THE ASSESSMENT OF INTERRATER AGREEMENT
FOR MULTIPLE ATTRIBUTE RESPONSES

by

Lawrence L. Kupper and Kerry B. Hafner

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1843

January 1988

THE ASSESSMENT OF INTERRATER AGREEMENT FOR MULTIPLE ATTRIBUTE RESPONSES

Lawrence L. Kupper and Kerry B. Hafner
Department of Biostatistics, School of Public Health,
University of North Carolina at Chapel Hill,
Chapel Hill, North Carolina 27599-7400, U.S.A.

SUMMARY

New methods are developed for assessing the extent of interrater agreement when each unit to be rated is characterized by *one or more* nominal attributes. For such multiple attribute response data, a two-rater concordance statistic is derived, and associated statistical inference-making procedures are provided. This concordance statistic is corrected for chance agreement based on an underlying hypergeometric model. Numerical examples are given to illustrate the proposed methodology, and extensions to situations involving more than two raters are briefly considered.

1. Introduction

In a variety of scientific disciplines, studies are conducted in which two or more "raters" independently examine a group of units (*e.g.*, human or animal subjects, verbatim transcripts, radiographs, etc.) to determine whether prespecified attributes are present or absent in each of the units. The usual goal of such studies is to evaluate how well the raters can discriminate among the attributes under investigation. Since the raters will invariably make some incorrect assessments, quantification of rater performance becomes an important statistical concern. When the set of attributes possessed by each unit is known without error, so that validity rather than agreement is the issue, then sensitivity and specificity analyses can be used to help achieve such quantification (Kleinbaum, Kupper, and Morgenstern, Chapter 12, 1982; Maclure and Willett, 1987). However, in some situations it may not be possible to make such error-free determinations, and hence measurement of the extent of agreement among the different raters is of primary interest.

Key Words: Chance-corrected agreement statistics; Non-central hypergeometric distribution; Kappa statistic; Conditional distribution theory; Nominal attributes.

Numerous measures of "interrater agreement" have been used to quantify the degree of concordance among different raters. Reviews of the pertinent literature have been provided by Fleiss (1975, 1981) and by Bartko and Carpenter (1976). When determination regarding the presence or absence of *exactly one* nominal attribute is sufficient to describe each unit, the putative measure of interrater agreement is the kappa statistic. The kappa statistic was proposed by Cohen (1960) as an appropriate agreement index because it is *adjusted* for agreement due purely to chance; methodology based on its use is still being developed (Gross, 1986; Jannarone, Macera, and Garrison, 1987). However, alternative measures of interrater agreement may be more appropriate than kappa in certain instances (see Maclure and Willett, 1987).

The problem to be addressed in this paper is that of quantifying interrater agreement when each unit is to be characterized by *one or more* nominal attributes. That is, given a set of $k \geq 2$ nominal attributes from which to choose, each of two or more raters independently assigns a subset of these attributes to each of the n units under investigation. To apply the methods proposed in this paper, these k attributes must be distinct from one another. Furthermore, for each of the n units, each rater must select *at least one* attribute from the specified set. In other words, it is assumed that each unit possesses at least one of the k attributes under consideration (*i.e.*, the set of attributes is "exhaustive").

With one exception, we also require that the set of attributes be defined in such a way that the selection of one attribute does not preclude the possible selection of any other. That exception pertains to the situation in which each unit is to be characterized by exactly one of the k prespecified traits. As an example, consider a study where two doctors (*i.e.*, raters) independently evaluate each of n subjects, the purpose of the study being to assess how well the doctors' diagnoses agree. Suppose that each doctor can only select *one* of $(k-1)$ specified diseases, or a k-th diagnosis of "disease-free", to characterize the i-th subject, $i=1, 2, \ldots, n$. Then, since only one of the k distinct diagnoses is being selected by each doctor, no logical contradiction can occur. The application of our methodology to these important special cases will be discussed in detail later in this paper.

The usual method of analysis for multiple attribute responses involves computing separate agreement indices (*e.g.*, kappa statistics) for each individual attribute (*e.g.*, see Armitage and Blendis, 1966). However, even though these separate agreement measures are of interest, it is not clear how they relate to the assessment of "overall agreement" since they ignore the underlying multivariate nature of the data. An average (or weighted average) of the separate agreement indices could be used as a summary index of agreement, but such a conglomerate score would be hard to interpret if the magnitudes of the individual indices vary considerably. In addition, the standard error of such an average would be extremely difficult to compute because the indices involved are statistically dependent. For the situation in which the attributes are scored on an ordinal scale, a repeated measures ANOVA is applicable if the appropriate assumptions are met (see Maxwell and Pillner, 1968).

2. Statistical Methods for Two Raters

2.1 *Estimation of Degree of Concordance*

Consider a study in which two equally trained raters, say rater A and rater B, independently examine each of n units. Following examination of the i-th unit, $i=1, 2,\ldots, n$, each rater must decide which subset, from a prespecified set of $k \geq 2$ nominal attributes, best describes the i-th unit. The objective of the study is to quantify the magnitude of the agreement between raters A and B, and subsequently to carry out appropriate statistical inferences.

Let $\mathcal{X}$ denote a prespecified set containing $k \geq 2$ nominal attributes, so that the cardinality of this set is $\mathrm{Card}(\mathcal{X}) = k$. Also, let the random variables A_i and B_i denote the numbers of attributes for the i-th unit chosen by raters A and B, respectively, where $1 \leq A_i \leq k$ and $1 \leq B_i \leq k$. Recall that the prespecified set $\mathcal{X}$ is assumed to be "exhaustive", so that each rater must necessarily select at least one attribute from $\mathcal{X}$ to characterize any particular unit.

Conditional on $A_i = a_i$ and $B_i = b_i$, consider the subsets $\mathcal{A}_i$ and $\mathcal{B}_i$, where $\mathcal{A}_i$ is the subset of attributes chosen by rater A to describe the i-th unit, and where $\mathcal{B}_i$ is defined analogously for rater B; note that $\mathrm{Card}(\mathcal{A}_i) = a_i$ and $\mathrm{Card}(\mathcal{B}_i) = b_i$. Then, let the random variable X_i denote the number of elements common to the sets $\mathcal{A}_i$ and $\mathcal{B}_i$, namely,

$$X_i = \mathrm{Card}(\mathcal{A}_i \cap \mathcal{B}_i)\ ,\ \ i=1, 2,\ldots, n.$$

Note that larger values of X_i indicate more "overlap" between the sets $\mathcal{A}_i$ and $\mathcal{B}_i$ and hence are reflective of more agreement between the two raters. In the extreme case of perfect agreement, $\mathcal{A}_i = \mathcal{B}_i$ and so $X_i = \mathrm{Card}(\mathcal{A}_i) = \mathrm{Card}(\mathcal{B}_i)$.

We define the "observed proportion of concordance" between rater A and rater B for the i-th unit to be

$$\hat{\pi}_i = X_i / \max(a_i, b_i)\ ,\ \ i=1, 2,\ldots, n. \tag{1}$$

As a simple example, let $\mathcal{X} = \{P, Q, R, S\}$ so that $\mathrm{Card}(\mathcal{X}) = 4$. Suppose that rater A chooses $\mathcal{A}_i = \{P\}$ and rater B chooses $\mathcal{B}_i = \{P, R\}$ to describe the i-th unit. Thus, the two raters agree on one attribute and $\hat{\pi}_i = 1/2$. With the possible exceptions in the special case $a_i = b_i = 1$ mentioned at the end of Section 1, $\hat{\pi}_i$ does not incorporate agreement on attributes that were *not* selected by either rater (attributes Q and S in our simple example). This feature of (1) is especially attractive when the set $\mathcal{X}$ contains several rarely occurring attributes.

The *overall* observed concordance between raters A and B can be computed as the average of the $\hat{\pi}_i$'s, namely,

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_i\ . \tag{2}$$

However, since some of this observed agreement will be due purely to chance, the quantity (2) should be adjusted in an appropriate manner. To do so, it is necessary to find the conditional distribution of X_i given that $A_i = a_i$ and $B_i = b_i$. If both raters choose their subsets of attributes for the i-th unit completely at random, then the null distribution of X_i is *central* hypergeometric, namely,

$$\mathrm{pr}_o(X_i = x_i) = \frac{\binom{a_i}{x_i}\binom{k-a_i}{b_i - x_i}}{\binom{k}{b_i}}\ ,\ i=1, 2,\ldots, n, \tag{3}$$

in which $\max(0, a_i + b_i - k) \leq x_i \leq \min(a_i, b_i)$.

Expressions for the null expected value and null variance of X_i based on (3) are well known. Using (1), it then follows that

$$E_o(\hat{\pi}_i) = \frac{\min(a_i, b_i)}{k},$$

and

$$V_o(\hat{\pi}_i) = \frac{(k-a_i)(k-b_i)\min(a_i, b_i)}{k^2(k-1)\max(a_i, b_i)}.$$

Thus, an overall measure of chance-corrected concordance between rater A and rater B is

$$\hat{C}_{AB} = \frac{\hat{\pi} - \pi_o}{1 - \pi_o}, \tag{4}$$

where $\pi_o = \frac{1}{nk}\sum_{i=1}^{n} \min(a_i, b_i)$. The above expression and the formulae to follow can easily be modified to permit k to vary with i, although we will continue to assume that $k_i = k$ for $i=1, 2,..., n$. Note that $\hat{C}_{AB}$ in (4) follows the general form of the chance-corrected agreement measures discussed by Fleiss (1975, p. 655).

Under the null hypothesis that both raters choose their attributes totally at random, $E_o(\hat{C}_{AB})=0$. The maximum value of $\hat{C}_{AB}$ is attained when there is perfect concordance between the two raters (*i.e.*, $\hat{\pi}_i=1$, $i=1, 2,..., n$), in which case $\hat{C}_{AB}=1$. The minimum value of $\hat{C}_{AB}$ is achieved when $\hat{\pi}_i=0$ for $i=1, 2,..., n$. When this occurs, $\min(\hat{C}_{AB}) = -(k-1)/(k+1)$ if k is an odd integer, and $\min(\hat{C}_{AB}) = -1$ if k is an even integer.

As mentioned earlier, an interesting special case of (4) results when each rater need only select one attribute from the set $\mathcal{X}$ to describe the i-th unit. For this situation, $a_i = b_i = 1$ for all i, and expression (4) simplifies to

$$\hat{C}_{AB} = \frac{(k\hat{\pi} - 1)}{(k-1)}, \tag{5}$$

in which $\hat{\pi}$ (or p_o in the notation of Fleiss, 1981, p. 213) is the *observed* proportion of agreement. When $k=2$ (*e.g.*, the presence or absence of a single specified trait), the chance-corrected measure of concordance given by (5) equals $(2\hat{\pi} - 1)$. This agreement index was initially proposed by Holley and Gillford (1964), and later by Maxwell (1977). Interestingly, in their developments they do not recognize that $(2\hat{\pi}-1)$ is, in fact, a chance-corrected statistic. Furthermore, it can be shown that (5) is always at least as large in value as the kappa statistic; the two measures are equal when each rater chooses each attribute an equal number of times (*i.e.*, for data in the format of a (k x k) frequency table, the marginal proportions are all 1/k).

(2.2) *Inference methods*

In practice, agreement indices are sometimes used only as descriptive statistics. However, it is often desirable to test hypotheses, and, more importantly, to construct confidence intervals. If the number of units n is moderately large, then an appropriate statistic to test H_0: $C_{AB}=0$ versus the alternative H_1: $C_{AB}>0$, where $E(\hat{C}_{AB})=C_{AB}$, is

$$Z=\frac{\hat{C}_{AB}}{\sqrt{V_o(\hat{C}_{AB})}}\ ;$$

under H_0, the asymptotic distribution of Z is standard normal, and the null variance of $\hat{C}_{AB}$ is

$$V_o(\hat{C}_{AB}) = \frac{\sum_{i=1}^{n}(k-a_i)(k-b_i)\min(a_i,\, b_i)/\max(a_i,\, b_i)}{(k-1)[n k-\sum_{i=1}^{n}\min(a_i,\, b_i)]^2}\ .$$

A test of H_0: $C_{AB}=0$ versus H_1: $C_{AB}>0$ assesses whether the observed agreement between raters A and B, as measured by $\hat{C}_{AB}$, is significantly greater than that due to chance alone.

To develop a $100(1-\alpha)\%$ confidence interval for C_{AB}, it is necessary to find an estimate for the non-null variance $V(\hat{C}_{AB})$ of $\hat{C}_{AB}$. Given that $A_i=a_i$ and $B_i=b_i$, it can be shown (see Cox, 1970; Kleinbaum, Kupper, and Morgenstern, 1982) that the non-null distribution of X_i is *non-central* hypergeometric, namely,

$$\mathrm{pr}(X_i=x_i;\ \psi_i)=\frac{\binom{a_i}{x_i}\binom{k-a_i}{b_i-x_i}\psi_i^{x_i}}{\sum_{u_i}\binom{a_i}{u_i}\binom{k-a_i}{b_i-u_i}\psi_i^{u_i}}\ . \tag{6}$$

In expression (6), ψ_i is the non-centrality parameter and the summation in the denominator is over all integer values of u_i such that $\max(0, a_i+b_i-k) \le u_i \le \min(a_i, b_i)$. Note that (6) reduces to (3) when $\psi_i=1$.

Unfortunately, a series of complex iterative calculations is necessary to estimate the non-null variance of X_i, and subsequently $V(\hat{C}_{AB})$. However, if one assumes that $\psi_i=\psi$ for $i=1, 2,\ldots, n$, then the number of iterations is considerably reduced. Given this homogeneous non-centrality parameter assumption, the estimated variance of X_i is obtained from (6) by evaluating $V(X_i;\, \psi)$ at $\psi=\hat{\psi}$, where $\hat{\psi}$ is the iterative solution to the equation

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} E(X_i;\, \psi) \;, \tag{7}$$

and where $E(X_i;\, \psi)$ and $V(X_i;\, \psi)$ are the non-null mean and variance of X_i based on (6).

A good approximation (Breslow, 1981) to $\hat{\psi}$ obtained from (7) is the Mantel-Haenszel estimate

$$\hat{\psi}_{mh} = \frac{\sum_{i=1}^{n} x_i\,(k-a_i-b_i+x_i)}{\sum_{i=1}^{n} (a_i-x_i)\,(b_i-x_i)} \,.$$

In practice, then, an approximate large-sample $100(1-\alpha)\%$ confidence interval for C_{AB} can be computed as

$$\hat{C}_{AB} \pm Z_{1-\alpha/2} \sqrt{\hat{V}(\hat{C}_{AB})} \;, \tag{8}$$

where

$$\hat{V}(\hat{C}_{AB}) = \frac{\sum_{i=1}^{n} V(X_i;\, \hat{\psi}_{mh}) / [\max(a_i, b_i)]^2}{[n - \frac{1}{k}\sum_{i=1}^{n} \min(a_i, b_i)]^2} \,. \tag{9}$$

As an important special case, consider again the situation where $a_i=b_i=1$ for all i, so that $\hat{C}_{AB} = (k\hat{\pi}-1)/(k-1)$. In this case, it can be shown that $\hat{\psi}=\hat{\psi}_{mh}=(k-1)\hat{\pi}/(1-\hat{\pi})$, and that a large-sample $100(1-\alpha)\%$ confidence interval for C_{AB} based on (8) and (9) is

$$\frac{(k\hat{\pi}-1)}{(k-1)} \pm Z_{1-\alpha/2}\left(\frac{k}{k-1}\right)\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\,. \tag{10}$$

The binomial-like variance term $\hat{\pi}(1-\hat{\pi})/n$ results when $a_i=b_i=1$ for all i because $\hat{\pi}_i=X_i$ is, from (6), a Bernoulli random variable with $pr(X_i=1)=\psi/(k-1+\psi)$.

3. Numerical Examples

3.1 *Example* 1

The hypothetical data displayed in Table 13.1 of Fleiss (1981) are considered in this example. Briefly, the data consist of n=100 subjects (units) that were evaluated by two raters, say A and B. Each subject was independently diagnosed by each rater as *either* psychotic (P), neurotic (N), or organic (O). In our notation, $\mathfrak{X}=\{P, N, O\}$ so that k=3. The goal of the study was to assess the extent of diagnostic agreement between the two raters.

Since each rater was required to assign exactly one of the k=3 attributes to each subject, estimation and inference methods for the special case when $a_i=b_i=1$, $i=1, 2, \ldots, n$, are applicable. In particular, the estimated concordance between rater A and rater B is found from (5) to be

$$\hat{C}_{AB}=\frac{(3\hat{\pi}-1)}{2},$$

in which $\hat{\pi}$ is the observed proportion of agreement between the raters. For these data, $\hat{\pi}=0.89$ and hence $\hat{C}_{AB}=0.835$; in contrast, the kappa value is 0.68 (see Fleiss, 1981, pg. 219). To assess whether the observed concordance given by $\hat{C}_{AB}$ is greater than that expected by chance alone (*i.e.*, to test H_0: $C_{AB}=0$ versus H_1: $C_{AB}>0$), the appropriate test statistic is, from Section (2.2),

$$Z=\frac{(3\hat{\pi}-1)/2}{\sqrt{1/n(k-1)}}\,.$$

For these data, Z=11.81, which is highly significant. The corresponding kappa-based Z value is 8.95 (see Fleiss, 1981, pg. 220). An approximate 95% confidence interval for C_{AB} is found to be (0.743, 0.927) by use of equation (10); the approximate 95% confidence interval for kappa (see Fleiss, 1981, pgs. 221-222) is (0.505, 0.847).

3.2 *Example* 2

In this example, dental X-ray data are used to illustrate the general concordance methods proposed in this paper for the analysis of multiple attribute response data (see Kantor, Reiskin, and Lurie, 1985, for details of the study). For our analysis, a sample of $n=44$ dental X-ray films was selected from the complete data set. The purpose of this study was to compare two different types of X-ray film: "U" denotes the standard film, and "E" denotes a faster speed film that reduces patient X-ray exposure by approximately fifty percent. Use of the "E" film would be strongly encouraged by the dental community if the X-ray quality was at least as good as that provided by the "U" film.

To assess the comparability of the two films, each of three raters (A, B, and C) independently evaluated the $n=44$ X-ray films (21 of type "U" and 23 of type "E") to determine which of the $k=14$ teeth on each X-ray had carious lesions (cavities). The data are displayed in Appendix 2 and show the film speed, the tooth reference numbers, and the particular raters that diagnosed cavities for each film-by-tooth combination. Teeth 1-7 refer to upper teeth and teeth 8-14 refer to lower teeth; also, the tooth reference numbers increase from the front of the mouth to the back. As one would expect, the front teeth are essentially cavity-free, whereas the premolars (teeth 4, 5, 11, 12) and molars (teeth 6, 7, 13, 14) had many carious lesions.

The methods developed in this paper are used to produce three pairwise concordance statistics ($\hat{C}_{AB}$, $\hat{C}_{AC}$, and $\hat{C}_{BC}$) and a three-rater concordance statistic ($\hat{C}_{ABC}$) for each film speed (see Appendix 1). These concordance statistics, Z-tests of H_0: $C=0$ versus H_1: $C>0$, and approximate 95% confidence intervals for each film speed are shown in Table 1. Note that the concordance statistics for the "U" film are larger than those for the "E" film, and that the estimated non-null variances $\hat{V}(\hat{C}_{AB})$ and $\hat{V}(\hat{C}_{BC})$ for the "E" film are approximately twice as large as those for the "U" film. For these data, all of the concordance statistics are significantly greater than what would be expected by chance alone. Computationally feasible methods for calculating non-null variances and associated confidence intervals using our three-rater concordance statistic are not yet available and hence are not reported in Table 1.

Insert Table 1 here.

An approximate test statistic to assess whether the expected concordances for the two film speed types are equal or not (*i.e.*, to test H_0: $C^U - C^E = 0$ versus H_1: $C^U - C^E \neq 0$) is

$$Z = \frac{\hat{C}^U - \hat{C}^E}{\sqrt{\hat{V}(\hat{C}^U) + \hat{V}(\hat{C}^E)}},$$

in which the superscripts delineate the two film speeds. For large n, this test statistic can be compared to appropriate percentiles of the standard normal distribution to assess its significance. In addition, an approximate $100(1-\alpha)\%$ confidence interval for the true concordance difference $(C^U - C^E)$ can be constructed as

$$(\hat{C}^U - \hat{C}^E) \pm Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{C}^U) + \hat{V}(\hat{C}^E)}\ .$$

The results of this comparative analysis for the two film speed types are displayed in Table 2. The test statistics and approximate 95% confidence intervals clearly suggest that there is no statistically significant difference between the concordance measures for the two film speed types. Thus, if our interrater concordance statistics are reasonable measures of X-ray quality, then the two different speed films can be considered to be equally readable by the three raters.

Insert Table 2 here.

4. Further Methodologic Issues and Discussion

The observed proportion of concordance (1) is clearly an appropriate measure of agreement when the i-th unit is *truly* characterized by no more than $\max(a_i, b_i)$ attributes. One possible alternative to (1) is the ratio $X_i/(a_i+b_i-X_i)$, where $(a_i+b_i-X_i)$ is the number of elements in the set $\mathcal{A}_i \cup \mathcal{B}_i$. Even though much of our methodology can be extended to consider this alternative ratio, we prefer (1) on the grounds of both mathematical tractability and interpretability (*e.g.*, our generalized concordance statistic (4) based on (1) specializes to an agreement statistic previously suggested in the literature).

It may be of interest to test H_0: $C_{AB} = C_o$ versus H_1: $C_{AB} > C_o$, where $C_o > 0$. An appropriate test statistic to use is

$$Z = \frac{\hat{C}_{AB} - C_o}{\sqrt{V_{C_o}(\hat{C}_{AB})}} ,$$

which is approximately $N(0, 1)$ for large n under H_0. Here, $V_{C_o}(\hat{C}_{AB})$ is given by expression (9) with $V(X_i; \psi_o)$ replacing $V(X_i; \hat{\psi}_{mh})$; ψ_o is the iteratively-obtained solution to the equation

$$\sum_{i=1}^{n} \frac{E(X_i; \psi_o)}{\max(a_i, b_i)} = n\,[\pi_o + C_o(1-\pi_o)] ,$$

and $E(X_i; \psi_o)$ and $V(X_i; \psi_o)$ are obtained using (6) with $\psi_i = \psi_o$.

For the important special case in which $a_i = b_i = 1$ for all i, our new chance-corrected concordance statistic (5) and its associated inference-making procedures are very appealing because of their computational simplicity. Furthermore, in some instances they are to be preferred on interpretative grounds to kappa statistic-based methods. For example, consider the following 2x2 table which depicts a situation in which the attribute D rarely occurs in the population of interest.

		Rater A: D	Rater A: $\bar{D}$
Rater B	D	1	2
	$\bar{D}$	2	95

For this table, the overall observed proportion of agreement between the two raters is $\hat{\pi}=0.96$, so that $\hat{C}_{AB}=0.92$ from (5). In contrast, the kappa statistic value is only 0.31, reflecting poor chance-corrected interrater agreement. This large discrepancy between the two measures occurs because the kappa statistic is very sensitive to highly skewed frequency distributions, as in the case of a very rare, or extremely common, attribute (see Bartko and Carpenter, 1976, pg. 309, for a related discussion). Furthermore, if five of the $\{\bar{D}, \bar{D}\}$ agreements had been $\{D, D\}$ agreements, then the kappa statistic would increase in value to 0.73, suggesting good agreement beyond chance. Thus, unlike (5), the kappa statistic is not invariant to the distribution of agreement types $\{D, D\}$ and $\{\bar{D}, \bar{D}\}$.

Although the procedures developed in this paper are applicable in a wide variety of situations, further research is needed to provide even greater generality. For example, to use our methods for multiple attribute response data, each rater is required to choose *at least one* attribute from the set $\mathcal{X}$, and the selection of any one attribute must not preclude the possible selection of any other attribute. Clearly, new methodology is needed to deal with responses for which no attributes are chosen by one or more raters. However, recall that this preclusion restriction is not an issue when each unit is to be characterized by one, and only one, of the k attributes. In that special case, our new concordance statistic (5) is arguably preferable (on both interpretative and computational grounds) to kappa, which is applicable *only* in this limited situation. Lastly, another area that needs further development pertains to the case of three (or more) raters. This paper provides a three-rater concordance statistic and its corresponding null variance; however, no attractive computational method is currently available to produce estimated non-null variances and associated confidence intervals.

Acknowledgements

The authors would like to thank Dr. Mel Kantor, DDS for providing the dental data used in *Example* 2. The authors would also like to thank the Members of the Evaluation Core for the Multipurpose Arthritis Center at the University of North Carolina in Chapel Hill for providing their practical insight into this problem. The second author was partially funded by N.I.E.H.S. training grant #2 T32 ES07018 while a Ph.D. student at the University of North Carolina in Chapel Hill.

Table 1. Concordance estimates and associated inferential statistics for each film speed.

Speed = "U"

Raters	$\hat{C}$	$V_o(\hat{C})$	Z	$\hat{\psi}_{mh}$	$\hat{V}(\hat{C})$	Approximate 95% C.I.
AB	0.619	0.0025	12.4	79.80	0.0023	(0.525, 0.714)
AC	0.582	0.0025	11.6	30.47	0.0030	(0.476, 0.689)
BC	0.548	0.0021	11.9	104.0	0.0012	(0.481, 0.615)
ABC	0.511	0.0004	25.6			

Speed = "E"

Raters	$\hat{C}$	$V_o(\hat{C})$	Z	$\hat{\psi}_{mh}$	$\hat{V}(\hat{C})$	Approximate 95% C.I.
AB	0.530	0.0024	10.8	34.64	0.0042	(0.403, 0.657)
AC	0.501	0.0021	10.9	23.76	0.0033	(0.389, 0.614)
BC	0.534	0.0020	11.9	40.00	0.0028	(0.431, 0.637)
ABC	0.465	0.0003	26.8			

Table 2. Inferences comparing film speed types.

Raters	Z	Approximate 95% C.I.
AB	1.10	(−0.069, 0.247)
AC	1.02	(−0.075, 0.237)
BC	0.22	(−0.110, 0.138)

REFERENCES

Armitage, P., Blendis, L.M., and Smyllie, H.C. (1966). The measurement of observer disagreement in the recording of signs. *Journal of the Royal Statistical Society* (*A*) 129, 98-109.

Bartko, J.J. and Carpenter, Jr. (1976). On the methods and theory of reliability. *The Journal of Nervous and Mental Disease* 163, 307-317.

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika* 68, 73-84.

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

Cox, D.R. (1970). *The Analysis of Binary Data.* London: Methuen.

Fleiss, J.L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics* 31, 651-659.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edition. New York: Wiley.

Gross, S.T. (1986). The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics* 42, 883-893.

Holley, J.W. and Guilford, J.P. (1964). A note on the G index of agreement. *Educational and Psychological Measurement* 24, 749-753.

Jannarone, R.J., Macera, C.A., and Garrison, C.Z. (1987). Evaluating interrater agreement through "case-control" sampling. *Biometrics* 43, 433-437

Kantor, M.L., Reiskin, A.B., and Lurie, A.G. (1985). A clinical comparison of X-ray films for detection of proximal surface caries. *Journal of the American Dental Association* 111, 967-969.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods.* Belmont, California: Lifetime Learning Publications.

Maclure, M. and Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology* 126, 161-169.

Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry* 130, 79-83.

Maxwell, A.E. and Pilliner, A.E. (1968). Deriving coefficients of reliability and agreement for ratings. *The British Journal of Mathematical and Statistical Psychology* 21, 105-116.

APPENDIX 1

In this appendix we consider the problem of assessing interrater agreement for multiple attribute response data when three or more raters individually evaluate each of n units. As in the two-rater case, each rater must select at least one attribute from a prespecified set $\mathcal{K}$ containing $k \geq 2$ attributes to characterize the i-th unit. Assumptions regarding the set $\mathcal{K}$ are the same as discussed in Section 1. Only the formal development for three raters is presented here, because extensions to more than three raters follow in an obvious manner.

Denote the three raters by A, B, and C, and let the random variables A_i, B_i, and C_i refer to the numbers of attributes chosen by the respective raters. Given that $A_i = a_i$, $B_i = b_i$, and $C_i = c_i$, let $\mathcal{A}_i$, $\mathcal{B}_i$, and $\mathcal{C}_i$ refer to the subsets of attributes chosen by the corresponding raters to describe the i-th unit, $i = 1, 2, \ldots, n$. Also, let X_i refer to the number of attributes common to *any* particular pair of these subsets, and let Y_i refer to the number of attributes common to all three subsets, *i.e.*,

$$Y_i = \mathrm{Card}(\mathcal{A}_i \cap \mathcal{B}_i \cap \mathcal{C}_i).$$

The *conditional* null distribution of Y_i, given that $X_i = x_i$ for some particular pair of raters, is again central hypergeometric. The conditional null mean and variance of Y_i can then be unconditioned using (3) to obtain the unconditional null mean and variance of Y_i, namely,

$$E_o(Y_i) = \frac{a_i b_i c_i}{k^2}$$

and

$$V_o(Y_i) = \frac{a_i b_i c_i}{k^4 (k-1)^2} \left[(k-1)(k-c_i)(k^2 - a_i b_i) + k(c_i - 1)(k - a_i)(k - b_i) \right].$$

It is important to note that the null mean and (with some algebra) null variance of Y_i do not depend on which pair of raters is chosen to define X_i.

Analogous to (1), we take $\hat{\pi}_i = Y_i / \max(a_i, b_i, c_i)$, and define an overall measure of concordance among raters A, B, and C to be

$$\hat{C}_{ABC} = \frac{\frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_i - \frac{1}{nk^2} \sum_{i=1}^{n} a_i b_i c_i / \max(a_i, b_i, c_i)}{1 - \frac{1}{nk^2} \sum_{i=1}^{n} a_i b_i c_i / \max(a_i, b_i, c_i)} .$$

Tests of hypotheses for relatively large n can be conducted like those described in Section (2.2). In particular, a test of H_0: $C_{ABC}=0$ versus H_1: $C_{ABC}>0$ can be based on the statistic

$$Z=\frac{\hat{C}_{ABC}}{\sqrt{V_o(\hat{C}_{ABC})}};$$

here, the limiting distribution of Z is standard normal, and the null variance of $\hat{C}_{ABC}$ is

$$V_o(\hat{C}_{ABC})=\frac{\sum_{i=1}^{n} V_o(Y_i)/[\max(a_i, b_i, c_i)]^2}{[n-\frac{1}{k^2}\sum_{i=1}^{n} a_i b_i c_i/\max(a_i, b_i, c_i)]^2}.$$

Finally, it should be noted that it is possible to have high two-way concordance, but low three-way concordance. As a general comment, we recommend that all pairwise and higher-order concordance statistics be reported for any study; each statistic supplies useful information and may provide insight into problem areas that require further investigation.

Appendix 2. Partial data from dental study comparing X-ray film quality between two film speed types, the standard film "U" and the faster speed film "E". The letters in the body of the table refer to specific raters that diagnosed a cavity for the tooth numbers shown.

	Tooth													
Speed	1	2	3	4	5	6	7	8	9	10	11	12	13	14
U						ABC								
U					ABC	ABC	A			C	ABC	ABC	C	ABC
U					ABC	C						C		
U				A	ABC	ABC	ABC				C	C	C	
U						AC	A						ABC	
U					ABC	BC	ABC				AC	A		C
U					A						A	AC		ABC
U				ABC	ABC	AB	C					ABC	ABC	C
U				AC	ABC						ABC	B	AC	
U				ABC	C	A	C							
U				ABC								ABC		
U				ABC	ABC	AC				ABC				
U					BC	ABC							BC	
U				AC	C							A	ABC	
U					BC	ABC								C
U											ABC	C		
U													ABC	
U				ABC	AC	AC	AC							ABC
U										AC			ABC	ABC
U				C		BC					ABC	A		C
U							C				ABC	ABC		C
E			AC	B							A	ABC	AB	B
E			AC	A	ABC								C	
E				ABC										
E			A	C			C					C	BC	
E					ABC									
E				ABC		ABC						ABC	C	C
E					ABC	ABC							AC	
E					ABC								ABC	ABC
E													ABC	
E					ABC						A	A		
E											BC	B	ABC	B
E			C	A	AC	BC							C	
E				C	C	B				ABC	ABC			
E			ABC				C							
E				A	C							C		ABC
E					C						C	ABC	ABC	
E				AB		ABC						ABC	ABC	C
E			ABC	ABC	ABC	C					ABC	ABC	C	B
E											ABC	BC		
E					C						A		B	AC
E				C							A	BC		
E				ABC	ABC	ABC					A	C	C	BC
E				AC	AC	A				A		ABC	ABC	