Modeling Approaches for the Analysis of Observer Agreement

JOHN S. UEBERSAX, PhD

PRACTICAL ISSUES CONCERNING the use of the kappa statistic to measure observer agreement were recently reviewed.[1] Given the many potential problems with kappa, one may wonder if there are better ways to analyze agreement data. In fact, there has been much recent work in this area. The kappa statistic and similar omnibus indices reduce all information about agreement to a single number. In contrast, recent authors have taken a modeling approach to agreement, which has many advantages. This article's purpose is to acquaint radiology and other medical researchers with new developments in agreement modeling. The subject's scope precludes detailed treatment here. The current goal is to provide an overview; more details are available in the primary sources cited.

After reviewing several modeling methods and noting their strengths and limitations, one method will be illustrated with actual data. Readers can use this example to consider the suitability of agreement modeling to their research.

General Considerations

We will consider here only agreement on ordered category ratings, which are common in radiology research. We will not discuss dichotomous or nonordered category ratings, although some of the methods described here can be adapted for their analysis. A more detailed review[2] includes discussion of dichotomous and nonordered category rating agreement.

The type of data we are concerned with occurs when N cases are independently diagnosed or classified by two or more procedures. We use the word "procedure" broadly. It may refer to different diagnostic techniques such as imaging modalities, or to different experts who review test results and classify cases accordingly. Data are summarized as a cross-classification table (Table 1), which shows the number of cases that receive each possible combination of ratings across the set of procedures.

There are three reasons for collecting and analyzing such data. The first reason is to *describe the amount of agreement* among different procedures. With this goal, a statistical method is preferred that not only assesses the amount of agreement but also helps identify the causes of disagreement.

Although simply quantifying agreement is sometimes the main function of such data, deeper concerns frequently motivate its collection. Often, what we really want to know is which procedure is most accurate. A second use of such data, then, is to evaluate diagnostic procedures in the absence of a definitive comparison classification. By comparing results of different procedures, one hopes to get a *general idea of the accuracy of each procedure*. A basic premise for this use of data is that if two procedures disagree, at least one must be wrong.

A third use of such data is to determine the *value of diagnosis based on joint use of two or more procedures*. In this case, the statistical approach should include a way to integrate the results of the procedures.

Approaches Considered

Table 2 summarizes recent contributions in the area. The grouping of approaches is somewhat arbitrary, but emphasizes important distinctions among them.

Latent trait and latent class models are commonly subsumed under the heading "latent structure analysis" (Lazarsfeld and Henry[3]). One also could view loglinear, association, and quasisymmetry models as forms of latent structure analysis. There is a good analogy between latent

TABLE 1. Cross-classification Table Summarizing the Results of Three Tests* for Liver Metastases†

Rating pattern (i)	Test 1	Test 2	Test 3	Observed frequency	Expected frequency‡
1	1	1	1	36	35.1
2	1	1	2	22	22.4
3	1	1	3	0	0.8
4	1	2	1	26	26.4
5	1	2	2	22	20.1
6	1	2	3	0	1.2
7	1	3	1	3	1.5
8	1	3	2	0	2.0
9	1	3	3	1	0.6
10	2	1	1	13	14.7
11	2	1	2	14	11.7
12	2	1	3	0	0.8
13	2	2	1	12	13.2
14	2	2	2	25	23.4
15	2	2	3	5	4.1
16	2	3	1	1	1.3
17	2	3	2	5	6.3
18	2	3	3	10	9.4
19	3	1	1	1	0.9
20	3	1	2	1	1.3
21	3	1	3	1	0.5
22	3	2	1	1	1.3
23	3	2	2	7	6.6
24	3	2	3	10	11.2
25	3	3	1	3	0.5
26	3	3	2	13	14.9
27	3	3	3	66	65.8

*1 = definitely negative result; 2 = marginal result; 3 = definitely positive result.
†Adapted from Henkelman et al.[5]
‡For Model M3 of Table 3.

TABLE 2. Recent Articles on Modeling Approaches to Agreement on Ordered Category Ratings

	References
Latent trait and signal-detection agreement models	Henkelman et al[5] Quinn[6] Uebersax and Grove[4]
Latent class agreement models	Clogg[9] Dawid and Skene[8] Uebersax[10]
Loglinear, association, and quasisymmetry agreement models	Agresti[12] Agresti and Lang[19] Becker[13,14] Darroch and McCloud[18] Tanner and Young[11,16]

"mild form," and "severe form." A rating of "mild form" would be made if a case's trait level falls above the procedure's threshold for "mild form" and below the threshold for "severe form."

This scheme elucidates three potential sources of disagreement. First, ratings may disagree because one procedure's thresholds are systematically higher or lower than another's; in this case, we may think of the procedures as differently *biased*. Second, procedures may have different inter-threshold distances, which correspond to different *rating category definitions*. Third, a case's latent trait level is assumed subject to random *measurement error*. That is, due to various sources of "noise," the apparent trait level of a case, as registered by a procedure, may vary from its actual latent trait level; consequently, the same case, viewed by different procedures, may yield different apparent trait levels.

The parameters of the complete LTAA model include: 1) terms that characterize the distribution of the latent trait; 2) the thresholds for each procedure; and 3) measurement error parameters. The threshold parameters are used to estimate bias and rating category definitions for each procedure. Measurement error parameters are used to assess the precision of the procedures.

Figure 1 illustrates the basic model components. The abscissa corresponds to latent trait level, which we denote by x. The function $f(x)$ represents the probability density function of latent trait levels; t_{ij} is the threshold that must be exceeded to elicit from procedure i a rating level of category j or above. The function $p_{ij}(x)$ gives the probability of a case with each latent trait level x exceeding t_{ij}. For an actual model, there would be many t_{ij} values and corresponding $p_{ij}(x)$ functions.

The shape of $p_{ij}(x)$ depends on the model of measurement error. The ordinary assumption is that measurement error causes latent trait levels to appear normally distributed about their true levels. This assumption results in the ogive shape for $p_{ij}(x)$ shown in Figure 1.

It is important to note that measurement error applies here only in a relative sense. Error is not assessed with respect to "true" disease level, which would require a definitive com-

Latent Trait and Signal Detection Agreement Models

Uebersax and Grove[4] described methods for the analysis of agreement based on latent trait analysis. Henkelman et al[5] and Quinn[6] described similar approaches based on signal detection theory. The latent trait and signal detection agreement models are essentially identical and best viewed as a unified approach. We use the term latent trait agreement analysis (LTAA) to refer to both approaches.

The LTAA model assumes that a continuous latent trait underlies ratings—for example, the latent trait may be disease severity or symptom salience. The model assumes that each diagnostic procedure has thresholds along this continuum that correspond to various rating levels. For example, possible ratings by a procedure may be "no disease,"

structure analysis and computed tomography. The idea is that observed ratings (manifest variables) are due to unobserved factors (latent variables). The first step of latent structure analysis is to posit a plausible model that relates observed ratings to latent variables. One then estimates the values of the latent variables that are most likely, given the observed data. Model fit is statistically examined by comparing the observed cross-classification frequencies with those predicted by the model.
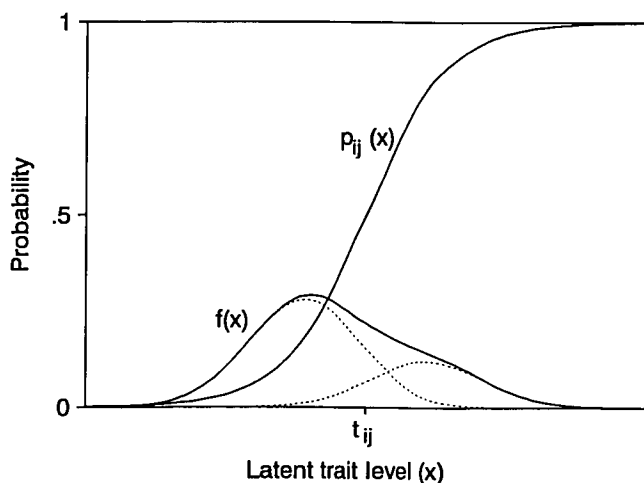
Fig. 1. Components of the latent trait agreement model. Latent trait distribution f(x) results from a mixture of disease-negative (left dotted line) and disease-positive (right dotted line) cases; $p_{ij}(x)$ is the probability that a case with latent trait level x exceeds diagnostic threshold t_{ij}.

parison criterion or "gold standard." Instead, error pertains to a procedure's ability to measure the thing that all the procedures measure in common. If we were dealing with continuous ratings, error would be the difference between one procedure's rating and the average of all procedures' ratings for the same case (assuming the ratings are comparably scaled). The analogy applies here, though the mechanics differ, because we are dealing with ordered category ratings. Consequently, if all procedures are inaccurate but highly correlated, true measurement error will be underestimated.

The LTAA approach has several advantages. One is that it is based on a familiar and plausible diagnostic model. Another is that the approach quantifies the factors of bias, category definitions, and measurement error that contribute to disagreement. Some LTAA models also permit inferences about rating sensitivity, specificity, positive and negative predictive validity, and the area under the ROC curve.[4,5] These rating validity inferences require special assumptions about the latent trait distribution; typically, one assumes that the case population consists of two types—disease-negative and disease-positive cases—and specifies the form of the latent trait distribution for each type. Another advantage of the LTAA approach is that it provides for the combination of ratings across procedures to yield a single score or measure of the latent trait.

There are potential limitations of the approach as well.[7] One is that the method restricts the latent trait distribution to some easily parameterized form (eg, a normal distribution or mixture of normal distributions). Another is that the procedure can be computationally demanding. Estimation algorithms require iterative evaluation of integrals. With many procedures and rating categories, calculations may

require several minutes using a desktop computer. Of course, this is more an inconvenience than a limitation, and as computers and algorithms become faster, its importance will lessen.

A factor complicating use of the LTAA approach is that the number of procedures and rating levels affects whether all model parameters can be uniquely estimated, that is, under certain conditions, different combinations of parameter values may "explain" observed data equally well. This requires an investigator to attend closely to the analysis, but poses few serious obstacles. When all parameters cannot be uniquely identified, it is usually possible to apply plausible constraints to some parameters that enable the remaining ones to be estimated.

The models of Uebersax and Grove[4] and Quinn[6] assume a single latent trait dimension. This is equivalent to the assumption that there is only one basic continuum of disease or symptom intensity to which procedures are sensitive. The Henkelman et al[5] model allows multiple latent trait dimensions; however, this increases computational complexity. A unidimensional model should be tried first. Experience suggests that a unidimensional latent trait model is often satisfactory.

Latent Class Agreement Models

Latent class agreement models assume that there is a set of case subtypes, or latent classes, such that cases within each latent class are essentially identical with regard to disease severity or symptom salience. For example, latent classes might correspond to different clinical stages or manifestations of a disorder. Latent classes also may represent rough gradations of disease severity.

Dawid and Skene[8] considered a general latent class agreement model for categorical ratings. Their model does not fully account for the ordered category nature of ratings such as we consider here, and may require estimation of more parameters than are necessary. Latent class agreement models for ordered category ratings are discussed by Clogg[9] and Uebersax.[10] The approach of Uebersax, termed "located latent class agreement analysis" (LLCAA) is the most theoretically based of these models.

The LLCAA model is essentially a discrete approximation of the LTAA model. The LTAA model portrays the latent trait distribution as a smooth function (eg, a mixture of normal distributions), whereas the LLCAA model portrays it as more of a histogram, with each latent class corresponding to a discrete location on the latent trait continuum and accounting for a certain proportion of cases. With enough latent classes, the LLCAA model can approximate the LTAA model to any degree of precision. In practice, though, a few latent classes often suffice to represent data well. The advantages of the LLCAA model relative to the LTAA model are that 1) it substantially reduces computational complexity, and 2) it does not restrict the latent trait

distribution to an arbitrary form such as a normal distribution or mixture of two normals.

Aside from how they represent the latent trait distribution, the LLCAA and LTAA models are equivalent. Bias, category definitions, and measurement error may be quantified with the LLCAA approach just as with the LTAA approach. The LLCAA model has not been used to estimate rating accuracy indices such as sensitivity, specificity, and area under the ROC curve, but it could be adapted for this purpose. The LLCAA model is illustrated in the example.

Loglinear, Association, and Quasisymmetry Agreement Models

Loglinear, association, and quasisymmetry agreement models are a related set of approaches that derive from standard techniques for categorical data analysis. Tanner and Young[11] discussed general loglinear agreement models; these models are better suited for dichotomous and nonordered category ratings than for ordered category ratings. For ordered category rating agreement, Agresti[12] and Becker[13,14] recommended an approach based on Goodman's[15] association model; Tanner and Young[16] discussed a similar model for ordered category agreement.

Association agreement models have so far only considered agreement between two procedures, but this is not an inherent limitation of the association model.[17] The model is easily understood. Let π_{ij} denote the probability that a randomly selected case receives rating level i by the first of two procedures and rating level j by the second procedure. The goal is to decompose $\ln(\pi_{ij})$ into specific components—this is true of loglinear modeling in general, of which the association model is a special case. With the association model, the components are: 1) the tendency of the first procedure to make rating level i; 2) the tendency of the second procedure to make rating level j; 3) for each (i,j), a term that reflects the confusability of the two rating categories; and 4) when i = j, an additional term that reflects the tendency of the procedures to apply the same rating category.

The confusability terms (component 3 above) are constructed from additional parameters. Specifically, for each procedure, each rating category is viewed as located on a continuum—the category's location is termed its scale score. The confusability of two categories is a function of the closeness of their scale scores. For example, the scale scores of four rating levels might be (1, 2, 3, 4) for the first procedure and (1.1, 2.7, 3.5, 4.2) for the second procedure. One might then expect many cases to be assigned rating level 3 by the first procedure and rating level 2 by the second procedure, because their respective scale scores for these categories (3 and 2.7) are relatively close. Scale scores are analogous to the thresholds of the LTAA and LLCAA models, but function differently. With the latter, category definitions correspond to the intervals between adjacent thresholds. With the association model, the scale scores themselves correspond to the rating categories.

The association agreement model has several advantages. It represents rating bias with the parameters for components 1) and 2) above. Estimated scale scores can be graphically portrayed and provide a convenient way to examine category definitions and how they may differ across procedures. The model also is computationally simpler and faster than the LTAA and LLCAA approaches.

In part, this simplification is possible because the association agreement model does not consider case latent trait levels. Because of this, it does not permit diagnosis based on multiple ratings or inferences about sensitivity, specificity, and other indices of rating accuracy. Another consideration is that—unlike the LTAA and LLCAA models, which attempt to represent the process of how ratings are made—the association agreement model is not theoretically based.

The quasisymmetry model for rating agreement (Darroch and McCloud[18]) is similar to the association agreement model. The model permits consideration of more than two procedures, but, for simplicity, we will suppose that agreement between only two procedures is considered. The quasisymmetry agreement model is again a specialized loglinear model, and the goal is to separate $\ln(\pi_{ij})$ into meaningful components. The components for the quasisymmetry model are the same as components 1), 2), and 3) above for the association agreement model—specifically, the tendencies of the two procedures to apply the various rating categories and the confusabilities of categories. The emphasis is on describing rating bias and category confusability.

Unlike the association agreement model, the quasisymmetry agreement model does not model category confusabilities in terms of more basic parameters or derive underlying category scale scores. Thus, the quasisymmetry agreement model does not, as with the association agreement model (and the LTAA and LLCAA models), provide a graphic representation of categories' definitions. Still, it is helpful to know which rating categories are most confused, which may suggest changes in rating nomenclature. Another difference is that the quasisymmetry model implicitly assumes that category definitions are the same for each procedure; with the association model, this is an optional assumption.

The quasisymmetry agreement model is based on a multiplicative signal detection model which is of some interest in its own right. It assumes that each case has a general tendency to elicit each rating category, and that each procedure has a general tendency to use each rating category. A given rating is assumed to be jointly determined by both factors. The model does not, however, estimate the former tendencies.

Intriguing connections among association, quasisymmetry, latent trait, and latent class models have been noted.

Agresti and Lang,[19] for example, showed how to include latent classes in the quasisymmetry agreement model. In some instances, association and quasisymmetry models produce identical results. Also, with dichotomous ratings, the quasisymmetry agreement model may produce results equivalent to LTAA and LLCAA models. These examples suggest that there are other mathematical connections among the models that have not yet been identified.

Example

This section illustrates one technique discussed above. The LLCAA approach is selected because it is particularly flexible and comprehensive. However, the example also illustrates the agreement modeling approach more generally.

We consider data presented by Henkelman et al[5] on three imaging techniques for diagnosis of liver metastases. The techniques are magnetic resonance imaging, computed tomography, and radionuclide scintigraphy; for convenience, we term these tests 1, 2, and 3, respectively. Table 1 shows the results of application of all three tests to each of 298 cases. Henkelman et al expressed each test's results on a scale with five gradations ranging from "definite negative result" to "definite positive result." To simplify, we collapse the three middle categories and express results on a three-level scale with categories "definite negative result," "marginal result," and "definite positive result." Column 5 of Table 1 shows the number of cases with each possible combination of ratings on the three tests.

Table 3 shows the results of several LLCAA models applied to the data. Model fit is assessed with the X^2 (Pearson chi-square) and G^2 (likelihood ratio chi-square) statistics. The X^2 statistic is calculated as $\Sigma_i(f_i - e_i)^2/e_i$, and the G^2 statistic is calculated as $2 \Sigma_i f_i \ln(f_i/e_i)$, where f_i is the observed frequency of the ith rating combination, e_i is the expected frequency of the ith rating combination given parameter estimates, and summation is across all I possible rating combinations (for these data, $I = 3 \times 3 \times 3 = 27$). The number of degrees of freedom is equal to $I - 1 - M$, where M is the number of independent estimated model parameters.

Good model fit is indicated when the chi-square statistics are close in value to their associated degrees of freedom.

One also typically evaluates the statistical significance of the X^2 and G^2 statistics. Probability values of approximately .1 or above generally indicate good fit.

Model M1 is a one-class model; it assumes case homogeneity with respect to the latent trait, which we view as the severity or salience of liver metastases. Model M1 also assumes that all tests have equal measurement error. This is the simplest model possible for the data and corresponds to the assumption of statistical independence of ratings. Because the model is implausible, its poor fit is reassuring.

Model M2 assumes that there are two latent classes of cases and two associated latent trait levels. Although fit is improved, the X^2 and G^2 statistics still show significant lack of fit. Model M3 assumes three latent classes and improves fit to a statistically acceptable level. It is important to keep in mind that the latent classes serve mainly a heuristic and pragmatic function. In reality, we would suppose the latent trait to be continuous. However, the fit of model M3 indicates that the presumably continuous distribution can be satisfactorily approximated by a three-class discrete distribution.

Table 4 and Figure 2 summarize the parameter estimates of model M3. The estimates in Table 4 (left) characterize the latent trait distribution. Results are arbitrarily scaled so that the lowest and highest latent classes have latent trait levels of -3 and 3, respectively. Most cases occupy the two extreme latent classes, which correspond to strongly negative and strongly positive cases, but 15.5% of cases fall in a third, ambiguous class.

Table 4 (right) summarizes measurement error for the three tests. Measurement error (or, strictly speaking, its absence) is expressed as the estimated correlation between true latent trait level and apparent trait level as registered by a procedure. The estimated correlation shows how much ratings are determined by signal rather than noise, and provides an index of rating precision. These correlations ordinarily range in value from 1 (perfect correlation; no noise) to 0 (no correlation; all noise). The results indicate that the three procedures are approximately equally precise.

Figure 2 plots the threshold estimates for model M3. For each test, t_2 marks the location of the threshold for the second rating level ("marginal result") and t_3 marks the location of the threshold for the third rating level ("definite positive result"). Some variability in threshold locations

TABLE 3. Results of Some LLCAA Models Applied to Data in Table 1

| Model | Description | Model fit | | df |
		X^{2*}	G^{2}†	
M1	1 class + EME	503.90	406.90	20
M2	2 classes	33.21	30.54	16
M3	3 classes	21.07	18.95	14

EME: equal measurement error across tests.
*Pearson chi-square.
†Likelihood ratio chi-square.

TABLE 4. Parameter Estimates for Model M3 of Table 3 Applied to Data in Table 1

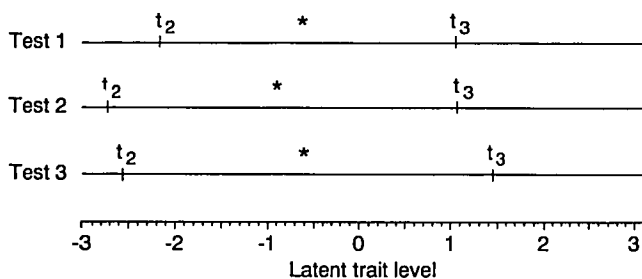| Latent trait distribution | | | Measurement error | |
Latent class	Trait level	Prevalence	Test	Correlation with latent trait
1	−3.000	0.489	1	0.851
2	−0.584	0.155	2	0.831
3	3.000	0.356	3	0.847

Fig. 2. Estimated threshold locations for Model M3 of Table 3 applied to data in Table 1.

across tests is evident. For example, test 1 has a higher threshold for the second rating category than the other two tests.

An asterisk marks the location of the mean threshold for each test, which provides an index of overall bias. Figure 2 shows that the tests are fairly similar in bias. Differences in threshold locations are often much more extreme than with these data.

After a well-fitting model is obtained, one often tests variations of the model. This permits statistical assessment of the importance of specific effects. For brevity, we omit such analyses here but can summarize the results of their application to these data. A simple variation of model M3 requires that measurement error be equal across tests. The fit of this model is not significantly worse than model M3. This is consistent with the results in Table 4 (right), but statistically verifies the observation that measurement error does not significantly differ across tests. Other submodel comparisons show that the tests do not differ significantly in terms of overall bias, but that differences in category definitions are statistically significant.

Finally, the model can be used to integrate the results of the three tests. One way to do so is to determine the probability of a case's membership in each latent class based on its ratings—this determination is made from model parameters with use of Bayes' rule. A case may then be assigned to the latent class to which it has the highest probability of membership. An alternative is to use ratings to calculate latent trait scores.[10] These scores are continuous measures that correspond roughly to the case's location on the latent trait continuum. Details are omitted here, but the method is very simple.

The conclusions provided by the analysis are as follows. First, from the poor fit of model M1, we concluded that the tests' ratings are not statistically independent. However, model M3, with three basic case types and three associated latent trait levels, represented the data well. The majority of cases appeared to fall at the extremes of the latent trait continuum. The fit of model M3 also affirmed the tenability of the assumptions of the LLCAA model, namely a unidimensional latent trait and a threshold model of diagnosis, for these data. The tests appeared to be approximately

equally precise and equally biased, but somewhat different in rating category definitions. If these were human observers, the results would suggest how to direct efforts to improve agreement. Specifically, the results indicate that efforts should aim to make category definitions more similar and to reduce measurement error overall.

Conclusions

This article reviewed several statistical modeling approaches for the analysis of rating agreement. Use of one approach was illustrated with actual data.

The modeling approach is quite flexible and can be readily adapted to the needs of a particular application. There are many specialized uses of agreement modeling that we have not considered. For example, one can use model parameters to estimate the relative precision of diagnosis based on different test combinations or sequences. In the future, we will likely see refinements of these methods. However, they are already sufficiently well developed to provide useful tools for analyzing agreement.

References

1. Ker M. Issues in the use of kappa. Invest Radiol 1991;26:78–83.
2. Uebersax JS. Quantitative methods for the analysis of observer agreement: toward a unifying model. Paper P-7686. Santa Monica, CA: The RAND Corporation; 1991a.
3. Lazarsfeld PF, Henry NW. Latent structure analysis. New York, NY: Houghton-Mifflin; 1968.
4. Uebersax JS, Grove WM. A latent trait finite mixture model for the analysis of rating agreement. Biometrics (in press).
5. Henkelman RM, Kay I, Bronskill MJ. Receiver operator characteristic (ROC) analysis without truth. Med Decis Making 1990;10:24–29.
6. Quinn MF. Relation of observer agreement to accuracy according to a two-receiver signal detection model of diagnosis. Med Decis Making 1989;9:196–206.
7. Begg CB, Metz CE. Consensus diagnoses and "gold standards." Med Decis Making 1990;10(1):29–30.
8. Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. Appl Stat 1979;28:20–28.
9. Clogg CC. Some latent structure models for the analysis of Likert-type data. Soc Sci Res 1979;8:287–301.
10. Uebersax JS. Statistical modeling of expert ratings on medical treatment appropriateness. J Am Statist Assoc (in press).
11. Tanner MA, Young MA. Modeling agreement among raters. J Am Stat Assoc 1985;80:175–180.
12. Agresti A. A model for agreement between ratings on an ordinal scale. Biometrics 1988;44:539–548.
13. Becker MP. Using association models to analyse agreement data: two examples. Stat Med 1989;8:1199–1207.
14. Becker MP. Quasisymmetric models for the analysis of square contingency tables. J Roy Stat Soc B 1990;52:369–378.
15. Goodman LA. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables (with discussion). Int Stat Rev 1986; 54:243–309.
16. Tanner MA, Young MA. Modeling ordinal scale disagreement. Psychol Bull 1985;98:408–415.
17. Sobel ME. Some models for the multiway contingency table with a one-to-one correspondence among categories. In: Clogg C, ed. Sociological methodology 1988. Washington, DC: American Sociological Association; 165–192.
18. Darroch JN, McCloud PI. Category distinguishability and observer agreement. Austr J Stat 1986;28:371–388.
19. Agresti A, Lang JB. Quasi-symmetric latent class models, with application to rater agreement. Biometrics (in press).