answered by simply counting the number of words in a newspaper article or the number of occurrences of the names of candidates in pre-election news coverage – which can be quite easily done with a computer – the coding process will often require the coders to make subjective judgments. Under those circumstances, readers of the research would demand the researchers to show that "those judgments, while subjectively derived, are shared across coders," which again confirmed the practical necessity of establishing intercoder agreement in content analysis (Potter & Levine-Donnerstein, 1999, p. 266).

Having illustrated the importance of intercoder agreement, the current status of correctly using and reporting intercoder agreement measures in communication journals is quite alarming. Riffe and Freitag (1997) found that only half of the 486 articles published in *Journalism and Mass Communication Quarterly* between 1971-1995 reported intercoder reliability. A recent study by Lombard et al. (2002) searched virtually all content analysis articles indexed in *Communication Abstracts* from 1994 to 1998. Of the 200 articles they found, only 69% ever mentioned intercoder reliability, and usually the methods for computing intercoder reliability were not reported. Of the 44% of all articles that did report the names of the specific methods, more than half of them relied on liberal indices that are not chance-corrected, such as percent agreement, which seriously undermined the effort of computing and reporting reliability coefficients (Lombard, et al., 2002, p. 596).

Given the current undesirable state of affairs of appropriately using and reporting intercoder agreement indices in the communication scholarship, the next section shall explicate intercoder agreement in the context of a two-stage coding process, namely, unitizing and categorizing. The importance of *co-termination* when unitizing textual data shall be presented.

Co-termination

When Guetzkow (1950) wrote about unitizing and categorizing, he presented a convincing case that in order for the entire content analysis to be reliable, one has to regard the assessment of the overall intercoder agreement as a two-stage process. Ideally one should compute agreement measures for unitizing first and then calculate the agreement indices for categorizing, with the "overall" reliability referring to the combined intercoder agreement in both stages. One should note that this overall reliability does not always have to be expressed in quantitative terms. It is possible that a particular content analysis consists merely of categorizing existing units, and then this two-stage notion would not be relevant. However, there are times when unitizing is a must, and under such circumstances, the intercoder agreement of unitizing becomes crucial. This paper does not attempt to develop any new agreement indices for the categorizing phase, as there are established methods already. Instead, the aim is on how the agreement of unitizing can be better summarized, and this goal cannot be achieved without first understanding the complexities of intercoder agreement in the unitizing phase.

The problem of agreement of unitizing focuses on how independent coders choose breaking points at various places in a continuous *segment* of textual content, be it a sentence, a paragraph, an article, or an entire television show. The segments are assumed to be clearly delineated from one another and are usually naturally given. This assumption is not unfair because most of the qualitative content that can serve as segments for coding has unambiguous endpoints. For example, if a newspaper article is chosen as a segment, where it ends is crystal clear. It is further assumed that segments are of two types, discrete and continuous. The reader should take note that such a distinction is probably quite artificial. The only reason behind such a distinction is the mathematics.

Discrete segments are composed of a finite number of *elements*. Defining what is an *element* is difficult, because it ultimately depends on the research question, and how detailed the researcher would like the content analysis to be, but an example should help illustrating this concept. For instance, a sentence from an online chat transcript is selected: "Apparently, from what I read, they haven't identified the dead body yet." It is convenient to define a word – anything in between two spaces – as an *element*. Therefore, this is a segment containing 12 elements. Thus defined, unitizing becomes an operation of grouping elements into units, and intercoder disagreement arises when coders define the groups differently.

The idea of elements is not applicable to continuous segments. For instance, a researcher may want to unitize audio/video recordings. It is probably hard to define what an element is within a continuous stream of audio/ video recording, but it is easy to deal with the relative *length* of a unit, perhaps expressed in terms of time. One can imagine the coder using a stopwatch to record the lengths of units, and intercoder disagreement occurs when the coders come up with different length readings. The idea of length is widely applicable and it is easy to see that one can actually express the discrete type of unitizing using lengths as well.

Consider this example: ABCD, a discrete segment of 4 elements, was to be coded by two judges by putting slashes at the breaking points. Judge 1 gave: A/B/CD, and judge 2 gave: A/BC/D. They both came up with three units for this segment, and they were said to be co-terminus for the first unit. The reliability data, using discrete terms, can be thought of as a set of binary streams: 1 1 0, for judge 1; and 1 0 1 for judge 2. The number of entries in the binary stream is the number of elements minus 1, representing the number of possible breaking points.

The 1s in a stream signify observed breaking points. The same data can be expressed in terms of lengths: 1 1 2 for judge 1; and 1 2 1 for judge 2. The numbers correspond to the number of elements in a particular unit, and the total number of entries equals the number of units.

Having defined the terms, it is natural to introduce the concept of *co-termination* and review what Guetzkow (1950) recognized as the two kinds of errors likely to be present when unitizing a stream of content: (1) failure to agree on the breaking points between the units, and (2) failure to attain the same number of units (p. 54). Co-termination, or co-terminability, a term introduced but not clearly defined in Guetzekow (1950), refers to the agreement among pairs of coders to break a given segment of content at the same points into the same number of smaller units. Note that this definition essentially contains two components: (1) the agreement on the breaking points, and (2) the agreement on the number of units. Such a definition of co-termination is said to be in a "strong" form because there will be perfect agreement of unitizing among coders when the strong form of co-termination is achieved. It is the necessary and sufficient condition for a "weaker" form to exist because it is possible that a pair of coders agree partially, such as for the first unit in the afore mentioned example, on how to choose the breaking points and yet at the same time do not agree on how many units there are in the segment of content. An example should help illustrating this point. Suppose two coders were instructed to break an article into smaller units containing one or more paragraphs. The two coders started out in perfect agreement as to how to group the paragraphs into units up to a certain paragraph after which things started to fall apart. As a result, the numbers of units were different, and certainly by definition of strong co-termination, they failed to achieve agreement. However, one has to acknowledge that at least the two agreed somewhat in the beginning, and a good agreement measure should give partial credit to what they agreed upon. It is conceivable that any measure of agreement based on the strong form of co-termination would necessarily be a conservative one and thus the existence of a weak form of co-termination is not an idea plucked out from the thin air.

The weak form of co-termination essentially depends on the sequential nature of content streams, i.e. one can only start unitizing from the beginning of a segment and proceed as the stream goes. Of course, going backwards from the end is not impossible, but this is makes little difference because one can then define the end as the beginning. Expressed in discrete terms, the weak form of co-termination between two-coders is defined as choosing breaking points so that at least the two coders grouped one set of elements in the same manner. Consider the first example again: a segment – ABCD, with 4 elements, and 3 coders were to unitize it. The result happened to be as follows: coder 1 – A/B/CD, coder 2 – AB/C/D, and coder 3 – A/B/C/D. There are three distinct pairs of coders: 1 vs. 2, 2 vs. 3, and 1 vs. 3. Clearly, none of the pairs achieved co-termination if the strong definition is used. Coders 1 and 2 gave the same number of units but were not co-terminus. Although coders 1 and 3 gave different numbers of units (3 and 4, respectively), they actually attained the weak form of co-termination for the groupings of elements A and B into the first and second unit. For coders 2 and 3, they achieved co-termination for C and D. The basic conceptualization of the measurement of co-termination would be to employ the strong definition when the coders agree on the number of units and to use the weak form when the numbers of units are different.

It is worthy of pointing out that according to Hubert (1977) there are three definitions of agreement when the number of coders goes beyond two: DeMoivre's definition, target-rater definition, and pair-wise definition. The first one refers to the unanimous agreement of all coders, and the second one refers to the joint agreement of all other coders with a "target-rater" who provides the "true" rating, and the third, which is also what is implied in the definition of co-termination, refers to the agreement between any pairings of coders. It is easy to see that

DeMoivre's definition tends to yield the most conservativeness. Most of the popular intercoder agreement indices that can handle three or more coders use the pair-wise definition, as does the new coefficient to be proposed in subsequent sections.

Having defined what co-termination is, it is not difficult to infer that the mere agreement on number of units does not imply co-termination. As to the relative importance of the two, Guetzkow (1950) remarked that the failure to achieve "co-terminability" is less likely to lead to confusions and low intercoder reliability in the subsequent categorizing of the coded units (p. 55). There is absolutely nothing wrong with this argument, because how far reliability assessment should go is a practical matter related to the nature of the specific study at hand. If the unit boundaries are relatively clear, or if slight inconsistencies in co-termination do not significantly affect the subsequent use of the coded units, one could worry less about co-termination and focus more on achieving a high level of agreement on the number of units. However, there are certain times when disagreement in co-termination may lead to different interpretations of the same data, even though the number of units are the same across coders. For instance, if two coders were to divide the sentence "Apparently, from what I read, they haven't identified the dead body yet," and the coders agreed that it contained two units, but the first coder put the division mark right after "apparently," while the second put it after "read." The interpretation of the two units would necessarily be different, because a stand-alone "apparently" would suggest confirmation, while "apparently, from what I read" would refer to the clear inferences that the chat user could make from what he or she read. This example is only a very trivial one. What is important is to realize that the mere agreement on number of units does not automatically imply reliability of unitizing.

Still using the previous example, suppose that the first coder divided the sentence after both "apparently" and "read," and the second coder only divided the sentence after "apparently," the number of units for the two coders are 3 and 2, respectively, and there seems to be much disagreement between the two, but in fact they did achieve co-termination, at least for the first unit. Given such results, at least the interpretation for the first unit – "apparently," is unambiguous. In the next section, the five most widely used indices of intercoder agreement shall be briefly reviewed. Most of them are intended for bivariate nominal level coding, and for discrete reliability data (binary streams) between two coders Cohen's κ can be used, but only to a limited extent because of the resulting gross underestimate of reliability, which will become clear in the sections to follow. However, for continuous content, no current indices are directly applicable and a new generalized measure based on Multi-response Randomized Blocks Permutation procedures (Mielke & Iyer, 1982) shall be presented.

Popular Indices of Intercoder Agreement

Percent Agreement and Holsti's Method

This is perhaps the most easily understood method for calculating intercoder agreement. It is simply the "percentage of all coding decisions made by pairs of coders on which the coders agree" (Lombard, et al., 2002, p. 590). This is not a chance corrected measure, and Krippendorff (1980) illustrated how chance could artificially inflate percent agreement with a neat example (pp. 133-135). In general, using percent agreement is a very poor practice that inflates reliability, and is not applicable to other higher levels of measurement than nominal level coding.

Holsti (1969) proposed a variation of the percent agreement measure, which is the same as percent agreement when two coders are coding the same segments of content. This is still not a chance-corrected measure and it suffers from the same drawbacks as percent agreement. It is interesting to note that even though some statisticians have argued against the use of chance-corrected measures (e.g., Goodman & Kruskal, 1956), supporters of chance-corrected measures "far outweigh detractors" (Berry & Mielke, 1988, p. 922).

Scott's π

This is a chance-corrected index first introduced by Scott (1955) primarily in the context of coding qualitative data obtained from surveys. In its original form, this index is only applicable to univariate nominal level coding and accommodates only two coders, although it is worth mentioning that Craig (1981) has given an extension of Scott's π to the case of multiple coders. Scott'*s* π is the first coefficient that considers both the number of categories and the marginal distributions, i.e. how the two coders distribute their classifications of the units. However, the π coefficient not only assumes that the column and row marginal distributions are identical to the "true" proportions, but also takes it a step further by assuming that the two coders share the same marginal distributions. Given the context of survey research, the former assumption is not unreasonable, as the "true" proportions are usually obtainable, and this assumption has given Scott's π a distinct edge over similar coefficients like Cohen'*s* κ, because π can still be computed when the two coders have coded different subsets of the content, while computation of κ requires that the pair of coders have coded the same units (Craig, 1981, p. 261). However, it is precisely the latter assumption of π that is more problematic. As Cohen (1960) pointed out, "one source of disagreement between a pair of judges is precisely their proclivity to distribution their judgments differently over the categories" (p. 41). Furthermore, the "true" proportions are not always available, thus making such an unrealistic assumption only hinders the general practicability of the π coefficient.

Cohen's κ

Cohen's (1960) κ is defined in much the same way as Scott's π. Usually it is assumed that two coders independently classify each of the n units into one of c established categories. The layout for computing such bivariate nominal level intercoder agreement as κ essentially involves the construction of a two-way cross-classification table, with entries in the table being the proportion of observations falling into one of the *c* by *c* cross-classifications. The marginal distributions are simply the column and row sums. The *κ* coefficient, and its variants for bivariate nominal data usually assumes the form of a ratio between observed and expected proportions $\kappa = (P_o - P_e)/(1 - P_e)$, with P_o given by the sum of the diagonal elements of the *c* by *c* cross-classification table, and P_e is found by first multiplying each column marginal with its associated row marginal and then taking the sum of the products.

Cohen's *κ* has enjoyed continued development by psychological methodologists. Cohen (1968) himself introduced a weighting procedure that accounts for the differential severity of disagreements. Fleiss (1971) gave its extensions to the case of multiple raters. Fleiss and Cohen (1973) established the equivalence of weighted *κ* and the intra-class correlation coefficient. Hubert (1977) introduced the underlying mathematical model of matching distributions in probability theory to users of the *κ* coefficient. Fleiss, Nee, and Landis (1979) worked out *κ*'s asymptotic variance. Conger (1985) extended it to measure agreement over time for continuous scales. As mentioned afore, *κ* can be used to assess co-termination for two coders and discrete type of unitizing, but results in an underestimate.

Krippendorff's α

When the number of coders is exactly two with nominal level coding assumed, Krippendorff's (1970) *α* coefficient is identical to Scott's *π* (cf. Krippendorff, 1980, p. 138). What makes the *α* coefficient more appealing than its competitors is that it offers an easy extension to measure the agreement of higher levels of measurement and of multiple coders. Recall that Guetzkow (1950) described the two kinds of errors in unitizing textual data. It appears that Krippendorff's *α* coefficient may well serve the purpose of calculating the intercoder agreement of the number of units of a given segment of content, but it is still lacking in its ability to detect co-termination for the continuous case.

Multi-response Randomized Blocks Layout and Intercoder Agreement

This section describes some of the details of the Multi-response Randomized Blocks Permutation procedure (MRBP) relevant to the assessment of agreement, and the details of the equations can be skipped without loss of continuity.

Multi-Response Permutation Procedure (MRPP) is a versatile analytic framework first outlined in Mielke, Berry, and Johnson (1976) as a robust and powerful tool for analyzing multivariate data from randomized experiments using the average within-group distance as the test statistic (descriptions of MRPP are given in Appendix I). The usual form of the symmetric distance function between two multi-dimensional responses is given by

$$\Delta_{I,J} = \left[\sum_{c=1}^{r} (x_{cI} - x_{cJ})^2 \right]^{v/2}, \tag{1}$$

where $(x_{1I}, \ldots, x_{rI})$ denote one r-dimensional response, I and J are distinct integers from 1 to N, and N is the total number of responses, i.e. the total sample size. It is easy to see that the distance between two multivariate responses is a power function of the summation of the squared distances between each dimension and therefore the choice of v gives rise to a variety of distance functions. The value of v determines the analysis space of the test and choice is somewhat arbitrary, but the most widely used two are $v = 1$ and $v = 2$, which corresponds to metric Euclidean distance and non-metric squared Euclidean distance. Some of the most widely employed tests such as the t-test, *ANOVA*, and their multivariate counterparts – Hotelling T^2, and Bartlett-Nanda-Pillai trace test in *MANOVA* all use squared Euclidean analysis space. Berry and Mielke (1988) pointed out that the choice of squaring the distances is "questionable at the best" (p. 922). They suggested $v = 1$ be used at all times based on its robustness against outliers, but

Janson and Olsson's (2001) modified statistic uses $v = 2$ and their main argument for the more conventional metric is the ease of interpretation. For binary streams of discrete reliability data, the choice does not matter because the square root of 1 is still 1, but for continuous content, as the reader will see later, $v = 2$ is the sometimes the only choice because of the vast reduction in computation time.

Multi-Response Randomized Blocks Permutation (MRBP) is a variation of MRPP statistic when a blocking variable is added into a one-way design. It is first introduced by Mielke & Iyer (1982) as a supplement to MRPP. In its original formulation, MRBP defines a b-block by g-treatment randomized experiment and within each block there is only one r-dimensional observation per treatment, taken as $n = 1$ for each cell. Using the MRBP layout, Berry and Mielke (1988) provided a re-formulation of Cohen's κ and a natural extension of κ to multiple coders as well as to higher levels of measurement.

In brevity, the original cross-classification layout of κ is transformed into a b-block by g-treatment MRBP layout. Assuming that two observers independently coded each of the g units into one of r categories. The usual cross-classification layout of κ would be an r by r table with the entries in the table being the proportions of cross-classifications in particular cells. The MRBP layout, however, would be a 2-block by g-treatment design with r-dimensional responses and v set to 1 when calculating distances. The extended measure of agreement is given by the equation

$$\kappa = 1 - \delta / \mu_{\delta}, \tag{2}$$

where δ denotes observed disagreement and μ_{δ} denotes expected proportion of disagreement by chance. Because MRBP is a based on permutation, μ_{δ} is found by permuting the data within each block across treatments. Such a formulation makes the extension of κ to higher levels of measurement and multiple coders easy.

Generally, assuming b coders independently rate g units of content, let $[x_{ijk}]$ denote the elements in an r-dimensional response vector (when $r = 1$, this is a scalar) from coder i for unit j, where $i = 1, \ldots, b, j = 1, \ldots, g$, and $k = 1, \ldots, r$, the within-treatment distance function is given by

$$\Delta_{p,q} = \left[\sum_{c=1}^{r} (x_{pjc} - x_{qjc})^2 \right]^{v/2}, \tag{3}$$

and the average within-treatment distance for all distinct pairs of coders is given by

$$\delta_{obs} = \left[g \binom{b}{2} \right]^{-1} \sum_{j=1}^{g} \sum_{p<q} \Delta_{p,q}, \tag{4}$$

where $p < q$ denotes the sum over all p and q such that $1 \leq p < q \leq b$, and basically this is to ensure that a response vector is not compared with itself. The definition of μ_δ reflects the addition of blocks because unlike MRPP, in randomized blocks designs data cannot be permuted across the blocks. Therefore the maximum number of permutations is $M = (g!)^b$ – the total number of permutations within each block to the bth power. Assuming the M permutations are equally probable, a theoretical definition of chance disagreement is given by

$$\mu_\delta = M^{-1} \sum_{i=1}^{M} \delta_i, \tag{5}$$

However, one does not need to enumerate all M permutations to arrive at μ_δ, a more efficient working formula for μ_δ is available due to the fact that the first moment of the permutation distribution is a constant multiple of g^2 elementary calculations (see Mielke & Iyer, 1982).

Using similar notations as in equations (3) and (4), then the within-and-between-units distance function between two rating vectors is given by

$$\Delta_{pi,qj} = \left[\sum_{c=1}^{r} (x_{pic} - x_{qjc})^2 \right]^{1/2} , \tag{6}$$

and the following equation gives the chance disagreement

$$\mu_\delta = \left[g^2 \binom{b}{2} \right]^{-1} \sum_{i=1}^{g} \sum_{j=1}^{g} \sum_{p<q} \Delta_{pi,qj} \, . \tag{7}$$

Equation (7) seems to be quite complicated. However, it is nothing but the average distance between any distinct pairings of response vectors. Berry and Mielke (1988) have named their extended κ coefficient as R, and have established the equivalence of this statistic with other known measures.

Formulation of the Proposed Coefficients of Co-termination

Assuming two coders are present, and they have broken a 7-word sentence into 3 units. The analysis of intercoder agreement, using discrete terms, may be expressed as a 2 block by 6 treatment MRBP layout. The entries are just 0s and 1s, and the design is summarized in Table 1.

Table 1

Example dataset

	Treatments					
Blocks (Coders)	1	2	3	4	5	6
1	1	0	0	0	1	0
2	1	0	0	0	0	1

If this sentence is of the form ABCDEFG, then the codings in Table 1 are: A/BCDE/FG for coder 1, and A/BCDEF/G for coder 2. If the cross-classification layout of κ is used, the design should be a 2 by 2 cross-classification table, and it would look like Table 2.

Table 2

Cross-classification Layout

	Coder 1		
Coder 2	1	0	Sums
1	1	1	2
0	1	3	4
Sums	2	4	6

Cohen's κ can be calculated from Table 2 using the usual way.

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{(1/6 + 3/6) - [(2/6 \times 2/6) + (4/6 \times 4/6)]}{1 - [(2/6 \times 2/6) + (4/6 \times 4/6)]} = .25$$

More generally, let $[x_{ij}]$ represent the value (0 or 1) from the cell corresponding to the *i*th block and *j*th treatment, where $i = 1, \ldots, b$, and $j = 1, \ldots, g$, the symmetrical MRBP distance function between any two cells $[x_{ij}]$ and $[x_{pq}]$ in a table similar to Table 1 can be simplified to

$$\Delta_{ij,pq} = (x_{ij} - x_{pq})^2 . \qquad (8)$$

Using equations (3) – (7), a reformulated κ can be expressed as

$$\kappa' = 1 - \frac{\delta'}{\mu_\delta'} \qquad (9)$$

where δ' is the average within-treatment disagreement and μ_δ' – the expected disagreement – can be found by averaging over all δ's obtained from permuting data within blocks. Table 3 is an example of a possible permutation. For instance, in Block 1, the 1s originally in the 1st and 5th treatments are swapped into the 2nd and 3rd places. For this permutation δ' = 4/6 = 2/3.

Table 3

A Possible Permutation

	Treatments					
Blocks (Coders)	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	0	0	1

To summarize, δ' can be calculated as (0+0+0+0+1+1)/6 = .333, μ_δ' = .444, and κ' is 1 – .333/.444 = .25, which is exactly the same as using the cross-classification table, but the MRBP approach can be easily extended to multiple coders.

The problem with this approach, as the reader probably has already noticed, is an underestimate of reliability. Without calculating any statistics, a visual examination of the codings: A/BCDE/FG for coder 1, and A/BCDEF/G for coder 2, reveal the fact that the codings are not much different yet the agreement measure indicates that it is only 25% agreement above chance, a value too low by any standards. Therefore, a remedy shall be presented and it makes use of the notion of continuous content.

One can express the data in Table 1 using lengths and the result is summarized below. For the moment, the reader is asked to ignore the lines corresponding to the "Cumulative" lengths. The usefulness of these identities will become clear later.

Table 4

Continuous measurement

	Treatments		
Blocks (Coders)	1	2	3
1	1	4	2
Cumulative 1	1	5	7
2	1	5	1
Cumulative 2	1	6	7

One can essentially apply equations (3) – (7) on the two blocks and follow computational formulae in Mielke and Iyer (1982) to obtain the reliability coefficient. With $v = 1$, $\kappa = 1 - \delta_{obs} / \mu_\delta$ = 1 – .667 / 1.778 = .625. With $v = 2$, $\kappa = 1 - \delta_{obs} / \mu_\delta$ = 1 – .666 / 5.111 = .87. One can see that by using continuous content, the agreement index is increased quite a bit. However, the direct application of continuous content is quite problematic given that coders usually do not agree on the number of units either. Not only the computational formulae are rendered useless, there are conceptual problems too. Consider the following coding:

Table 5

Continuous measurement

	Treatments						
Blocks (Coders)	1	2	3	4	5	6	7
1	10.3	19.7	10.1	9.8	4.6	5.0	10.5
Cumulative 1	10.3	30.0	40.1	49.9	54.5	59.5	70.0
2	10.2	19.8	8.5	12.0	9.5	10.0	–
Cumulative 2	10.2	30.0	38.5	50.5	60.0	70.0	70.0

The first coder came up with 7 units for this continuous piece of content and the second one came up with only 6. One can easily think of replacing the missing cell in the last column with zero and applying the computational formulae thereafter, but then a theoretical problem arises because when the permutation within the second block is conducted, that imputed zero may appear, for example, in the 3rd column. It makes little sense because one can hardly imagine a unit of length zero in between two other units of positive lengths. After all, the communication content is sequential stream that does not stop until the endpoint.

This problem can be offset by fixing the trailing zero(s), should there be one or more missing cells in the last a few columns, when conducting the within block permutations. Therefore, the total number of possible permutations in the given example is only (7!)(6!) = 3,628,800, instead of $(7!)^2$ = 25,401,600, as the trailing zero at $[x_{27}]$ will remain un-permuted. More generally, the total number of permutations is given by

$$M = \prod_{j=1}^{b} g_j!, \tag{10}$$

and when all g_j's are equal, equation (10) is equal to $(g_{max}!)^b$. This change essentially reflects the usefulness of the so-called *reference subsets* described in Edgington (1987). If the set of $(g_{max}!)^b$

data permutations is taken as the primary reference set, then the computation of μ_δ under the condition when all g_js are equal would be using the reference distribution for the general-null hypothesis, whereas when not all g_js are equal, and thus equation (10) yields a smaller value than $(g_{max}!)^b$, the computation of μ_δ would be comparable to the test of a restricted mull hypothesis (see Edgington, 1987, pp. 305-316).

When not all g_js are equal, it is useful to define the following computational expressions for computer implementation. A GAUSS (Aptech Systems Inc., 1997) procedure which implements these formulae is in Appendix II.

Let x_{pi} denote a row of g_i unit length data from coder i, where $p = 1, \ldots g_i$, and let M be given as in equation (10), define:

$$C_1(i,j) = \frac{M}{\max(g_i, g_j)}, \tag{11}$$

$$C_2(i,j) = M\left[1 - \frac{\min(g_i, g_j)}{\max(g_i, g_j)}\right], \tag{12}$$

$$C_3(i,j) = \begin{cases} i & \text{if } g_i > g_j, \\ j & \text{if } g_i < g_j. \end{cases} \tag{13}$$

$$D_1(i,j) = \sum_{p=1}^{g_i} \sum_{q=1}^{g_j} \Delta_{pi,qj}, \tag{14}$$

$$D_2(i,j) = \sum_{p=1}^{g_{C_3(i,j)}} x_{pC_3(i,j)}, \tag{15}$$

$$E(i,j) = C_1(i,j)D_1(i,j) + C_2(i,j)D_2(i,j), \tag{16}$$

$$\mu_\delta = M^{-1} \sum_{i<j} E(i,j), \tag{17}$$

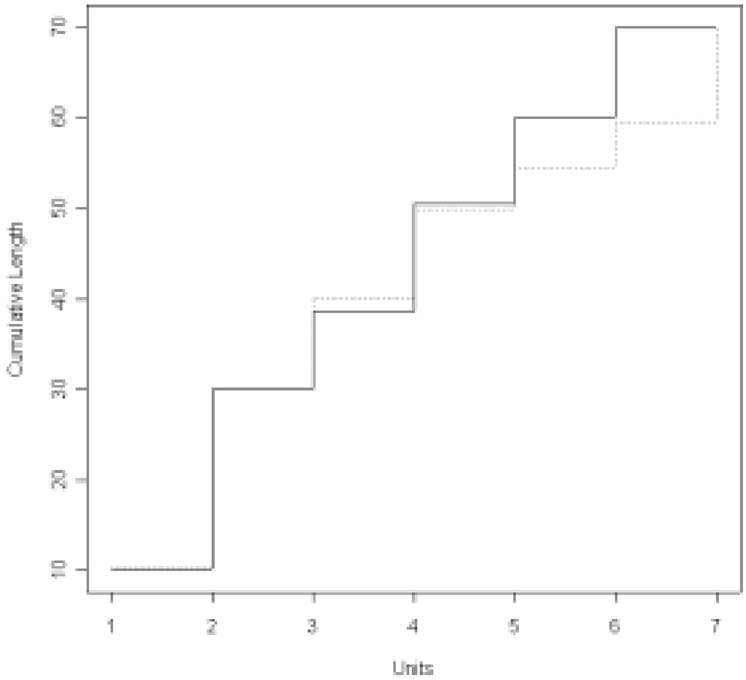
where $i < j$ denotes the summation over all i and j such that $1 \le i < j \le b$.

With $v = 1$, $\kappa = 1 - \delta_{obs} / \mu_{\delta}$ = 1 – 3.486/ 5.265 = .338. With $v = 2$, $\kappa = 1 - \delta_{obs} / \mu_{\delta}$ = 1 – 23.811/ 50.174 = .525.

A conceptually much simpler way makes use of the “Cumulative” lengths. Borrowing the concept of empirical cumulative distribution function (ECDF) in elementary mathematical statistics, one can envision the disagreement between two coders when unitizing continuous content as the difference between two cumulative length functions. A plot should help illustrating this point.

Figure 1

Two Cumulative Length Functions



The dotted line corresponds to coder 1 and the solid line corresponds to coder 2. In Table 5, when coder 2 has used up all available content, the cumulative length is 70.0 and it remains at 70.0 regardless of how many missing cells there may be.

Using similar notations as above, let $[x_{ij}]$ represent the value from the cell corresponding to the *i*th Block and *j*th Treatment, where $i = 1, \ldots, b$, and $j = 1, \ldots, g$, where g is the maximum number of units given by the coders. One may consider $g = \max(g_1, g_2, \ldots, g_b)$, and the squared distance ($v = 2$) between two cumulative lengths corresponding to cells $[x_{ij}]$ and $[x_{kj}]$ in a table similar to Table 5 can be expressed as

$$\Delta_{ij,kj} = \left[\left(\sum_{p=1}^{j} x_{ip}\right) - \left(\sum_{p=1}^{j} x_{kp}\right)\right]^2 = \left[\sum_{p=1}^{j}(x_{ip} - x_{kp})\right]^2, \qquad (18)$$

and the average observed disagreement is

$$\delta = \binom{b}{2}^{-1} \sum_{i<k} \sum_{j=1}^{g} \Delta_{ij,kj}, \qquad (19)$$

where $i < k$ denotes the summation over all i and k such that $1 \le i < k \le b$.

The computational formulae for expected disagreement are so cumbersome that precludes presentation here due to the complexities involved. However, a GAUSS (Aptech Systems Inc., 1997) procedure that implements this method is given in Appendix III. If one is willing to assume equal probability being placed on every permutation of these cumulative lengths, everything else then follows as what Berry and Mielke (1988) described.

The numerical results for the example datasets in Tables 4 and 5 are as follows:

$\kappa = 1 - \delta_{obs} / \mu_{\delta} = 1 - 1 / 10.222 = .902$.

$\kappa = 1 - \delta_{obs} / \mu_{\delta} = 1 - 143.43 / 551.197 = .74$.

Tests of Significance

Since κ is merely a linear function of δ_{obs}, a test of significance of κ is equivalent to the test of δ_{obs}. Mielke and Iyer (1982) gave formulae for the first three moments of the MRBP null distribution, and using the mean and variance, δ_{obs} can be standardized and the associated probability of δ_{obs} can be approximated via a Pearson type III distribution (see Mielke and Berry, 2001). This p-value is associated with the test whether κ is significantly different from zero. Since no random sampling assumption is involved, this test of significance is non-asymptotic and is different from what most computer packages do. Each one of the n segments in a reliability study would therefore have a p-value, and by looking at the set of p-values, the researcher should be able to infer whether the coders' overall agreement is due to chance or not.

A test of significance may also be conducted for the coefficient that uses the cumulative lengths via a random sample of all possible permutations (see Edgington, 1987). The exact moments of the null distribution can also be derived along the same lines as equations (11) – (17), but would be very cumbersome.

In summary, the assessment of co-termination is an important issue for content analysts. Coefficients of co-termination under various circumstances were considered, including discrete and continuous content, binary stream data and length data, $v = 1$ and $v = 2$, and the special case involving cumulative lengths. Generally speaking, for discrete data, multiple-coder κ can be used directly, but results in underestimate of agreement. Choosing $v = 2$ over $v = 1$ increases agreement. The coefficients of the continuous type can handle unequal number of units, especially the coefficient that makes use of cumulative lengths. It is also the least conservative among all indices. Depending on the nature of the study, researchers now possess a family of intercoder agreement indices for unitizing textual data, based on the Multi-response Randomized Blocks Permutation procedure.

References

Aptech Systems, Inc. (1997). GAUSS (version 3.2.30) [computer program]. Maple Valley, WA: Author.

Berelson, B. (1952). *Content analysis in communication research.* Glencoe, IL: Free Press.

Berry, K. J., & Mielke, P. W. (1983a). Computation of finite population parameters and approximate probability values for multi-response permutation procedures (MRPP). *Communications in Statistics – Simulation and Computation, 12*, 83-107.

Berry, K. J., & Mielke, P. W. (1983b). Moment approximations as an alternative to the *F* test in analysis of variance. *British Journal of Mathematical and Statistical Psychology, 36*, 202-206.

Berry, K. J., & Mielke, P. W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48*, 921-933.

Berry, K. J., & Mielke, P. W. (1992). A family of multivariate measures of association for nominal independent variables. *Educational and Psychological Measurement, 52*, 41-55.

Berry, K. J., & Mielke, P. W. (1997a). Agreement measure comparisons between two independent sets of raters. *Educational and Psychological Measurement, 57*, 360-364.

Berry, K. J., & Mielke, P. W. (1997b). Measuring the joint agreement between multiple raters and a standard. *Educational and Psychological Measurement, 57*, 527-530.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Conger, A. J. (1985). Kappa reliabilities for continuous behaviors and events. *Educational and Psychological Measurement, 45*, 861-868.

Craig, R. T. (1981). Generalization of Scott's index of intercoder agreement. *Public Opinion Quarterly, 45*, 260-264.

Edgington, E. S. (1987). *Randomization tests* (2nd ed.). New York, NY: Marcel Dekker.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613-619.

Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin, 86*, 974-977.

Fowler, G. L. (1986). Content and teacher characteristics for master's level research course. *Journalism Quarterly, 63*, 594-599.

Frey, L. R., Botan, C. H., & Kreps, G. L. (2000). *Investigating communication: An introduction to research methods* (2nd ed.). Boston: Allyn & Bacon.

Goodman, L. A., & Kruskal, W. H. (1956). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732-764.

Guetzkow, H. (1950). Unitizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology, 6*, 47-58.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 84*, 289-297.

Iyer, H. K., Berry, K. J., & Mielke, P. W. (1983). Computation of finite population parameters and approximate probability values for multi-response randomized block permutation procedures (MRBP). *Communications in Statistics – Simulation and Computation, 12*, 479-499.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61*, 277-289.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability of data. *Sociological Methodology, 2*, 139-150.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Newbury Park, CA: Sage.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587-604.

Mielke, P. W. (1979). On asymptotic non-normality of null distributions of MRPP statistics. *Communications in Statistics – Theory and Methods, A8*, 1541-1550.

Mielke, P. W. (1984). Meteorological applications of permutation techniques based on distance functions. In P. R. Krishnaiah and P. K. Sen (Eds.), *Handbook of statistics, volume 4* (pp. 813-830). Amsterdam: North-Holland.

Mielke, P. W., & Berry, K. J. (1982). An extended class of permutation techniques for matched pairs. *Communications in Statistics – Theory and Methods, 11*, 1197-1207.

Mielke, P. W., & Berry, K. J. (1994). Permutation tests for common locations among samples with unequal variances. *Journal of Educational and Behavioral Statistics, 19*, 217-236.

Mielke, P. W., & Berry, K. J. (1999). Multivariate tests for correlated data in completely randomized designs. *Journal of Educational and Behavioral Statistics, 24*, 109-131.

Mielke, P. W., & Berry, K. J. (2001). *Permutation methods: A distance function approach.* New York, NY: Springer-Verlag.

Mielke, P. W., & Iyer, H. K. (1982). Permutation techniques for analyzing multi-response data from randomized block experiments. *Communications in Statistics – Theory and Methods, 11*, 1427-1437.

Mielke, P. W., & Sen, P. K. (1981). On asymptotic non-normal null distributions for locally most powerful rank test statistics. *Communications in Statistics – Theory and Methods, A10*, 1079-1094.

Mielke, P. W., & Yao, Y. C. (1990). On g-sample empirical coverage tests: Exact and simulated null distributions of test statistics with small and moderate sample sizes. *Journal of Statistical Computation and Simulation, 35*, 31-39.

Mielke, P. W., Berry, K. J., & Johnson, E. S. (1976). Multi-response permutation procedures for *a priori* classifications. *Communications in Statistics – Theory and Methods, A5*, 1409-1424.

Mielke, P. W., Berry, K. J., Brockwell, P. J., & Williams, J. S. (1981). A class of nonparametric tests based on multiresponse permutation procedures. *Biometrika, 68*, 720-724.

Moffett, E. A., & Dominick, J. R. (1987). Statistical analysis in the JOB 1970-85: An update. *Feedback, 28*, 13-16.

O'Reilly, F. J., & Mielke, P. W. (1980). Asymptotic normality of MRPP statistics from invariance principles of *U*-statistics. *Communications in Statistics – Theory and Methods, A9*, 629-637.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*, 258-284.

Riffe, D., & Freitag, A. A. (1998). A content analysis of content analysis: Twenty-five years of Journalism Quarterly. *Journalism and Mass Communication Quarterly, 74*, 873-882.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Stewart, T. D. (2002). *Principles of research in communication*. Boston: Allyn & Bacon.

Tucker, D. F., Mielke, P. W., & Reiter, E. R. (1989). The verification of numerical methods with multivariate randomized block permutation procedures. *Meteorology and Atmospheric Physics, 40*, 181-188.

Wilhoit, F. G. (1984). Student research productivity: Analysis of Journalism Abstracts. *Journalism Quarterly, 61*, 655-661.

Wimmer, R. D., & Dominick, J. R. (1994). *Mass media research: An introduction* (4th ed.). Belmont, CA: Wadsworth.

Zimmerman, G. M., Goetz, H., & Mielke, P. W. (1985). Use of an improved statistical method for group comparisons to study effects of prairie fire. *Ecology, 66*, 606-611.

Appendix I

Permutation tests represent the “ideal” situations where one can derive the exact probabilities rather than approximate values obtained from common probability distributions, such as the *t*, *F* and χ^2 (Mielke & Berry, 2001, p. 1). Carrying out randomization or permutation of the collected data rather than relying on the often times unreasonable assumption of random sampling or normality not only make a test more data-dependent, but also enhances the practicability of a test, as the practitioners have full control over the stochastic component of the statistical model.

Subsequent development of this procedure includes: (1) the asymptotic behavior of the null distributions of the MRPP statistics (Mielke, 1979; O’Reilly & Mielke, 1980; Mielke & Sen, 1981), (2) specifications of the MRPP model when analyzing univariate rank data (Mielke et al., 1981), (3) appropriate techniques for analyzing Multi-Response data from Randomized Blocks layout (MRBP) (Mielke & Iyer, 1982), (4) a class of techniques for matched-pairs *t* test (Mielke & Berry, 1982), (5) computational procedures for finite population parameters of MRPP and MRBP (Berry & Mielke, 1983a; Iyer, Berry, & Mielke, 1983), (6) moment approximations as an alternative to the *F* test for detection of location and scale shifts under variance heterogeneity (Berry & Mielke, 1983b; Mielke & Berry, 1994), (7) generalization of Cohen’s *κ* to multiple coders and higher levels of measurement (Berry & Mielke, 1988), (8) goodness-of-fit empirical coverage tests (Mielke & Yao, 1990), (9) multivariate measures of association for nominal and higher levels of variables (Berry & Mielke, 1992), (10) intercoder agreement measure comparisons between two independent sets of coders (Berry & Mielke, 1997a), (11) measuring the agreement between coders and a standard (Berry & Mielke, 1997b), (12) multivariate tests for correlated dependent variables in randomized experiments (Mielke & Berry, 1999). This method has been successfully applied in the fields of meteorology and ecology (Mielke, 1984;

Tucker, Mielke, & Reiter, 1989; Zimmerman, Goetz, & Mielke, 1985). In addition, least sum of absolute deviations regression, coupled with MRPP yields a versatile and robust linear model that is capable of handling data from complex factorial experimental designs with covariates (Mielke & Berry, 2001).

Assuming an r-dimensional k group design with the combined sample size equaling N, and group sizes equaling n_i where $i = 1, \ldots, k$, and $\sum_{i=1}^{k} n_i = N$, let $(x_{1I}, \ldots, x_{rI})$ denote the r-dimensional responses where $I = 1, \ldots, N$, and let S_i, where $i = 1, \ldots, k$ denote the k groups of responses, or using the terms of Mielke and Berry (2001), the "exhaustive partitioning" of the N responses into k disjoint sets (p. 12). The basic formulation of the MRPP family of statistics involves the definition of a symmetric distance function of the form

$$\Delta_{I,J} = \left[\sum_{c=1}^{r} (x_{cI} - x_{cJ})^2 \right]^{v/2},$$

as a measure of the multivariate distance between the two observations x_I and x_J. For notational simplicity both the "excess group" and the truncation of distance to a preset maximum value shall not be discussed in the present paper (for details see Mielke & Berry, 2001). The choice of v is arbitrary, but the two choices $v = 1$ and $v = 2$ seems most reasonable. When $v = 1$, the distance is metric Euclidean distance and this distance function has nice theoretical properties of being robust and much less influenced by outliers (Mielke & Berry, 2001). When $v = 2$, the distance is defined in a non-metric squared Euclidean space because the triangle inequality fails in this analysis space, and it is known through both theoretical and simulative studies that this choice leads to a less robust test (Mielke & Berry, 1994). However, the choice of $v = 2$ yields an easier explanation of the test results, because many popular tests essentially involve the use of squared distance.

The MRPP statistic can be thought of as a weighted average of within-group distances. Intuitively, a smaller value of the MRPP statistic would mean higher concentration within each *a priori* classified group (Mielke, 1984, p. 815). Such an interpretation is also in line with the geometric interpretation of the conventional multivariate analysis of variance (see Edgington, 1987, pp. 190). Therefore, in terms of detecting between group differences, a smaller value of the MRPP statistic is necessarily "better."

The MRPP statistic is given by

$$\delta_{obs} = \sum_{i=1}^{k} C_i \xi_i ,$$

where C_i is the group weight for $i = 1, \ldots, k$, and $\sum_{i=1}^{k} C_i = 1$, and

$$\xi_i = \binom{n_i}{2}^{-1} \sum \Delta_{I,J} \psi(x_I)\, \psi(x_J) ,$$

is the average within-group distance for all distinct pairs of responses in the ith group. $\psi(\cdot)$ is an indicator function given by

$$\psi(x_I) = \begin{cases} 1 & \text{if } x_I \in S_i, \\ 0 & \text{if } x_I \notin S_i. \end{cases}$$

The choice of group weights is extensively discussed in Mielke (1984), but $C_i = n_i / N$, and $C_i = (n_i - 1) / (N - k)$ are two sensible choices for $v = 1$ and $v = 2$, respectively.

The formal test of significance of δ_{obs} is carried out by assuming the null hypothesis of equal probabilities being placed upon each one of the

$$M = \frac{N!}{\prod_{i=1}^{k} n_i} ,$$

possible permutations of the *N* responses into the *k* groups, each permutation yielding a realized value of δ. The probability value associated with δ_{obs} is a ratio of the number of δ s being smaller than or equal to δ_{obs} and *M*, formally written as $P\ (\delta_{obs}) = \{\#\ \delta \leq \delta_{obs}\}\ /\ M$.

Because *M* is usually a very large number even for relatively small sample sizes, the exact reference distribution of the MRPP statistic is difficult to obtain, therefore, Mielke, Berry and Johnson (1976) have provided efficient computational methods for the first three cumulants of the MRPP null distribution, upon which a moment approximation using Pearson type III distribution may be utilized. Generally this approximation is excellent. For details please refer to Mielke & Berry (2001).

Appendix II

The following is a GAUSS procedure that implements the missing-cell situation of continuous reliability data. The input is a *b*-coder by *g*-unit matrix similar to Table 4. The output is a 3-by-1 vector, call it *resultv*, with the first element being the observed disagreement, the second element being the expected disagreement (un-averaged), and the third element being the denominator of expected disagreement. So agreement is simply 1 – *resultv*[1]/(*resultv*[2]/ *resultv*[3]).

```
proc discrete(x,v);
  local M, delta, distance, temp, x1, x2, n1, n2, distance1, factor1, factor2, resultm;
  M = prodc(sumc((x .gt 0)')!);
  delta = 0;
  distance = 0;
  for idxI (1, rows(x)-1, 1);
    for idxJ (idxI+1, rows(x), 1);
      x1 = x[idxI,.]';
      x2 = x[idxJ,.]';
      n1 = sumc(x1 .gt 0);
      n2 = sumc(x2 .gt 0);
      if (n1 lt n2);
        temp = n1;
        n1 = n2;
        n2 = temp;
        temp = x1;
        x1 = x2;
        x2 = temp;
      endif;
      x1 = x1[1:n1];
      x2 = x2[1:n1];
      delta = delta + sumc(abs(x1-x2)^v);
      distance1 = 0;
      for i (1, n1, 1);
        for j (1, n2, 1);
          distance1 = distance1+abs(x1[i] - x2[j])^v;
        endfor;
      endfor;
      factor1 = M*(1/n1);
      factor2 = M*(1-n2/n1);
      distance = distance + distance1*factor1+sumc(x1^v)*factor2;
    endfor;
  endfor;
  resultm = zeros(3,1);
  resultm[1] = delta/(rows(x)!/(rows(x)-2)!/2);
  resultm[2] = distance/(rows(x)!/(rows(x)-2)!/2);
  resultm[3] = M;
  retp(resultm);
endp;
```

Appendix III

The following is a GAUSS procedure that implements the method based on cumulative lengths. The input is a *b*-coder by *g*-unit matrix similar to Table 5. The output is a 3-by-1 vector, call it *resultv*, with the first element being the observed disagreement, the second element being the expected disagreement (un-averaged), and the third element being the denominator of expected disagreement. So agreement is simply 1 – *resultv*[1]/(*resultv*[2]/ *resultv*[3]).

```
proc continuous(x);
  local delta, distance, M, temp, x1, x2, n1, n2, distm, tempdistm, part2,
        indexm, minor, minorv, factor, resultm;
  delta = 0; distance = 0;
  M = prodc(sumc((x .gt 0)')!);
  for idxI (1, rows(x)-1, 1);
    for idxJ (idxI+1, rows(x), 1);
      x1 = x[idxI,.]';
      x2 = x[idxJ,.]';
      n1 = sumc(x1 .gt 0);
      n2 = sumc(x2 .gt 0);
      if (n1 lt n2);
        temp = n1;
        n1 = n2;
        n2 = temp;
        temp = x1;
        x1 = x2;
        x2 = temp;
      endif;
      x1 = x1[1:n1];
      x2 = x2[1:n1];
      for i (1, n1, 1);
        delta = delta + (sumc(x1[1:i])-sumc(x2[1:i]))^2;
      endfor;
      distm = zeros(n1, n1);
      for i (1, n1, 1);
        distm[.,i] = x1[.,1] - x2[i,1];
      endfor;
      for j (1, n2, 1);
        distance = distance+sumc(vec(distm[.,1:n2]^2))*(1/n1/n2)*M*(n1+1-j);
      endfor;
      if (n2 ge 2);
        for i (1, n2-1, 1);
          for j (i+1, n2, 1);
            tempdistm = distm[.,1:n2];
            for col (1, n2-1, 1);
              for row (1, n1, 1);
                indexm = zeros(n1,n2);
                indexm[row,.] = ones(1,n2);
                minor = delif(tempdistm,indexm);
                minorv = tempdistm[row,col].*vec(minor[.,col+1:n2]);
                factor = 4*(1/n1)*(1/n2)*(((n2-1)*(n1-1))^(-1))*M*(n1+1-j);
                distance = distance + factor*sumc(minorv);
              endfor;
            endfor;
          endfor;
        endfor;
      endif;
      if ((n1 - n2) ge 1);
        for j (n2+1, n1, 1);
          distance = distance + sumc(distm[.,j]^2)*(1/n1)*M*(n1+1-j);
        endfor;
        for i (1, n2, 1);
          for j (n2+1, n1, 1);
            tempdistm = distm[.,1:n2];
            part2 = distm[.,j];
            for col (1, n2, 1);
              for row (1, n1, 1);
                indexm = zeros(n1,1);
                indexm[row] = 1;
                minor = delif(part2,indexm);
                minorv = tempdistm[row, col].*minor;
                factor = 2*(1/n1)*(1/n2)*((n1-1)^(-1))*M*(n1+1-j);
                distance = distance + factor*sumc(minorv);
              endfor;
            endfor;
          endfor;
        endfor;
        for i (n2+1, n1-1, 1);
          for j (i+1, n1, 1);
            tempdistm = distm[.,i];
            part2 = distm[.,j];
            for row (1, n1, 1);
              indexm = zeros(n1,1);
              indexm[row] = 1;
              minor = delif(part2,indexm);
              minorv = tempdistm[row].*minor;
              factor = 2*(1/n1)*((n1-1)^(-1))*M*(n1+1-j);
              distance = distance + factor*sumc(minorv);
            endfor;
          endfor;
        endfor;
      endif;
    endfor;
  endfor;
  resultm = zeros(3,1);
  resultm[1] = delta/(rows(x)!/(rows(x)-2)!/2);
  resultm[2] = distance/(rows(x)!/(rows(x)-2)!/2);
  resultm[3] = M;
  retp(resultm);
endp;
```