J Clin Epidemiol Vol. 46, No. 9, pp. 1055–1062, 1993
Printed in Great Britain. 

0895-4356/93 $6.00 + 0.00


# THE ANALYSIS OF ORDINAL AGREEMENT DATA: BEYOND WEIGHTED KAPPA

PATRICK GRAHAM[1] and RODNEY JACKSON[2]

[1]Department of Community Health and General Practice, Christchurch School of Medicine, P.O. Box 4345, Christchurch, New Zealand and [2]Department of Community Health, University of Auckland, Auckland, New Zealand

*(Received 3 December 1992; received for publication 14 April 1993)*

**Abstract**—The weighted kappa statistic has been used as an agreement index for ordinal data. Using data on the comparability of primary and proxy respondent reports of alcohol drinking frequency we show that the value of weighted kappa can be sensitive to the choice of weights. The distinction between association and agreement is clarified and it is shown that in some respects weighted kappa behaves more like a measure of association than an index of agreement. In particular, it is demonstrated that the weighted kappa statistic is not always sensitive to differences in the observed proportion in exact agreement and that high values of weighted kappa can be observed even when the level of agreement is low. We illustrate the use of statistical models in the analysis of epidemiologic agreement data and conclude that modelling ordinal agreement data produces insights which cannot be obtained through the use of weighted kappa statistics.

Kappa Agreement Epidemiologic methods

## INTRODUCTION

Studies of the reliability of epidemiologic survey instruments and of observer reliability usually involve analysis of agreement amongst paired measurements. In clinical research the question of diagnostic agreement has also received considerable attention [1]. In both these situations the study of agreement is the main issue and we follow Becker in describing such studies as agreement studies [2].

The natural representation of categorical agreement data is a two-way table such as Table 1, which is a cross-classification of primary respondent and proxy reports of alcohol drinking frequency. The data reported in Table 1 were drawn from the control series of the Auckland Heart Study, a community based case–control study of coronary heart disease. In this study, a randomly selected sub-sample of the non-fatal myocardial infarction controls (primary respondents) were asked if their next of kin (proxy respondents) could also be interviewed about them (the myocardial infarction controls) [3].

In this paper we focus on methods for analysing agreement on the classification of individuals rather than the issue of agreement between the marginal distributions. A natural measure of the degree of individual level agreement is the probability that any random selection from a set of paired measurements yields a pair who are in exact agreement. In the analysis of categorical agreement data it has become customary to use the kappa statistic [4] which discounts the observed proportion of all pairs in exact agreement by the proportion expected by chance. The proportion of pairs in agreement expected by chance is the proportion expected if the two measurements or reports are, in fact, made independently of one another. The kappa statistic is often referred to as a measure of chance corrected agreement or agreement beyond chance.

When the data to be analysed are measured on an ordered categorical scale, the weighted

kappa statistic has been advocated as one of the preferred methods for the analysis of agreement data [5]. The weighted kappa statistic incorporates the concept of "partial credit" for near agreements with the amount of credit dependent on the magnitude of the discrepancy and the weighting system employed.

Recent examples of the use of weighted kappa include the analysis of the reproducibility and primary respondent–proxy respondent reliability of dietary questionnaires [6, 7], the reproducibility of components of a measure of illness severity [8] and agreement between clinicians in the assessment of asthma severity in children and between methods of asthma severity assessment [9]. However, we demonstrate below that there are serious problems with the use of weighted kappa as an agreement index.

Statistical modelling has been suggested as an alternative to the use of indices in the analysis of ordinal agreement data [2, 10]. We illustrate the use of modelling in the context of epidemiologic agreement data and show that this approach provides insights not easily obtained through the use of the weighted kappa statistic.

The problems with weighted kappa and the application of modelling to agreement data are illustrated using the data shown in Table 1, concerning the comparability of primary and proxy respondent reports of alcohol drinking frequency. An analysis of this dataset, based on the weighted kappa statistic and estimates of category distinguishability [11], has been reported elsewhere [12].

## WEIGHTED KAPPA: WHICH WEIGHTING SYSTEM SHOULD BE USED?

The unweighted kappa statistic measures agreement beyond chance. Chance agreement refers to the fact that even if the row and column classifications in an agreement table were independent a number of pairs would be located on the leading diagonal. The unweighted kappa statistic gives zero weight to all disagreement cells. Weighted kappa generalizes unweighted kappa by employing differential cell weights which reflect differences in the magnitude of disagreement. More formally, let $n_{ij}$ denote the number of observations (pairs) in the *ith* row and *jth* column of a two-way table, $n_{i+}$ the $i$th row marginal total, $n_{+j}$ the $j$th column marginal total, $n_{++}$ the total sample size, and $w_{ij}$ the weight associated with the $ij$th cell. The weighted kappa statistic is defined as follows:

$$\kappa_w = (p_{ow} - p_{ew})/(1 - p_{ew})$$

where

$$p_{ow} = 1/n_{++} \sum_i \sum_j w_{ij} n_{ij},$$

the observed weighted proportion of pairs in agreement and

$$p_{ew} = (1/n_{++})^2 \sum_i \sum_j (w_{ij} n_{i+} n_{+j}),$$

the weighted proportion of pairs in agreement expected under a model of statistical independence.

There are many possible weighting systems but two systems mentioned by Fleiss [4] are: the squared error weights

$$w_{ij} = 1 - (i-j)^2/(r-1)^2,$$

where $r$ is the number of categories and $i$ and $j$ are category ranks ($1 \leqslant i, j \leqslant r$); and the absolute error weights:

$$w_{ij} = 1 - |i-j|/(r-1).$$

Under both weighting schemes the cells on the leading diagonal have weight equal to one and cells representing extreme disagreement are given zero weight. However for all other cells,

Table 1. Observed primary respondent and proxy reports of alcohol drinking frequency ($n = 456$)*

| | Primary respondent reports | | | | |
|---|---|---|---|---|---|
| Proxy reports | Never drinker | Ex drinker | ⩾1/month <1/week | ⩾1/week <1/day | ⩾1/day |
| Never drinker | 47 | 13 | 19 | 4 | 0 |
| Ex drinker | 5 | 6 | 2 | 1 | 2 |
| ⩾1/month, <1/week | 15 | 6 | 76 | 19 | 4 |
| ⩾1/week, <1/day | 1 | 1 | 23 | 54 | 22 |
| ⩾1/day | 0 | 0 | 4 | 33 | 99 |

*The data consist of reports from 456 primary respondent–proxy respondent pairs drawn from a case–control study of coronary heart disease (the Auckland Heart Study). The primary respondents were coronary heart disease controls and the proxies were their next of kin.

Table 2. Estimates of weighted kappa under various weighting schemes

| Weighting scheme | Kappa | SE |
|---|---|---|
| Unweighted | 0.50 | 0.029 |
| Squared error* | 0.79 | 0.020 |
| Absolute error† | 0.67 | 0.023 |

*Weight for cell $(ij)$ given by $w_{ij} = 1 - (i-j)^2/(r-1)^2$; where $r$ is the dimension of the table (5, in this example).
†Weight for cell $ij$ given by $w_{ij} = 1 - |i-j|/(r-1)$.

squared error and absolute error weights differ. Weighted kappa obtains its maximum value, one, when there is 100% exact agreement. When the row and column classifications are independent, the weighted kappa statistic is equal to zero. The special case of weighted kappa with cells on the leading diagonal given the maximum weight (one) and all other cells given zero weight is the unweighted kappa statistic.

The values of weighted kappa computed from Table 1 under various weighting schemes are shown in Table 2. Clearly, the values of weighted kappa vary considerably according to the weighting scheme employed. For example, the squared error version of weighted kappa is about 6 standard errors greater than the absolute error version. Thus the first problem to be resolved in using weighted kappa is to settle on a weighting system. In the event that investigators use different weighting systems, comparison of weighted kappa statistics from different studies would prove difficult.

Maclure and Willett suggest that the most sensible choice of weights is the squared error weighting system [13]. Fleiss and Cohen have shown that under this weighting scheme, weighted kappa is asymptotically equivalent to the intraclass correlation computed using the category ranks (1, 2, 3 . . . 5 in Table 1) to score responses [14]. The intraclass correlation has been advocated as an agreement index for both continuous and ordinal data [5].

The squared error version of weighted kappa is the version most commonly used in practice and we focus on this version of the statistic in the discussion below. Damiano *et al.* provide a recent example of the use of the absolute error version of weighted kappa although no justification is given for this choice of weights [8].

## WEIGHTED KAPPA CALCULATED WITH SQUARED ERROR WEIGHTS: INDEX OF AGREEMENT OR ASSOCIATION?

Broadly defined, the term association, in the context of a two-way table, includes any departure from independence of the row and column classifications. However, when dealing with ordered data, interest usually centres on departures from independence which are in a particular direction. For example, an investigator may be interested in the extent to which high scores on one variable are predictive of high scores on the other. Complete agreement, which occurs when all the data is concentrated on the leading diagonal, is a special case of association.

In general, however, perfect association does not imply perfect agreement. Table 3 serves to illustrate this point. In this hypothetical table, observer 2 tends to report exactly one category higher than observer 1. The reports of the two observers are strongly associated. In fact, given the first observer's reports the second observer's reports are perfectly predictable. However agreement is very poor as the two observers are in agreement for only 20% of the entities classified. The value of the unweighted kappa statistic for this table is zero while the value of weighted kappa calculated with squared error weights is 0.8.

It seems clear, from this example, that the weighted and unweighted versions of the kappa statistic cannot be measuring the same thing. The high degree of association in Table 3 has produced a high value of weighted kappa in a situation where the level of exact agreement is poor. Damiano *et al.* describe weighted kappa as "the proportion of weighted agreement corrected for agreement attributable to chance" [8]. Logically, the proportion of *weighted* agreement and hence the *weighted* kappa statistic, must be regarded as measures of association rather than measures of exact agreement.

Table 3. Hypothetical example of strong association but poor agreement

| | Observer 2 | | | | |
|---|---|---|---|---|---|
| Observer 1 | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 100 | 0 | 0 | 0 |
| 2 | 0 | 0 | 100 | 0 | 0 |
| 3 | 0 | 0 | 0 | 100 | 0 |
| 4 | 0 | 0 | 0 | 0 | 100 |
| 5 | 0 | 0 | 0 | 0 | 100 |

Squared error weighted kappa: 0.8 (SE: 0.008).
Absolute error weighted kappa: 0.5 (SE: 0.013).
Unweighted kappa: 0 (SE: 0.017).

Table 4. Hypothetical data with identical marginal distributions to Table 1, smaller proportion in exact agreement but identical value of squared error weighted kappa ($n = 456$)

| | Primary respondent reports | | | | |
|---|---|---|---|---|---|
| Proxy reports | Never drinker | Ex drinker | ⩾1/month <1/week | ⩾1/week <1/day | ⩾1/day |
| Never drinker | 57 | 12 | 13 | 1 | 0 |
| Ex drinker | 5 | 3 | 7 | 1 | 0 |
| ⩾1/month, <1/week | 6 | 9 | 63 | 32 | 10 |
| ⩾1/week, <1/day | 0 | 2 | 29 | 38 | 32 |
| ⩾1/day | 0 | 0 | 12 | 39 | 85 |

Squared error weighted kappa: 0.79 (0.019).
Absolute error weighted kappa: 0.62 (0.023).
Unweighted kappa: 0.39 (0.030).

Further evidence that weighted kappa should be regarded as a measure of association rather than agreement is provided by the hypothetical data in Table 4. Both the row and column marginals of Table 4 are identical to those of Table 1 but the proportion in exact agreement in Table 4 is 53.9% compared to 61.8% in Table 1. Despite this difference the two tables yield identical estimates of squared error weighted kappa.

The reason for the identical values of squared error weighted kappa in Tables 1 and 4 follows directly from the definition of the statistic. For, any general weighting system $\{w_{ij}\}$, two tables $\{n_{ij}\}$ and $\{m_{ij}\}$ which have identical marginal distributions (i.e $n_{i+} = m_{i+}$, $n_{+i} = m_{+i}$, $i = 1, \ldots r$) give rise to the same value of the weighted kappa statistic whenever

$$(1/n_{++}) \sum_i \sum_j w_{ij} n_{ij} = (1/m_{++}) \sum_i \sum_j w_{ij} m_{ij}.$$

Under the squared error weighting system and the assumption that the two tables have the same row and column marginals this condition is equivalent to

$$(1/n_{++}) \sum_i \sum_j (ij) n_{ij} = (1/m_{++}) \sum_i \sum_j (ij) m_{ij} \quad (1)$$

It is easily verified that Tables 1 and 4 satisfy this condition.

Given the other conditions, condition (1) is equivalent to the requirement that the correlation between the row and column responses be identical in the two tables. Clearly, two tables can satisfy all the above conditions but differ in the proportion of pairs in exact agreement. The condition (1) does not guarantee that

$$(1/n_{++}) \sum_i n_{ii} = (1/m_{++}) \sum_i m_{ii},$$

i.e. identical correlation coefficients do not imply identical proportions in exact agreement. Thus, amongst tables with the same marginal distributions, squared error weighted kappa is dependent only on the overall correlation between row and column classifications and is not directly dependent on the propensity for exact agreement. A corollary of this is that weighted kappa can appear insensitive to differences in the proportion in exact agreement.

We do not claim that squared error weighted kappa is always insensitive to differences in the proportion in exact agreement but rather, that it can be insensitive. This is an undesirable characteristic for an agreement index. The above examples suggest that weighted kappa should be regarded as a measure of association rather than an index of agreement.

The propensity for pairs to be in agreement should be the focus of agreement analyses. When dealing with ordered data however, the presence of off-diagonal association will usually also be of some interest. For example, when a high level of agreement is observed in a study of reproducibility, the presence of off-diagonal association may further strengthen claims about the underlying quality of a survey instrument. It is not clear how a single index such as weighted kappa can reflect both differences in exact agreement and differences in off-diagonal association. In the next section we briefly discuss the use of statistical models which focus attention on the leading diagonal of an agreement table while also allowing the strength of off-diagonal association to be assessed.

## MODELLING ORDINAL AGREEMENT DATA

Agresti proposed the following quasi-association model for studying ordinal agreement data [10]:

$$\log(m_{ij}) = \mu + \lambda_i^x + \lambda_j^y + \beta u_i u_j + \delta_i I(i = j),$$

where $m_{ij}$ is the expected cell count in the cell defined by the $i$th row and $j$th column, $\mu$, $\lambda_i^x$ and $\lambda_j^y$ are mean, row and column effect parameters respectively, $\beta$ is the association parameter, the $\{u_i\}$ are the category scores and the $\{\delta_i\}$ are agreement parameters which take account of the special features of diagonal cells. $I(i=j)$ is an indicator function which takes the value 1 when $i=j$, i.e when a cell lies on the leading diagonal, and 0 elsewhere. The model can be characterized as "agreement plus linear by linear association" [10, 15].

Not all of the diagonal parameters, $\delta_i$, need be unique. Equality constraints can be imposed on all the $\{\delta_i\}$ or on a subset of these diagonal parameters. Agresti emphasized the model with only one diagonal parameter i.e. $\delta_i=\delta$, $i=1,\ldots r$. [10, 15]. Models with two or more diagonal parameters may be useful if variation by category in the propensity for agreement is of interest. Strictly speaking, the term quasi association refers specifically to the model with all diagonal parameters distinct. However, for convenience, our usage of the term includes models with equality constraints on the diagonal parameters.

Category scores for quasi-association models must either be specified by the investigator or estimated from the data. If the scores are estimated, the quasi-association model is log non-linear and cannot be fitted using conventional statistical software. The integer scores, $(1, 2, 3 \ldots r)$ will often be a sensible choice for the category scores, however when the ordinal scale is derived by grouping an underlying continuous scale, the category midpoints provide a sensible alternative set of scores.

The special case of Agresti's quasi-association model with $\beta=0$ is the quasi-independence model proposed by Tanner and Young for studying nominal scale agreement [16]. The special case with $\delta_i=0$, $i=1,\ldots r$ is the linear by linear association model, which is a useful model for studying the association between two ordinal variables [15]. The special case of quasi association with $\beta=0$ and $\delta_i=0$, $i=1,\ldots r$ is the independence model. Thus, the quasi-association model can be used to study the extent of agreement beyond chance, the extent of agreement beyond linear by linear association ($\delta_i>0$) and also the extent of off-diagonal association ($\beta$ non-zero). The model partitions beyond chance agreement into a component due to linear by linear association and a component due to agreement beyond linear by linear association.

The linear by linear association, quasi-independence, and quasi-association models are more fully described in the cited references.

The likelihood ratio chi-square statistic can generally be used to compare the fit of the quasi-association model with the fit of special cases of the model such as linear by linear association and quasi independence. When one model is a special case of another, more general model, it is usually the case that the difference in likelihood ratio statistics for the two models has a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom for the two models. An exception to this rule is the comparison of estimated category scores models with the independence model. Likelihood ratio differences for such models do not in general follow a chi-square distribution and consequently cannot easily be used to compare goodness of fit [2, 15]. The likelihood ratio statistic can be used to compare estimated scores models with their fixed scores counterparts however [2].

Standard methods can be used to test parameter estimates for significant departures from a null value. For example, when the one diagonal parameter version of quasi association is fitted, the hypothesis $H_0$: $\delta=0$ can be tested against $H_A$: $\delta>0$ by referring $\hat{\delta}/s(\hat{\delta})$ to the standard normal distribution, where $\hat{\delta}$ is the maximum likelihood estimate of $\delta$ and $s(\hat{\delta})$ is the estimated asymptotic standard error of $\hat{\delta}$.

*Application to Table 1*

Table 5 summarizes the results of fitting independence, quasi independence, linear by linear association and quasi-association models

Table 5. Goodness of fit statistics for various models fitted to Table 1

| Model | $G^{2*}$ | $df$† |
|---|---|---|
| Independence | 470.78 | 16 |
| Quasi independence‡ | 156.92 | 15 |
| Linear by linear association | 69.91 | 15 |
| Quasi association‡ | 41.61 | 14 |
| Estimated scores models: | | |
| linear by linear association | 33.02 | 12 |
| quasi association‡ | 17.06 | 11 |

*Likelihood ratio chi-square statistic, for nested models the difference in likelihood ratio chi-square statistics has a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom for the models.

†Residual degrees of freedom.

‡Both quasi-independence and quasi-association models include only one diagonal parameter.

to the data of Table 1. In the interests of simplicity of interpretation, only the special cases of quasi independence and quasi association with a single diagonal parameter are reported in Table 5. Fixed scores versions of the linear by linear association model and the quasi-association model were fitted using the integer scores (1–5) as category scores. All models were fitted by maximum likelihood. Proc Catmod in SAS [17] was used to fit all the models except the estimated scores models for which a SAS program adapted from Becker's algorithm [18] was written.

Comparing the goodness of fit of the quasi-independence model to that of the independence model reveals clear evidence that the number of pairs on the leading diagonal differs markedly from the number expected under an independence model ($\chi^2 = 313.86$, $df = 1$, $p < 0.00001$). Since the data is ordinal it is more noteworthy that the fit of the quasi-association model was significantly better than the fit of the linear by linear association model ($\chi^2 = 28.31$, $df = 1$, $p < 0.00001$). The estimate of the agreement parameter, $\delta$, was 0.734 (asymptotic standard error: 0.136). Thus, there is strong evidence of agreement beyond linear by linear association. There is also strong evidence of off-diagonal association since the quasi-association model fitted significantly better than the quasi-independence model ($\chi^2 = 115.31$, $df = 1$, $p < 0.00001$). The estimate of the association parameter, $\beta$, under the quasi-association model was 0.616 (asymptotic standard error: 0.081). Overall, none of the above models fit particularly well.

The estimated scores models appear to fit much better. A comparison of the fit of the estimated scores models with that of their fixed score counterparts yields chi-square statistics with 3 *df* which are highly significant ($p < 0.00001$). The improved fit can be attributed to differences between the estimated scores and the equal interval integer scores. For the quasi-association model the estimated scores, which were constrained to have a minimum value of 1 and a maximum value of 5 in order to facilitate comparison with the integer scores, were 1, 1.24, 2.44, 3.79, 5. The estimated scores suggest that the first two categories, never-drinker and ex-drinker are much closer than indicated by the equal interval integer scoring. This reflects the relatively poor agreement between primary and proxy respondents regarding never-drinker vs ex-drinker status (see Table 1).

The difference between the likelihood ratio statistics for the estimated scores versions of linear by linear association and quasi association was also highly significant ($\chi^2 = 15.96$, $df = 1$, $p < 0.00001$). The estimate of $\delta$ under the estimated scores version of quasi association was 0.603 (jackknife estimate of standard error: 0.152) and the estimate of $\beta$ was 0.648 (jackknife estimate of standard error: 0.091). As with the fixed scores version of quasi association, the parameter estimates are strongly suggestive of both agreement beyond association and off-diagonal association. The improved fit of the estimated scores model compared to the fixed scores model lends more credence to inferences based on the estimated scores model.

Although the likelihood ratio chi-square statistic for the estimated scores version of quasi association suggests that the model fits quite well, examination of the Pearson chi-square statistic ($\chi^2 = 48.68$, 11 *df*) does not yield the same conclusion. This serves as a reminder of the sparse nature of Table 1 at the extremes of disagreement and the sensitivity of the Pearson chi-square to small cell counts. Most of the lack of fit is due to the cell: primary respondent reports one drink or more per day, proxy reports ex-drinker. The observed count for this cell is 2, while the predicted cell count under the estimated scores version of quasi association is 0.1.

## DISCUSSION

There are several inherent problems in the use of the weighted kappa statistic for the analysis of ordinal agreement data. The choice of weighting scheme can greatly influence the estimated value of the statistic. Unless a standard weighting scheme is used the comparison of weighted kappa statistics from different studies would be very difficult. Maclure and Willett [13] and Fleiss and Cohen [14] have suggested a sensible choice of standard weights, but even with this choice of weights, we have shown that weighted kappa is not always sensitive to differences in the observed proportion in exact agreement. While it is true that weighted kappa achieves its maximum value when the data are concentrated on the leading diagonal, it is also true that high values of weighted kappa can be achieved when the proportion in exact agreement is low (see Table 3).

In general it would appear that weighted kappa should be regarded primarily as a measure of association, a conclusion also arrived at by Bloch and Kraemer, in the somewhat different context of dichotomous data [19]. Altman and Bland argued that the Pearson correlation coefficient should not be regarded as an agreement index for continuous data as it measures association, not agreement [20]. A similar argument could be applied to the weighted kappa statistic.

Some investigators have used unweighted kappa to analyse ordinal agreement data [21–23]. While this avoids the need to arbitrarily choose a weighting scheme and although unweighted kappa is generally sensitive to changes in the proportion in exact agreement, this is an unsatisfactory solution. By definition, unweighted kappa is completely blind to off-diagonal association. Therefore, unweighted kappa cannot detect differences in off-diagonal association and consequently cannot, by itself, provide a complete description of ordinal agreement data.

A further difficulty with the use of unweighted kappa with ordinal data is that it appears to be strongly dependent on the number of categories used [13]. This is particularly important when an ordinal scale is derived from a continuous measurement, in which case both the definition and number of categories are often determined arbitrarily.

The modelling of agreement data yields insights not easily obtained through the use of kappa statistics. By comparing the fit of the appropriate models and examining the appropriate parameter estimates, it becomes possible to establish whether a given table exhibits agreement beyond that expected under a linear by linear association model and also whether there is evidence of positive association off the leading diagonal. In addition, models with estimated category scores can provide further information, such as the identification of pairs of categories for which agreement is relatively poor.

Darroch and McCloud [11] argued that the issue of agreement for polychotomous data is closely related to the distinguishability of individual pairs of categories. They provide a technical definition of category distinguishability based on odds ratios for certain sub-tables of an agreement table. In addition, they demonstrate the connection between the modelling of nominal agreement data and estimation of category distinguishability. A similar connection exists for ordinal data [2, 10]. For example, the version of the quasi-association model with a single diagonal parameter and fixed, equal interval category scores assumes that all pairs of adjacent categories are equally distinguishable. Estimated scores models make no such assumption. Category distinguishability can be estimated for each pair of categories in an agreement analysis, either by direct estimation based on the observed data, or via modelling. Although direct, as opposed to model-based, estimation of category distinguishability does not explicitly take account of category ordering in the estimation process, the resulting estimates will usually reflect any ordering inherent in the data. Typically, estimated category distinguishability increases with the distance between categories [12].

Given the problems we have outlined with weighted kappa and the availability of alternative methods of analysis we feel that the continued use of weighted kappa as an agreement index is questionable. If the weighted kappa statistic is to be used it should be supplemented with some other analytical strategy such as the modelling approach outlined in this paper or the estimation of category distinguishability referred to above.

*Acknowledgements*—Supported by the Health Research Council of New Zealand, the National Heart Foundation of New Zealand and the Canterbury Medical Research Foundation.

## REFERENCES

1. Feinstein A. A bibliography of publications on observer reliability. **J Chron Dis** 1985; 38: 619–632.
2. Becker MP. Using association models to analyse agreement data: Two examples. **Stat Med** 1989; 8: 1199–1207.
3. Jackson R, Scragg R, Beaglehole R. Alcohol consumption and risk of coronary heart disease. **Br Med J** 1991; 303: 211–216.
4. Fleiss JL. **Statistical Methods for Rates and Proportions**, Chap. 13, 2nd edn. New York: Wiley; 1981.
5. Nelson LM, Longstreth WT Jr, Koepsell TD *et al.* Proxy respondents in epidemiologic research. **Epidemiol Rev** 1990; 12: 71–86.
6. Thompson FE, Lamphiear DE, Metzner HL *et al.* Reproducibility of reports of frequency of food use in the Tecumseh diet methodology study. **Am J Epidemiol** 1987; 125: 658–671.
7. Metzner Hl, Lamphiear DE, Thompson FE *et al.* Comparison of surrogate and subject reports of dietary practices, smoking habits and weight among married couples in the Tecumseh diet methodology study. **J Clin Epidemiol** 1989; 42: 367–375.
8. Damiano AM, Bergner M, Draper EA *et al.* Reliability of a measure of severity of illness: Acute physiology of

chronic illness evaluation—II. **J Clin Epidemiol** 1992; 45: 93–101.
9. Bishop J, Carlin J, Nolan T. Evaluation of the properties and reliability of a clinical severity scale for acute asthma in children. **J Clin Epidemiol** 1992; 45: 71–76.
10. Agresti A. A model for agreement between raters on an ordinal scale. **Biometrics** 1988; 44: 539–548.
11. Darroch JN, McCloud PI. Category distinguishability and observer agreement. **Aust J Stat** 1986; 28: 371–388.
12. Graham P, Jackson R. Primary versus proxy respondents: Comparability of questionnaire data on alcohol consumption. **Am J Epidemiol** In press.
13. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. **Am J Epidemiol** 1987; 126: 161–169.
14. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. **Educ Psychol Meas** 1973; 33: 613–619.
15. Agresti A. **Categorical Data Analysis**. New York: Wiley; 1991: 261–370.
16. Tanner MA, Young MA. Modelling agreement among raters. **JASA** 1985; 80: 175–180.
17. SAS Institute Inc. **SAS/STAT™ User's Guide, Release 6.03 Edition**. Cary NC:SAS Institute Inc.; 1988.
18. Becker MP. Algorithm AS 253. Maximum likelihood estimation of the RC(M) association model. **Appl Stat** 1990; 39: 52–167.
19. Bloch DA, Kraemer HC. $2 \times 2$ Kappa coefficients: Measures of agreement or association. **Biometrics** 1989; 45: 269–287.
20. Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. **Statistician** 1983; 32: 307–317.
21. Marshall J, Priore R, Haughey B *et al.* Spouse–subject interviews and the reliability of diet studies. **Am J Epidemiol** 1980; 112: 675–683.
22. Humble CG, Samet JM, Skipper B. Comparison of self and surrogate reported dietary information. **Am J Epidemiol** 1984; 119: 86–98.
23. Magaziner J, Simonsick EM, Kashner TM *et al.* Patient proxy response comparability on measures of patient health and functional status. **J Clin Epidemiol** 1988; 41: 1065–1074.