AMERICAN

Journal of Epidemiology

Formerly AMERICAN JOURNAL OF HYGIENE

VOL. 126 AUGUST 1987 NO. 2

# Reviews and Commentary

## MISINTERPRETATION AND MISUSE OF THE KAPPA STATISTIC

MALCOLM MACLURE[1] AND WALTER C. WILLETT[1,2]

As more studies of the reproducibility and validity of epidemiologic tools are reported in the *American Journal of Epidemiology*, the Kappa statistic ($\kappa$) conceived by Cohen (1) is increasingly seen (2–14). Originally proposed as a measure of agreement between two observers classifying subjects into two nominal categories, it has been extended to multicategory classifications and used to assess not only reproducibility but also validity. A survey of its usage in the *Journal* over a three-year period revealed frequent misinterpretation and, in our opinion, some misuse.

Kappa was meant to be an improvement on the simpler measure, per cent of agreement, because it discounts the proportion of agreement which is expected by chance alone ($P_e$). Instead of the total proportion of observations on which there is agreement ($P_o$) being compared as a ratio with its maximum value (100 per cent), the attributable proportion ($P_o - P_e$)—the fraction of observations for which agreement can be attributed to the reproducibility of the observations rather than to mere chance—is compared as a ratio with its maximum possible value ($1 - P_e$). Thus,

$$\kappa = \frac{P_o - P_e}{1 - P_e}.$$

While Kappa is arguably an improvement on alternative measures of the reproducibility of dichotomous nominal categorizations (15), it does not follow that extending its application to polytomous nominal or ordinal categorizations, or to the assessment of validity, is also an improvement on available alternatives. We suggest here that: 1) with continuous data grouped into ordinal categories, usually just for the convenience of the investigator, Kappa is so arbitrary it is virtually meaningless; 2) with naturally ordinal data, the intraclass correlation coefficient (16) is superior to Kappa; 3) with polytomous nominal data, the use of several Kappas for different combinations of dichotomies may be more informative than an overall Kappa for the polytomy; and 4) when assessing validity, there are better alternatives to Kappa, i.e., for nominal data, there are sensitivity and specificity, or predictive value (17), and for ordinal and continuous data, there are the mean and standard deviation of the difference between the new measurement and the valid reference measurement (18) or the

[1] Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. (Send reprint requests to Dr. Maclure at this address.)

[2] Channing Laboratory, Department of Preventive Medicine and Clinical Epidemiology, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115.

Supported in part by the American Cancer Society (SIG 10-11) and Public Health Service Research Career Development Award HL-01018 from the National Institutes of Health.

product-moment (interclass) correlation coefficient (16).

The principal weakness of Kappa stems from it being a measure of the frequency of exact agreement rather than a measure of the degree of approximate agreement. This is not a weakness with dichotomous data since observations then either agree exactly or they totally disagree. With polytomous nominal data, however, one pair of categories may be considered more dissimilar than another, so some instances of disagreement are worse than others. Similarly, with ordinal data, a pair of observations two categories apart is a greater disagreement than a pair in adjacent categories. Kappa treats all instances of disagreement, large or small, as identical. Weighted Kappa (19), developed to address this weakness, has its own potential weakness: it allows weights to be arbitrary in relative magnitude, which means the magnitude of weighted Kappa may be arbitrary. To avoid this arbitrariness, standard weights should be used. It turns out, however, that a logical choice of standard weights makes weighted Kappa equivalent to the intraclass correlation coefficient (20).

## Continuous data

First let us consider the most obvious misuse of Kappa—its application to continuous data. Data which are continuous, or virtually continuous, are often grouped, for the convenience or needs of the investigator, into categories of arbitrary number and size. To illustrate this and subsequent points, we use data on sucrose intake from the evaluation of a semiquantitative food frequency questionnaire administered twice to 173 nurses one year apart (21). The questionnaire was meant to enable estimation of relative risks for various diseases according to level of intake of various nutrients, including sucrose. To compute relative risks, it is necessary to group subjects by level of nutrient intake. Therefore, it is convenient to think of reproducibility in terms of the groupings by sucrose intake rather than of the exact value of sucrose intake scores. For example, the reproducibility of our sucrose intake data may be shown as in table 1, a 12 × 12 array in which the 173 nurses are cross-classified by their ranks on the first and the second occasions they completed the questionnaire.

Given such an array, it may be tempting to compute a Kappa statistic. The problem with such a computation is that the definition of exact agreement is arbitrary. As a result, the magnitude of Kappa is dependent more upon how the categories were defined than upon the degree of reproducibility of observation methods. This can be shown algebraically.

When we divide a continuous distribution of observations (say, sucrose intake scores derived from our questionnaire) into $j$ intervals by equal percentiles, the probability of an observation falling into any given interval is $j^{-1}$. If the sucrose data happened to be completely unreproducible, the joint probability of two observations from the same subject on two occasions falling in the same interval would be the square of this probability, $j^{-2}$. Since there are $j$ intervals in which exact agreement can occur, the probability of two observations falling in a same interval by mere chance is $j^{-1}$. This is the expected proportion of exact agreement: $P_e = j^{-1}$. In the limit as $j \longrightarrow \infty$ (assuming the sample size $n \longrightarrow \infty$), $P_e = 0$ and therefore $\kappa = P_o$.

Now what happens to $P_o$ as the number of categories arbitrarily increases? Only if the questionnaire unfailingly yields replicate observations in exact agreement, to the smallest decimal place permitted by the data, will $P_o$ be independent of the number of categories chosen ($P_o = 1$). Realistically, the two methods yield pairs of observations which differ by some discrepancy, $\delta_i$. As the number of intervals ($j$) increases and, consequently, the width of each interval ($W_i$) decreases until it is small relative to $\delta_i$, the proportion of discrepancies which exceed the widths of an interval rises. In the limit as $j \longrightarrow \infty$, $\Pr(W_i > \delta_i) = 0$ and $P_o = 0$, so $\kappa = 0$. In other words, as the definition of exact agreement is arbitrarily narrowed, the proportion of observations which exactly agree is arbitrarily reduced.

## Ordinal data

The same sort of problem occurs with ordinal data and can be readily illustrated by the data in table 1.

The total number of observations which exactly agree in the 12 × 12 table is 31 (the diagonal from top left to bottom right), and the corresponding $P_o$ is 0.18. Kappa equals 0.10. If every consecutive pair of categories is merged so as to form a 6 × 6 table, the definition of exact agreement is relaxed and the number of concordant observations rises to 62, with a $P_o$ of 0.36. Kappa now equals 0.23. Similarly, by collapsing categories to make a 4 × 4 table, a 3 × 3 table, or a 2 × 2 table, the proportion of agreement arbitrarily rises to 0.50, 0.58, or 0.77, respectively. Paralleling this progression are higher values for Kappa—0.33, 0.38, and 0.55. Clearly, in this situation, the values for Kappa are so greatly influenced by the number of categories that a four-category Kappa for ordinal data cannot be compared with a three-category Kappa. This is not widely appreciated, as we will see below.

The intraclass correlation coefficient using marginal scores (16) is less sensitive to changes in the number of categories, and tends to increase rather than to decrease with the number. For the 2 × 2, 3 × 3, 4 × 4, 6 × 6, and 12 × 12 tables, the intraclass correlation coefficients are 0.55, 0.55, 0.63, 0.64, and 0.66, respectively, scoring each observation by the rank of the marginal category into which it is grouped (e.g., scores 1 to 3 for the 3 × 3 table and 1 to 12 for the 12 × 12 table). The coefficient for the continuous data, before they were grouped into ordinal categories, was 0.61 (and 0.70 after a natural log transformation).

Perhaps the lack of appreciation of Kappa's dependence on number of categories is due to the fact that the "overall Kappa" for a polytomy can be shown to be a weighted average of the "individual Kappas" for all the alternative dichotomies which can be made by preserving one category and combining all others (15). Thus, Kappa for the above mentioned 3 × 3 table (0.38) is a weighted average of the Kappas for the three 2 × 2 tables with the following three dichotomies as margins: the top third vs. the bottom two-thirds of the sucrose distribution ($\kappa = 0.44$), the bottom third vs. the top two-thirds ($\kappa = 0.48$), and the middle third vs. the two extreme thirds ($\kappa = 0.21$). The weights are the denominators of the individual Kappas. Note that the third dichotomy is the principal reason why the

Table 1

*Cross-classification of subjects by dodeciles of sucrose intake measured by a food frequency questionnaire administered twice, one year apart, to 173 Boston-area female registered nurses aged 34–59 years in 1980–1981 (21)*

| Second questionnaire dodeciles | First questionnaire dodeciles | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 7 | 4 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 3 | 3 | 4 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 3 | 1 | 0 | 2 | 2 | 2 | 3 | 1 | 2 | 1 | 0 | 1 | 0 |
| 4 | 1 | 3 | 2 | 3 | 3 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 5 | 0 | 2 | 1 | 2 | 1 | 5 | 0 | 1 | 2 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 3 | 1 | 3 | 3 | 3 | 0 | 0 | 1 | 0 |
| 7 | 1 | 0 | 1 | 0 | 1 | 2 | 3 | 0 | 3 | 1 | 1 | 1 |
| 8 | 0 | 0 | 2 | 1 | 2 | 1 | 0 | 1 | 2 | 5 | 0 | 1 |
| 9 | 1 | 1 | 2 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 3 | 1 |
| 10 | 0 | 0 | 0 | 1 | 0 | 0 | 4 | 2 | 3 | 1 | 1 | 3 |
| 11 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 5 | 1 | 5 |
| 12 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 2 | 5 | 4 |

overall Kappa for the trichotomy is much less than Kappa for the dichotomy—the top half vs. the bottom half—for which $\kappa = 0.55$. This illustrates a distinction between the use of Kappa for ordinal vs. nominal data. With ordinal data, an intermediate category will often be subject to more misclassification than an extreme category because there are two directions in which to err away from an intermediate position but only one direction in which to err away from the extremes. With nominal data, there are no extremes since there is no directionality, so no individual Kappa will have an inherent tendency to be larger or smaller than any other.

If Kappa's weakness is its blindness to differences in degree of disagreement, perhaps a modification of Kappa which accounts for severity of discordance or size of discrepancy would be better. Weighted Kappa (19) is just such a modification. It permits one to give a unique weight to each possible combination of discordant observations. Weighted Kappa (and also unweighted Kappa) can be expressed in two equivalent ways: 1) as an index of agreement in which weights are maximal for exact agreement and minimal for extreme disagreement, or 2) as an index of disagreement in which the weights are minimal (perhaps zero) for exact agreement and increase with the degree of disagreement. The latter approach gives the simplest formula:

$$\kappa_w = 1 - \frac{\sum_{ij} w_{ij} o_{ij}}{\sum_{ij} w_{ij} e_{ij}},$$

where $o_{ij}$ is the observed frequency in cell $ij$, $e_{ij}$ is the frequency which would be expected in that cell due to chance, and $w_{ij}$ is the weight given to that cell. Typically, $w_{ij} = 0$ for all instances of exact agreement ($i = j$), but it may take any value in the off-diagonals ($\imath \neq j$). If the latter are all given equal weights, weighted Kappa is found to equal the unweighted Kappa we have been discussing until now.

While weighted Kappa is, in theory, a solution to the limitations of unweighted Kappa, it poses a new practical problem: what weights should be chosen? Since the magnitude of weighted Kappa is greatly influenced by the relative magnitudes of its weights, it will be uninterpretable unless there is some standardization in usage of weights by different investigators. The goal, after all, is to report an index of agreement which is interpretable by a wide readership.

For ordinal variables, an intuitively appealing standard usage would be to weight each instance of disagreement by the square of the deviation of the pair of observations from exact agreement. In our 12 × 12 array of subjects by their sucrose intake, we would assign a weight of zero to the 12 cells on the diagonal ($i = j$), a weight of 1 to the 22 cells immediately adjacent to the diagonal (where $|i - j| = 1$), and weights of 4, 9, 16, 25, and so on, to cells for which $|i - j| = 2, 3, 4, 5$, and so on, respectively. This yields a weighted Kappa of 0.66. As Cohen (19) noted, the weighted Kappa calculated with these weights approximately equals the product-moment correlation coefficient for the array of data scored by the ranks of the marginal categories (i.e., sc̦ored 1 to 12 in our sample). Moreover, he observed that, under the condition of identical distributions in the margins, the weighted Kappa is exactly equivalent to the intraclass correlation coefficient (19, 20).

Thus, we find that a logical modification of Kappa for ordinal data is one which leads us back to the correlation coefficient. This is a reassuring result since the main reason for studying reproducibility is to assess the degree to which a measure of effect might be underestimated due to nondifferential random misclassification (22). As Walker and Blettner (23) have shown, the correlation coefficient can be used with the assumption of bivariate normality to construct "misclassification matrices". These matrices may in turn be used to infer the degree of nondifferential misclassification bias.

## Nominal polytomous data

If the categories of a polytomy are natural or fixed by convention and they have no inherent order (e.g., blood type, marital status, or type of hepatitis virus), then an overall unweighted Kappa for two sets of observations is free of the above-mentioned arbitrariness and may be the best measure of overall agreement. Still, it should be understood by authors and readers that the overall Kappa is an average and that, as such, it may hide the fact that one category accounts for most of the misclassification. Moreover, it should be appreciated that in choosing among alternative nominal classification schemes (e.g., white/black/other vs. non-Hispanic white/Hispanic/black/Asian/other), the Kappa for the more detailed classification scheme will be lower. In addition, one should be alert to the possibility that an apparently nominal polytomy might be better viewed as ordinal (e.g., the categories yes/no/don't know, or never smoked/current smoker/ex-smoker might be better treated as ordinal trichotomies yes/don't know/no, or never smoked/ex-smoker/current smoker).

An alternative to reporting an overall Kappa for a nominal polytomy is to report Kappas for the component dichotomies. This circumvents the problem of treating all types of disagreement equally because it allows the different Kappas to be weighted at the reader's discretion. The main disadvantage is that more summary statistics need to be displayed in tables which may already be busy. But such a price may be worth paying for the additional information.

A characteristic of Kappa which should be mentioned is its variation with changes in prevalence of the phenomenon being measured. (This is analogous to the intraclass correlation coefficient's dependence on the variance of the quantity being measured (20)). To correct for this, Walter (24) proposes the use of "maximum Kappa": it is the maximum among all Kappas for different prevalences of the characteristic, given fixed sensitivities and specificities of the two measurements. While this is an ingenious solution, more often than not reproducibility is assessed in the absence of information on the actual sensitivities, so their values can only be assumed.

## Assessment of validity

Kappa was proposed as a measure of reproducibility. When the assessment of a method focuses instead on its validity—as a surrogate for, or predictor of, the truth as estimated by a more valid reference measurement—there is asymmetry between the two measurements, so indices of agreement other than Kappa may be more informative. In the case of dichotomous data, there are the sensitivity and specificity, or predictive values positive and negative (17). In the case of continuous data, there are the mean and standard deviation of the difference (18) or the product-moment correlation coefficient (16). (Validity of ordinal data may be measured in the same way as continuous data if the values for the ordinal categories are a standard numerical scale.) Choosing among these alternatives, the trade-off is between generalizability and intelligibility.

If the investigators' goal is to report the validity of a method for dichotomous classification, which they hope will be widely generalizable and have utility in diverse populations among whom the prevalence of the phenomenon varies markedly, then sensitivity and specificity have the advantage of being invariant with changes in prevalence (17). Predictive values for any specific population can always be calculated given an estimate of prevalence. However, if the prevalence is relatively invariant among the populations in which the investigators hope the method will have utility, then the predictive values positive and negative obtained from the validation study are fairly generalizable. Moreover, from the viewpoint of the user of the method, the predictive values are more intelligible and practically usable measures of validity than sensitivity and specificity because, by incorporating information on both accuracy and prevalence, they describe the performance of the method in an actual population. In a clinical setting where a test is applied to patients with expected prevalences markedly different from the general population, the predictive value of a test assessed in a comparable clinical setting is applicable, whereas the predictive value from the same test assessed in the general population is not.

Phi ($\phi$), the correlation coefficient for the 2 × 2 table, is convenient because it is a single number which is generalizable among populations with similar prevalences. However, the use of sensitivity, specificity, and predictive values remains preferable in most cases for the abovementioned reasons.

Analogous conditions exist for ordinal or continuous data. If the investigators' goal is to report the validity of an exposure measurement method, in the hope that it will be widely used among populations with diverse exposure distributions, then it is useful to report the mean and standard deviation of the difference between the surrogate measurement being assessed and the valid reference value (18). These measurements, like sensitivity and specificity, are independent of the exposure distribution in the study population (if the measurement error is constant across the range of true values). However, if the investigators' goal is to report the accuracy of an exposure assessment method as concisely and intelligibly as possible, with the understanding that the summary statistic is generalizable only to populations with similar exposure distributions, then the product-moment correlation coefficient (16) is preferable. The greater intelligibility of the correlation coefficient derives from its being a single number and dimensionless, so there is no need to remember typical ranges of values for every exposure variable measured. Furthermore, for epidemiologic purposes, the correlation coefficient is directly related to misclassification and statistical power (23). By contrast, the mean difference and its standard deviation cannot be interpreted without comparing them to the range of values in a typical population, and their use to obtain a prediction interval about any particular measurement is not straightforward.

The interpretation of a regression coefficient and the residual standard deviation, from a regression of the surrogate on the reference measurement, also requires knowledge of the true range. The choice between using the mean difference versus the regression coefficient has been debated (18, 25) but space forbids a discussion of the issues here. One benefit of the latter approach is that, given the residual sums of squares (RSS) about the regression line from a published validation study done in population A (in which there was a wide range of exposure levels), and the total sum of squares (TSS) of the surrogate measurements in a particular population B, one can estimate the correlation coefficient ($r$) which would have been seen if the validation study were repeated in population B, since $r^2$ = (TSS − RSS)/TSS. (This is analogous to estimating the predictive values given knowledge of a particular prevalence, plus the sensitivity and specificity.)

Since data are often put to uses other than those intended by the investigators, it makes sense to report results of validation studies more than one way. The reader is then free to make the choice between generalizability and intelligibility.

## How kappa has been used

We surveyed the issues of the *American Journal of Epidemiology* during the three-year period July 1983 to June 1986, and we found 13 papers which used Kappa statistics (2–14). The papers by van Leeuwen et al. (3) and van Staveren et al. (14) were the only ones to use Kappa to compare continuous variables (nutrient scores) after grouping subjects into discrete categories. However, the authors clearly stated that their categorization schemes were by tertiles and they presented Kappas only as supplements to correlation coefficients and other measurements of degree of disagreement. Moreover, since the marginal distributions were constant, the Kappas were mutually comparable. Nevertheless, one must question their meaningfulness. As the tables show, in contrast to the correlation coefficients, the Kappas were relatively insensitive to the amount of misclassification between opposite tertiles (a more serious type of disagreement than misclassification between adjacent tertiles).

In six reports (2, 5, 9–11, 13), the authors explicitly or implicitly compared Kappas for dichotomous data with Kappas for ordinal polytomous data. This was done 1) by making comparative statements in the text of the paper, 2) by linking Kappa values to qualitative terms proposed by Landis and Koch (26) (e.g., excellent, fair, and poor), and/or 3) by presenting different types of Kappa together in one table without commenting on their dependence on number of categories. The tables in Brilliant et al. (2) demonstrate nicely how Kappas for polytomies tend to be lower than those for dichotomies.

The same problem occurs more surreptitiously when Kappa is used to assess reproducibility of food frequency questionnaires without grouping subjects by percentiles (4, 7, 10, 11, 13). Foods which are consumed rarely, such as liver, tend to yield higher Kappas than foods with much greater variation in intake, such as carrots. This is largely because Kappa for liver typically is a Kappa for a dichotomy (≤1/month vs. 2–3/month), whereas that for carrots may be a five-category Kappa (≤1/month, 2–3/month, 1–2/week, 3–6/week, ≥1/day). The use of correlation coefficients can give quite a different picture. In the data of Lerchen and Samet (13), the question on carrots yielded a correlation coefficient of 0.52 compared with a coefficient of 0.39 for liver, whereas the Kappas were 0.21 for carrots and 0.24 for liver.

The data of Vitéz et al. (5) illustrate how Kappas for three different combinations of dichotomies are more informative than an overall Kappa for a nominal trichotomy. Their Kappa of 0.52 for exact agreement of their trichotomous classification of study subjects by discriminant analysis, compared with their original classification of subjects as either drinkers, abstainers, or controls, leaves the reader unsatisfied. The question which immediately comes to mind is what would the Kappas for the three dichotomies—drinkers vs. others, abstainers vs. others, and controls vs. others—have been using the same discriminant model? Using the raw numbers from their 3 × 3 table, we calculate those Kappas to be 0.53, 0.27, and 0.67, respectively, and thereby obtain a better idea of the performance of their discriminant analysis than we had from the overall Kappa for the trichotomy. The low Kappa for the abstainers compared to others is due to the fact that "abstainer" is a category intermediate between "drinker" and "control", if the trichotomy is viewed as ordinal.

Significance tests of Kappas were reported in three papers (4, 12, 13). This seems inappropriate since the purpose of a validation study is to estimate the degree of agreement, not to test a null hypothesis of no agreement. Significance can always be achieved by increasing sample size.

Weighted Kappa was used in one paper (8). Each of six types of disagreement between two ordinal trichotomous classifications was assigned a unique weight according to the investigators' opinions of the seriousness of that type of disagreement, with weights ranging from 0.95 to 0.005. Weighted Kappa was calculated in two instances as 0.72. Although the weighting scheme was rational, it was nonstandard. It could even be considered arbitrary insofar as higher or lower values for Kappa could be obtained by choosing a different set of weights. Therefore, when the authors described their Kappa of 0.72 as evidence of "excellent" agreement, it was not clear how legitimate this was, nor how 0.72 would compare with other weighted Kappas. It would have been more interpretable and therefore more informative to report the weighted Kappas with squared deviations from agreement as weights. We calculate them to be 0.71 and 0.72—no different, as it happens, but more interpretable because they approximate the standard intraclass correlation coefficients.

Kappas were sometimes used in ways which might be interpreted as assessments of validity rather than reproducibility. In the paper by Vitéz et al. (5), sensitivity and specificity might have been more appropriate. Clinical interview data on whether a mother was a drinker, an abstainer, or a normal control during pregnancy were taken as the benchmarks against which a discriminant model based on fetal characteristics was evaluated, as if the latter were a diagnostic test. We calculate that their model had a sensitivity of 57 per cent to drinkers and corresponding specificities of 24 per cent among abstainers and 4 per cent among controls.

We found several other papers (3, 8, 9, 12, 14) in which can be found applications of Kappa to the assessment of agreement between two methods, one of which is viewed as more definitive than the other. If the investigator considers the assessment of agreement to be a matter of validity rather than reproducibility, then it would be more appropriate to use sensitivity, specificity, and predictive value, the mean and standard deviation of the difference or the product-moment correlation coefficient. The last paper (14) is a good illustration of the use of the mean and standard deviation of the difference as measures of validity.

A summary statistic is meant to be a tool for communication. If it does not communicate well because the weighting scheme is nonstandard, or there is a large component of arbitrariness in its definition or, worse still, if it misleads the reader by appearing to say something it does not, then it is not a good tool. Our conclusion is that unweighted Kappa is a satisfactory measure of agreement only for dichotomous variables, and weighted Kappa is best when it equals the intraclass correlation coefficient.

## REFERENCES

1. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
2. Brilliant LB, Lepkowski JM, Musch DC. Reliability of ophthalmic diagnoses in a epidemiologic survey. Am J Epidemiol 1983;118:265–79.
3. van Leeuwen FE, de Vet HCW, Hayes RB, et al. An assessment of the relative validity of retrospective interviewing for measuring dietary intake. Am J Epidemiol 1983;118:752–8.
4. Humble CG, Samet JM, Skipper BE. Comparison of self- and surrogate-reported dietary information. Am J Epidemiol 1984;119:86–98.
5. Vitéz M, Korányi G, Gönczy E, et al. A semiquantitative score system for epidemiologic studies of fetal alcohol syndrome. Am J Epidemiol 1984; 119:301–8.
6. Holman CDJ. Analysis of interobserver variation on a programmable calculator. Am J Epidemiol 1984;120:154–60.
7. Little RE, Worthington-Roberts B, Mann SL, et al. Test-retest reliability of diet and drinking estimates for pregnancy and post partum. Am J Epidemiol 1984;120:794–7.
8. Pfeffer RI, Kurosaki TT, Chance JM, et al. Use of the Mental Function Index in older adults: reliability, validity, and measurement of change over time. Am J Epidemiol 1984;120:922–35
9. Tilley BC, Barnes AB, Bergstralh E, et al. A comparison of pregnancy history recall and medical records: implications for retrospective studies. Am J Epidemiol 1985;121:269–81.
10. Herrmann N. Retrospective information from questionnaires. I. Comparability of primary respondents and their next-of-kin. Am J Epidemiol 1985;121:937–47.
11. Herrmann N. Retrospective information from questionnaires. II. Intrarater reliability and comparison of questionnaire types. Am J Epidemiol 1985;121:948–53.
12. Byers T, Marshall J, Fiedler R, et al. Assessing nutrient intake with an abbreviated dietary interview. Am J Epidemiol 1985;122:41–50.
13. Lerchen ML, Samet JM. An assessment of the validity of questionnaire responses provided by a surviving spouse. Am J Epidemiol 1986;123: 481–9.
14. van Staveren WA, West CE, Hoffmans MDAF, et al. Comparison of contemporaneous and retrospective estimates of food consumption made by a dietary history method. Am J Epidemiol 1986; 123:884–93.
15. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley and Sons, 1981.
16. Fisher RA. Statistical methods for research workers. 14th ed. New York: Hafner, 1973.
17. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown and Co., 1985.
18. Bland JM, Altman DG. Statistical methods for measuring agreement between two methods of clinical measurement. Lancet 1986;1·307–10.
19. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–20.
20. Fleiss JL, Cohen J. The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability Educ Psychol Meas 1973;33:613–19.
21. Willett WC, Sampson L, Stampfer MJ, et al. Reproducibility and validity of a semiquantitative food frequency questionnaire. Am J Epidemiol 1985;122:51–65.
22. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair

studies. Int J Epidemiol 1983;12:93-7.

23. Walker AM, Blettner M. Comparing imperfect measures of exposure. Am J Epidemiol 1985; 121:783-90.
24. Walter SD. Measurıng the reliability of clinical data: the case for using three observers. Rev Epidemiol Sante Publique 1984;32:206-11
25. Rawler J. Regression analysıs. (Letter.) Lancet 1986;1:614.
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrıcs 1977;33:159-74.