A RAND NOTE

Validity Inferences From Interobserver Agreement

John S. Uebersax

March 1989

RAND

This publication was supported by The RAND Corporation as part of its program of public service.

Published by The RAND Corporation
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138

A RAND NOTE

N-2927-RC

Validity Inferences From Interobserver Agreement

John S. Uebersax

March 1989

RAND

Validity Inferences From Interobserver Agreement

John S. Uebersax
Behavioral Sciences Department, The RAND Corporation

Methods for measuring rater agreement and making inferences about the accuracy of dichotomous ratings from agreement data are described. The first section presents a probability model related to latent class analysis that is applicable when ratings are based on a discrete trait. The second section extends these methods to situations in which ratings are based on a continuous trait, using a model related to signal detection theory and item response theory. The values obtained by these methods provide either direct or upper-bounds estimates of rating accuracy, depending upon the nature of the rating process. Formulas are shown for combining the opinions of multiple raters to classify cases with greater accuracy than simple majority or unanimous opinion decision rules allow. Additional technical refinements of the probability modeling approach are possible, and it promises to lead to many improvements in the ways that ratings by multiple raters are analyzed and used.

Classifications based on dichotomous observer or expert ratings are used routinely in psychological research. For example, an investigator may observe episodes of children's behavior and classify them according to whether or not they constitute play, a clinician may rate whether a particular trait or symptom is present, or an experimental psychologist may wish to classify learning strategies according to some typology. Common to all of these is the need to evaluate the quality of observer ratings. Typically, the assessment of rating quality takes one of two forms. If some criterion measure is available, the accuracy of ratings can be assessed directly by comparing observers' ratings to cases' true states as determined by the criterion; the accuracy of ratings thus measured, or their tendency to indicate the true state of cases, is also commonly referred to as their *validity.* Another common strategy for assessing the quality of ratings is to compare those of one rater with those of another. Thus, for example, a researcher may obtain independent ratings from two or more raters for each of a set of cases and determine the extent to which ratings of the same case agree in order to measure what is commonly referred to as their *reliability.*

While there is some degree of general agreement about what statistical methods should be used to measure rating validity, there is much more debate concerning the methods that should be used to measure reliability. Cohen (1960) proposed the kappa coefficient as a measure of interrater agreement, arguing that it had certain advantages over simply considering the proportion of times raters agree. Many limitations of the kappa coefficient have been pointed out, however. Maxwell (1977), for example, noted the arbitrary method by which it attempts to correct levels of observed agreement for an amount attributable to chance. More recently, Carey and Gottesman (1978), Grove, Andreason, McDonald-Scott, et al. (1981), Spitznagel and Helzer (1985), and Uebersax (1987) observed that the same rating procedure may result in different, and potentially very disparate, values of kappa when the proportions of cases belonging to various categories vary from population to population. A consequence of this is that it may be difficult to compare results across studies or to generalize from the results of a single study. Several of these authors also noted the desirability of measuring agreement on presence and absence of a trait separately, which the kappa coefficient does not do.

Other statistical indices proposed to measure observer agreement include the random error coefficient (Janes, 1979; Maxwell, 1977; Zwick, 1988), Yule's Y (Spitznagel & Helzer, 1985), and the log-odds ratio (Sprott & Vogel-Sprott, 1987). Many of the criticisms raised in connection with the kappa coefficient apply in varying degrees to these also.

At the same time, a more fundamental question about the relationship between rating validity and rating agreement has not been resolved. There is often a tendency to view validity and reliability as separate and unrelated issues. It is clear, though, that they must be related in some way. We know, for example, that the correspondence of ratings to an external criterion is limited by their reliability (Shrout, Spitzer, & Fleiss, 1987). There has been recent interest in the possibility of making direct inferences about the accuracy of individual ratings on the basis of observed levels of rater agreement. Several articles have discussed methods for making such inferences. However, possible limitations of these methods, in particular, the circumstances under which they provide information about rating validity as opposed to simply expressing rater agreement in a different, though perhaps very useful way, require greater clarification.

The purpose of this article is threefold. First, it reviews the conceptual basis for methods proposed to make inferences concerning rating validity from agreement data. Second, restrictions on the applicability of these methods, most importantly, that the trait on which ratings are based must be viewed as discrete, are noted. Methods extending the approach to situations

I thank the following people for comments and information helpful in the preparation of this article: Joseph Fleiss, William Grove, Shelby Haberman, Elizabeth Lewis, Hubert Schouten, Stephen Walter, and two anonymous referees.

Correspondence concerning this article should be addressed to John S. Uebersax, Behavioral Sciences Department, The RAND Corporation, 1700 Main Street, Santa Monica, California 90406-2138.

Reprinted from *Psychological Bulletin*, Vol. 104, No. 3, November 1988, pp. 405–416,

in which the trait is more accurately viewed as continuous are then presented. Third, the relationship between the information provided by these methods and rating validity as it is more traditionally understood is discussed.

Discrete Case

Conceptual Basis

The nature of the problem of making inferences about rating validity from observer agreement can be illustrated by means of a hypothetical example. Consider the following problem:

> An urn contains a large number of black and white marbles. There are also two bowls, each containing a large number of black and white stones, the proportions of black and white stones in the two bowls being different. A two-stage procedure is defined as follows: First, a marble is drawn from the urn; then, if the marble is black, three stones are randomly drawn from the first bowl, but if the marble is white, three stones are drawn from the second bowl. You are informed neither of the color of the marble drawn nor from which of the bowls the stones are taken. You are told only the numbers of black and white stones drawn each time, for example, "two blacks and a white." Given that this procedure is repeated many times, you are to determine (a) the proportion of black and white marbles in the urn, and (b) the proportion of black and white stones in each of the two bowls.

This problem can in fact be solved easily. To see this, consider what would happen if there were only one bowl instead of two. In that case, the frequency of outcomes involving various numbers of black and white stones would follow a binomial distribution. Because the shape of a binomial distribution is uniquely determined by the probabilities of the alternative outcomes, it is possible to infer these probabilities from an observed distribution, given a sufficient number of observations, or in terms of this example, to infer the proportion of black and white stones in the bowl from the results observed. The fact that the stones may be drawn from either of two bowls adds only slightly to the complexity of the problem. The observed pattern of results then represents the mixture of two binomial distributions. However, again the shape of the combined distribution is determined uniquely by the proportion of black and white stones in each of the two bowls and by the proportion of times that stones are selected from one or the other of them. Thus, given enough repetitions to accurately characterize the distribution, it is possible to estimate the desired values.

There is a direct analogy between this example and that of inferring the accuracy of a rating procedure, given multiple ratings for each of a sample of cases. The urn corresponds to a population of cases consisting of a certain proportion with and a certain proportion without some characteristic or trait of interest; the bowls and their contents represent differing probabilities of a case being rated as having or not having the trait, depending upon whether it does or does not have the trait; the three stones drawn correspond to the ratings of three independent observers of a particular case. Thus, inferring the contents of the urn and the bowls from the numbers of black and white stones observed over many trials is analogous to estimating the prevalence of cases in a population with and without a trait, and the probabilities of positive and negative ratings, given either type of case, on the basis of several ratings of each of a sample of cases. Because these probabilities are the same as the probabilities of correct or incorrect ratings, this is equivalent to inferring the accuracy or validity of ratings on the basis of observer agreement.

Several authors have discussed the application of probability models related to this subject. Kaye (1980) and Kraemer (1982) considered the possibility of directly estimating rater accuracy on the basis of ratings by two or more raters. Using related methods, Hui and Walter (1980) showed that given rating procedures with different levels of accuracy and two samples drawn from populations with different trait prevalences, it is possible to estimate the accuracy of the two procedures and the prevalence of the trait in both populations. Gelfand and Solomon (1974, 1975) have also discussed the use of related models in the analysis of votes by jury members in criminal and civil trials. Though not directly identified as such, the models developed in those articles are closely related to the set of techniques generally known as latent class analysis (Goodman, 1974; Lazarsfeld & Henry, 1968). More formal discussions of the application of latent class analysis to the study of observer agreement have been made by Dawid and Skene (1979), Dillon and Mulani (1984), and Walter (1984).

Kaye (1980), Kraemer (1982), and Gelfand and Solomon (1974, 1975) differ somewhat from Dawid and Skene (1979), Dillon and Mulani (1984), and Walter (1984) in terms of the type of rating design considered. The approach of the former authors was more consistent with an interpretation of ratings being made by panels of raters selected for each case separately by some random process. Thus, the raters rating one case may not necessarily be the same as those rating another. In contrast, the rating paradigm considered by the latter authors corresponds more closely to a fully crossed design in which a fixed panel of raters evaluates each case, so that each case is observed by each rater. An advantage of the latter approach is that it makes it possible to consider the level of rating accuracy associated with each individual rater. However, in order to do this, it is necessary that it be possible to associate with each rating of a particular case the identity of the observer making the rating, information that may not always be available to the investigator.

We shall consider here the first type of rating paradigm, that is, varying rater panels, in part because it involves fewer parameters and is therefore somewhat simpler but also because it resembles more closely the paradigm to be considered in conjunction with the continuous model in the next section. The reader should note, however, that with the application of appropriate constraints, the general latent class model can be adapted to designs in which raters vary from case to case and will yield results identical to those of the methods described here. A discussion of this is to be found in Dillon and Mulani (1984).

Latent Class Analysis

Latent class analysis has been used in many diverse applications in psychological research (for recent examples, see Dillon, Madden, & Kumar, 1983; Young, 1983). No satisfactory description of the theory, methods, or extensive literature concerning latent class analysis can be made here, and for this the reader is referred to Lazarsfeld and Henry (1968) and Goodman (1974). For present purposes, it suffices to note that the

basic premise of latent class analysis is that an observed pattern of results, for example, responses or scores of cases with respect to a set of categorical variables, derives from the cases' membership in two or more latent categories. From the observed data, it is possible to apply mathematical procedures to estimate the prevalence of these latent categories, the extent to which each variable indicates membership in one category or another, and the probable membership of each case.

We shall suppose that cases are rated on the same trait by several observers. Specifically, let us assume that the data correspond, as in the marble example, to a two-stage Bernoulli process. First, N cases are assumed to be randomly drawn from a population whose members belong to one of c latent categories. In general, we shall suppose these categories to be two in number, one with a higher probability of eliciting a positive rating and corresponding, though not necessarily exactly, to presence of the trait and the other, with a lower probability of eliciting a positive rating, corresponding to absence of the trait; we shall refer to these as the *positive* and *negative* latent categories, respectively. In the formal development of the probability model, however, we shall allow the number of latent categories to vary, to allow for instances in which it may be useful to consider more than two. The observer ratings themselves, though, will be considered dichotomous. Second, for each case, a panel of k observers is assumed to be selected by a random process from a population of potential raters, who independently rate the case. We note, also, that this need not be a panel in the literal sense of the term; alternatively, multiple ratings of each case may come, for example, from replications of the same rating procedure.

Two assumptions are made in conjunction with the discrete model considered here. The first is that the probability of a positive rating, given a case belonging to a particular latent category, is the same for each observation of the case. This is not the same as requiring that all raters have the same probabilities of making a positive rating for a given latent class. Probabilities of individual raters may vary; however, given that they do, it is essential that random sampling of raters be used to assure that any given rating of a case has the same probability of being positive.

The second assumption, which is common to all forms of latent class analysis, is that of conditional independence. This asserts that within each latent category, the probability of one rating being positive is statistically independent of another being positive. This is equivalent to stating that cases within each latent category are identical with respect to their probability of eliciting a positive rating and is essentially a restatement of the definition of latent categories. The discrete model of observer agreement thus requires that the probabilities of eliciting positive ratings for all cases take only a finite, and presumably relatively small, number of values. This may be a reasonable assumption in some applications, as, for example, when the trait considered is truly dichotomous. However, in other situations the trait may be more accurately regarded as continuous, with corresponding continuously varying values in the probability of eliciting a positive rating. This will be considered in the next section.

Probability Model

To begin, let the ith rating of a case be denoted by u_i, a value of 1 being assigned if the rating is positive and a value of 0 if the rating is negative. Let X be a discrete variable indicating one of c latent categories to which a case may belong. We define the probability of the ith rating of a case being positive, given that the case is a member of latent category s ($s = 1, \cdots, c$), as

$$P\{u_i = 1 | X = s\} = p_s. \tag{1}$$

We shall assume in general that the categories are numbered such that the first has the lowest probability of eliciting a positive rating and that the last is associated with the highest probability, that is, $p_1 < p_2 < \cdots < p_c$ or, in the two-category case, simply that $p_1 < p_2$. Let us also denote the population prevalence of each latent category, s, by r_s, such that $\Sigma r_s = 1$. From this we easily derive a formula for the probability that, given k randomly selected ratings of a case, exactly j will be positive. Consider the hypothetical results of a case rated by five raters, represented by an arbitrary ordering as the vector 1, 0, 1, 1, 0. Given a case belonging to the first latent category, the probability of this response pattern occurring is $p_1(1 - p_1)p_1p_1(1 - p_1)$. We note, however, that there are 5!/(3! 2!) orderings containing exactly three positive ratings, each with the same probability of occurring. Thus we see that the probability of exactly j out of k ratings being positive, given a case belonging to latent class s, is

$$P\{u_\cdot = j | X = s\} = \binom{k}{j} p_s^j (1 - p_s)^{k-j}, \tag{2}$$

where $u_\cdot$ indicates the total number of positive ratings for that case. It follows that the probability of exactly j out of k positive ratings for a randomly selected case, regardless of latent category, is

$$P\{u_\cdot = j\} = \binom{k}{j} \sum_{s=1}^{c} r_s p_s^j (1 - p_s)^{k-j}. \tag{3}$$

This leads directly to a specification of the likelihood function for a set of results obtained in a multirater agreement study. Let the results of N cases rated by k raters each be represented by the vector $n_0, n_1, \cdots, n_k$, with n_j indicating the number of cases receiving exactly j out of k positive ratings. Given the independence assumptions noted in the preceding section, the likelihood of a particular pattern of observed frequencies, given a set of prevalence and rating probability parameters, is

$$L = \prod_{j=0}^{k} P\{u_\cdot = j\}^{n_j}. \tag{4}$$

Estimation

From observed results summarized in terms of the vector n_0, $n_1, \cdots, n_k$, the goal is to estimate the probabilities of positive ratings $p_1, p_2, \cdots, p_c$ and the prevalences $r_2, r_3, \cdots, r_c$ (r_1, or any other one of the prevalences, need not be estimated, because they must sum to 1) for each of the latent classes.

Iterative maximum likelihood procedures provide a general method for obtaining estimates of these parameters from a set of observed data. By definition, the maximum likelihood estimates of the latent parameters are those that maximize the value of Equation 4. To facilitate computation, we take the logarithm of Equation 4, resulting in

$$F = \sum_{j=0}^{k} n_j \log \sum_{s=1}^{c} r_s p_s^j (1 - p_s)^{k-j}, \quad (5)$$

as the function to be maximized, noting that the logarithm of the combination term need not be considered because it is constant. A convenient way of obtaining the parameter values that maximize this function is by means of the multivariate generalization of the Newton-Raphson algorithm (Lazarsfeld & Henry, 1968). By this iterative method, first derivatives of Equation 5 relative to each of the latent parameters and second partial derivatives are calculated for a given set of parameter estimates. From these, new estimates of the latent parameters, closer to the maximizing values, are calculated, derivatives are again computed, and the process is repeated until the solution converges, that is, until two successive sets of estimates result in a sufficiently small change in Equation 5. A useful feature of this method is that it leads directly to estimates of the asymptotic standard errors of the parameters. Specifically, these values are calculated following a method discussed by Haberman (1978, 1979), from the inverse of the information matrix, which is generated in the course of deriving iterative approximations.

A computer program implementing this algorithm in the case of two latent categories to estimate the parameter p_1, p_2, and r_2 has been developed (Uebersax, in press). Convergence is very rapid, usually taking two or three iterations. One of the factors affecting the efficiency of the algorithm is the initial estimates of p_1, p_2, and r_2 selected. Advantage can be taken of the fact that the results of $k \geq 4$ raters can be reformulated in terms of the expected results, given three ratings per case. Let n_j denote, as before, the number of times a case receives exactly j positive ratings, given k opinions. In addition, let n'_i ($i = 0, 1, \cdots, k'$) denote the expected results of sampling only k' ($k' < k$) ratings for each case. Then

$$n'_i = \sum_{j=0}^{k} \frac{\binom{k'}{i}\binom{k-k'}{j-i}}{\binom{k}{j}} n_j. \quad (6)$$

For the special case of $k = 3$ raters and $c = 2$ latent categories, Lazarsfeld and Henry (1968, pp. 27–31) presents an algebraic solution for the estimation of latent parameters. The values for n'_0, n'_1, n'_2, and n'_3 derived by Equation 6 can be used in conjunction with this method to provide initial estimates for p_1, p_2, and r_2, which then form the basis for a convergent maximum likelihood solution.

Parameter estimates may also be obtained by iterative proportional fitting procedures, such as those discussed by Goodman (1974).

Goodness of Fit

The correspondence of a model defined by a set of parameter estimates to observed results can be assessed either by means of the χ^2 goodness-of-fit statistic or a χ^2 test based on the likelihood ratio statistic. Let e_j ($j = 0, 1, \cdots, k$) be the vector of expected frequencies of cases with various numbers of positive ratings. These are obtained for a particular model by multiplying the number of cases rated, N, by the value of $P\{u_\cdot = j\}$ given in Equation 3 for each number of positive ratings, j. The χ^2 goodness-of-fit statistic is then equal to $\Sigma_j[(n_j - e_j)^2/e_j]$, the number of degrees of freedom for the test being equal to the number of possible values of j, or $k + 1$, minus 1 (since $\Sigma n_j = N$) minus the number of latent parameters estimated. Because the parameters estimated, given c latent categories, are c rating probability estimates and $c - 1$ prevalence estimates, the total number of parameters estimated is $2c - 1$. Thus the degrees of freedom for the goodness-of-fit test is $k - 2c + 1$ or, in the two-category case, $k - 3$. Alternatively, the value of χ^2 based on the likelihood ratio statistic, equal to $2\Sigma_j n_j \log (n_j/e_j)$, can be used to assess model fit, the degrees of freedom being the same as with the goodness-of-fit test.

Most of the cautions concerning the use of goodness-of-fit tests in general apply to this case as well. Because, unlike most applications of statistical tests, the goal of a goodness-of-fit test is to obtain results that are not significant, the typical procedure of setting a low value for α becomes nonconservative rather than conservative. A comparatively high value for α may therefore be more appropriate. Also, if the sample size is extremely large, then even a slight departure of the results expected for a particular model from the observed results may be statistically significant. Thus, in using these approaches to assessing model fit, a researcher must exercise judgment in interpreting results.

If a satisfactory fit is not obtained with a two-category model, it may be because the trait itself is not discrete rather than because it has a greater number of discrete levels, so that it may be useful to consider approaches such as that presented in the following section that view the trait as a continuous variable. Alternatively, particularly when theoretical considerations warrant it, improved fit may be achieved by considering discrete models with three or more latent categories. The maximum number of categories possible to consider is limited by the number of raters. Specifically, it is necessary that k be greater than or equal to $2c - 1$ to obtain estimates and greater than or equal to $2c$ if a χ^2 test is also to be used. Instances may occur in which two or more models may both yield a satisfactory fit with the data. In such cases, parsimony, given that there are no a priori theoretical considerations lending support to the higher-category model, would argue in favor of the model involving fewer categories. In general, parameter estimates associated with models with fewer categories will also have smaller standard errors than those associated with models involving more categories.

Example

Suppose that a researcher is studying recognition of facial expressions. Let data be gathered in which 600 photographs are rated according to whether the expression of the figure in the picture is perceived as smiling. Four subjects are selected randomly to rate each photograph, and the ratings are made according to the discrete model outlined in the previous section. Results are tabulated in the form of a vector of frequencies with which various numbers of positive ratings occur, specifically, $n_0 = 71$, $n_1 = 39$, $n_2 = 46$, $n_3 = 144$, and $n_4 = 300$.

From Equation 6, the expected frequencies for the results of three-rater panels are approximately $n'_0 = 81$, $n'_1 = 52$, $n'_2 =$

131, and $n'_3 = 336$, leading from Lazarsfeld and Henry's (1968) equations to initial estimates of $p_1 = .132$, $p_2 = .889$, and $r_2 = .796$. Using these as starting values, the iterative estimation algorithm results in values of $p_1 = .125$, $p_2 = .884$, and $r_2 = .803$, with standard errors of .0234, .0088, and .0189. Application of Equation 3 results in expected frequencies for cases receiving varying numbers of positive ratings of $e_0 = 69.5$, $e_1 = 42.2$, $e_2 = 38.7$, $e_3 = 154.9$, and $e_4 = 294.7$. In comparing these to the observed frequencies by means of the goodness-of-fit test, a value of χ^2 equal to 2.524 is obtained, which with $4 - 3$ or 1 *df* is nonsignificant at the .1 level. Thus, it could be concluded that the data are adequately represented by a model positing the existence of two latent categories, one with a higher, .884 probability of leading to a positive rating and one with a lower, .125 probability.

Relationship to Rating Accuracy

Before considering the relationship between the latent parameter estimates and rating accuracy, it will be helpful to review more generally methods for measuring accuracy. Several indices are used for this purpose. Two that are very common are sensitivity and specificity. Both may be defined in terms of conditional probabilities. *Sensitivity* (Se) is the conditional probability that a case will be rated as having a trait, given that the trait is present. Equivalently, it may be interpreted as the probability of detecting a positive case or the proportion of positive cases correctly rated. *Specificity* (Sp) measures the accuracy of negative ratings, being defined as the conditional probability that a case will be rated as not having the trait, given that the trait is absent. Thus, it may similarly be viewed as the probability of detecting a negative case or the proportion of negative cases correctly rated.

Two related indices are what are often termed the predictive value of positive and negative ratings. These represent the converse conditional probabilities of sensitivity and specificity. The predictive value of positive ratings is equal to the conditional probability of a case having the trait, given that it receives a positive rating, and the predictive value of negative ratings is the conditional probability of a case not having the trait, given that it is rated as not having the trait. As is evident from their definitions, these indices bear a strong relationship to one another. The application of Bayes' theorem shows, for example, that the sensitivity of ratings is equal to the probability of a positive rating being made, multiplied by the predictive value of a positive rating, divided by the prevalence of positive cases. In a corresponding way, rating specificity is equal to the probability of a negative rating multiplied by the predictive value of negative ratings, divided by the prevalence of negative cases.

An additional index occasionally used to express the validity of ratings is overall accuracy or percent correct, which is simply the proportion of cases, either positive or negative, that are correctly rated. The drawback of using overall accuracy as a validity index, however, is that it does not distinguish between the accuracy of positive and negative ratings, information that the researcher may find useful.

In the two-category case it is clear that there is a close connection between the probabilities of positive ratings, given membership in the latent categories, p_1 and p_2, and rating accuracy. Specifically, if the latent categories correspond exactly to presence and absence of the trait, p_2 may be taken as a direct estimate of rating sensitivity and $1 - p_1$ as an estimate of rating specificity. By application of Bayes' theorem, the predictive value of positive ratings is then equal to $r_2p_2/(r_1p_1 + r_2p_2)$, the predictive value of negative ratings is equal to $r_1(1 - p_1)/[r_1(1 - p_1) + r_2(1 - p_2)]$, and the percent of correct ratings is equal to $r_1(1 - p_1) + r_2p_2$.

Kaye (1980) and Walter (1984), in their discussions of validity inferences from agreement data, assumed this identity between latent classes and true categories. Again, this is not implausible in many cases. If there are no systematic factors causing raters to agree other than their mutual ability to detect presence or absence of a trait, that is, no sources of shared error, latent and true categories would be expected to correspond perfectly. Moreover, in many applications, particularly those connected with psychological research, what is often investigated is precisely the capacity of a stimulus to elicit a certain response in a rater. For example, the rating process considered may involve judgments concerning whether a sound is audible or whether an anecdote is funny. In such cases, a latent consensus of observers' judgments may be regarded as the criterion of interest.

In other instances, however, the latent classes may not correspond exactly to actual presence or absence of the trait. For example, raters may have some decision criterion or definition of the trait in common that corresponds imperfectly with true presence or absence of the trait, or some cases may have a related characteristic causing them to appear positive to all raters, hence belonging to the positive latent category, yet actually be negative. In this case, the relationship between p_1 and p_2 and rating accuracy depends on the probabilities of latent class membership, given true category membership. Specifically, rating sensitivity is equal to the probability of a positive rating, given a case belonging to the positive latent category (p_2), multiplied by the probability of a case belonging to the positive latent category, given that it is positive, plus the probability of a positive rating, given a member of the negative latent category (p_1), multiplied by the probability of membership in the negative latent class, given a positive case. Because the probabilities of positive and negative latent class membership, given a positive case, must sum to 1, and because $p_2 > p_1$, it follows that Se must be less than or equal to p_2. By similar reasoning, it is seen that $\text{Sp} \leq 1 - p_1$.

For $c > 2$ latent categories, explicit formulas can be derived giving the accuracy of individual ratings based on the estimated values of p_s and r_s. These require, however, that the investigator be able to specify on the basis of theoretical considerations or a priori knowledge the probability of a positive or negative case falling in each latent class or, equivalently, the proportions of cases in each latent class that are positive and negative. However, following reasoning similar to that in the two-category case, it may be shown that sensitivity is bounded by the largest value of p_s, that is, $\text{Se} \leq p_c$, and specificity by 1 minus the lowest value, that is, $\text{Sp} \leq 1 - p_1$.

The validity parameters obtained by means of latent class modeling, both in cases in which they directly estimate rating sensitivity and specificity and in cases in which they provide upper-bounds estimates, are useful in a variety of practical re-

search situations. For example, they may serve as a basis for estimating minimal sample sizes necessary to attain requisite statistical power for a comparison of mean differences between two groups, when group membership is determined on the basis of rater classifications. Although positive and negative cases may differ on a particular variable of interest, the effect of inaccurate ratings is to create groups that contain both positive and negative cases. The mean difference between groups as given by the rating procedure may therefore tend to underestimate the true difference between the groups. By using the estimates of sensitivity, specificity, and prevalence obtained by the methods shown, it is possible to estimate the effects of misclassification on observed mean differences and statistical power and to determine a suitable sample size to compensate for this effect.

Having noted more generally the factors that must be considered in deriving validity inferences from rating agreement, we shall now focus on the case of $c = 2$ latent classes and perfect correspondence between actual and latent categories.

Value of Multiple Opinions

An important feature of the methods discussed here is that they make it possible to attach a precise probabilistic meaning to the ratings of several observers. It follows from the initial independence assumptions that the probability of a case belonging to the positive latent category receiving k unanimous positive ratings is equal to p_2^k. Similarly, the probability of a case belonging to the negative latent category receiving k unanimous negative ratings is $(1 - p_1)^k$. By extension, the application of Bayes' theorem shows that the probability of a case belonging to the positive latent class, given k unanimous positive ratings, is

$$P\{X = 2|u_\cdot = k\} = \frac{r_2 p_2^k}{r_1 p_1^k + r_2 p_2^k}, \tag{7}$$

and the probability of a case belonging to the negative latent class, given k unanimous negative ratings, is

$$P\{X = 1|u_\cdot = 0\} = \frac{r_1(1 - p_1)^k}{r_1(1 - p_1)^k + r_2(1 - p_2)^k}. \tag{8}$$

Given that the latent categories derive only from trait presence and absence, these values are also interpretable as the predictive values of unanimous positive and negative ratings. More generally, consider a pattern of opinions consisting of j positive and $k - j$ negative ratings. The probability of positive class membership, given this pattern, is then

$$P\{X = 2|u_\cdot = j\} = \frac{r_2 p_2^j (1 - p_2)^{k-j}}{r_1 p_1^j (1 - p_1)^{k-j} + r_2 p_2^j (1 - p_2)^{k-j}}, \tag{9}$$

and the probability of negative class membership is 1 minus this amount. Again, because we have assumed a perfect correspondence between latent categories and trait states, these values also estimate the probability of a positive or a negative case, given the observed ratings.

It is easy to see many practical applications for these formulas. For example, a researcher may wish to integrate the information provided by multiple observers to derive an optimal classification for each case. Accordingly, by Equation 9 it is possible to determine on the basis of observed ratings whether a case is more probably a member of the positive or the negative latent category. If, as is commonly the case, equal misclassification costs are assumed, that is, if the cost associated with assigning a positive case to the negative category is equal to that of assigning a negative case to the positive category, the optimal classification would be to the category to which the case has the higher probability of membership. If misclassification costs are different, the product of the probability of a case belonging to each category and the cost associated with failing to identify a case belonging to that category should be calculated and the case assigned to the category for which this product is higher. For a discussion of this in a related context, see Macready and Dayton (1977).

Equations 7–9 may also be helpful in planning the number of ratings per case necessary to classify cases with a desired degree of accuracy. Although an investigator may feel that the opinion of one rater or expert is an insufficient criterion to classify a case for research or clinical purposes, there are often few guidelines for determining how many ratings per case would be more appropriate. These equations provide a formal basis for such decisions and make it possible to estimate the incremental predictive value of each additional observer. Of potential interest is that these equations also lend themselves to adaptive rating strategies, in which ratings are solicited from successive observers until a case can be assigned to one category or the other with a sufficient degree of accuracy.

Majority Opinion

Many authors have considered the use of majority opinions by a panel of raters in assigning a case to one category or another (see, e.g., Schouten, 1985, 1986). By simple extension, the formulas shown here can be applied to estimate the sensitivity, specificity, and predictive values of majority-based decisions. Let $u_\cdot \geq j'$ denote at least j' out of k positive ratings of a case, where j' is the minimal number constituting a majority. The probability of a majority positive opinion, given a member of the positive latent category, corresponding to the sensitivity of a majority positive opinion, is thus

$$P\{u_\cdot = j'|X = 2\} + \cdots + P\{u_\cdot = k|X = 2\},$$

or

$$P\{u_\cdot \geq j'|X = 2\} = \sum_{j=j'}^{k} \binom{k}{j} p_2^j (1 - p_2)^{k-j}. \tag{10}$$

Similarly, the probability of positive latent category membership, given a majority of positive ratings, corresponding to the predictive value of a positive classification based on majority decision, is

$$P\{X = 2 \,|\, u_\cdot \geq j'\} = \frac{r_2 \sum_{j=j'}^{k} P\{u_\cdot = j|X = 2\}}{\sum_{j=j'}^{k} P\{u_\cdot = j\}}, \tag{11}$$

where the component probabilities are given by Equations 2 and 3.

Agreement Due to Error

In the original article concerning the kappa coefficient, Cohen (1960) suggested that the proportion of pairs of rater opinions in agreement is misleading because it does not consider the role that chance may play in determining agreement. The exact nature of chance as it operates in rater decision making, and especially as it pertains to agreement measurement, however, has never been clearly formulated. Maxwell (1977) suggested that cases can be divided into two categories: those for which a rating can be made with absolute certainty and those for which the status of the case is not clear and the rater must guess. According to this view, it would be expected that two raters guessing on the latter type of case would agree a certain proportion of the time, that is, agreements would occur by chance. In practice, of course, it is unlikely that all cases are such that trait presence or trait absence is either perfectly apparent or completely unknown. It is unclear, in fact, whether guessing affects the rating process to a significant degree at all. One might instead suppose that raters generally make their decisions on the basis of the observed information and according to rules that they consider valid. Information concerning a case may be misleading or a rater's decision rules inaccurate, leading to a rating that is in error, but this is not the same as making a rating by chance.

It is of interest to consider the implications of the methods shown here for the problem of agreement in which both raters are incorrect. The expected proportion of pairs of ratings in agreement, given values of p_1, p_2, r_1, and r_2, can be expressed as

$$P\{u_i = u_j\} = r_1[p_1^2 + (1 - p_1)^2] + r_2[p_2^2 + (1 - p_2)^2]. \quad (12)$$

It is readily seen that agreements can be divided into two categories: those in which both raters assign a case to the correctly corresponding latent category, that is, rate a member of the positive latent class as positive or a member of the negative latent class as negative, or $r_1(1 - p_1)^2 + r_2p_2^2$, and those associated with mutual error, in which both raters assign a case to the incorrect latent category, or $r_1p_1^2 + r_2(1 - p_2)^2$. These may be divided into components reflecting mutual error or accuracy with regard to each separate category. Furthermore, these components may be made independent of prevalence and hence comparable across studies. Thus, it is seen that p_2^2 estimates the probability of a nonerror agreement and $(1 - p_2)^2$ the probability of an agreement due to error, given a member of the positive latent class, and that $(1 - p_1)^2$ and p_1^2 represent corresponding values, given a member of the negative latent class.

The indices thus derived provide an approach to measuring rater agreement that is advantageous in that (a) it incorporates a theoretically based model about the role of error in determining agreement and how it should be corrected for, (b) it expresses agreement on positive and negative ratings separately, and (c) agreement can be expressed in a way that is independent of prevalences, facilitating the comparison of studies.

Application to a Continuously Varying Trait

Though useful in developing a basic framework for approaching the problem of making inferences about rating accuracy from agreement data, limitations on the applicability of the preceding methods are posed by the assumption of discrete latent categories. Kraemer (1982) considered several sets of data and found the fit of a two-category discrete model to be low. Although, as Walter (1984) noted, better fit could be obtained by allowing for additional latent categories, it is important to consider whether it makes sense to model many rating processes in terms of discrete categories. In many applications it would be more accurate to view the trait on which ratings are based as continuous, although the ratings themselves may be dichotomous. Methods similar to those presented in the previous section can be developed for the continuous case. These methods, it will be noted, are related in many respects both to signal detection theory (Swets, 1973, 1986) and to item response theory (Hulin, Drasgow, & Parsons, 1983; Lord & Novick, 1968).

The basic components of the continuous model are illustrated in Figures 1 and 2. We begin by assuming the existence of a continuous latent trait denoted by the variable θ. Although we refer to this as a trait, the term is used broadly and may refer either to a single continuous trait (e.g., activity level) or to a more complex characteristic consisting of an aggregation of several traits (e.g., degree of schizophrenic symptomatology). In addition, it is assumed that there are two types of cases in the population, which we shall refer to as *positives* and *negatives*. Both are, at least initially, assumed to follow normal distributions with regard to θ. Cases are again sampled randomly, and each is independently evaluated by k randomly selected observers, who indicate whether they consider it to be positive or negative.

Let $g_1(\theta)$, with mean μ_1 and standard deviation σ_1, be a probability density function describing the distribution of negative cases at each level of θ, and let $g_2(\theta)$, with mean μ_2 and standard deviation σ_2, be the probability density function of positive cases with respect to θ. Scaling these by the prevalences of both types of cases in the population, $1 - P$ and P, respectively, we define the functions

$$f_1(\theta) = (1 - P)g_1(\theta)$$

and

$$f_2(\theta) = Pg_2(\theta).$$

By letting $f_3(\theta) = f_1(\theta) + f_2(\theta)$, it is seen that $f_3(\theta)$ is a probability density function giving the probability of a randomly selected case having any trait level θ (Table 1).

We next consider the role that a rating threshold plays in the decision process of a rater. Each rater is assumed to have a characteristic threshold, corresponding to a minimal trait level that a case must display in order to warrant a positive rating. The concept of a rating threshold is thus closely related to the idea of a cutting score in test theory, and many of the ideas familiar from that context apply here as well. For example, the proportion of the area under $f_2(\theta)$ falling to the right of a rater threshold corresponds to the true positive rate for that decision criterion, and the proportion to the left corresponds to the false negative rate. Similarly, the proportion of the area under $f_1(\theta)$ to the left

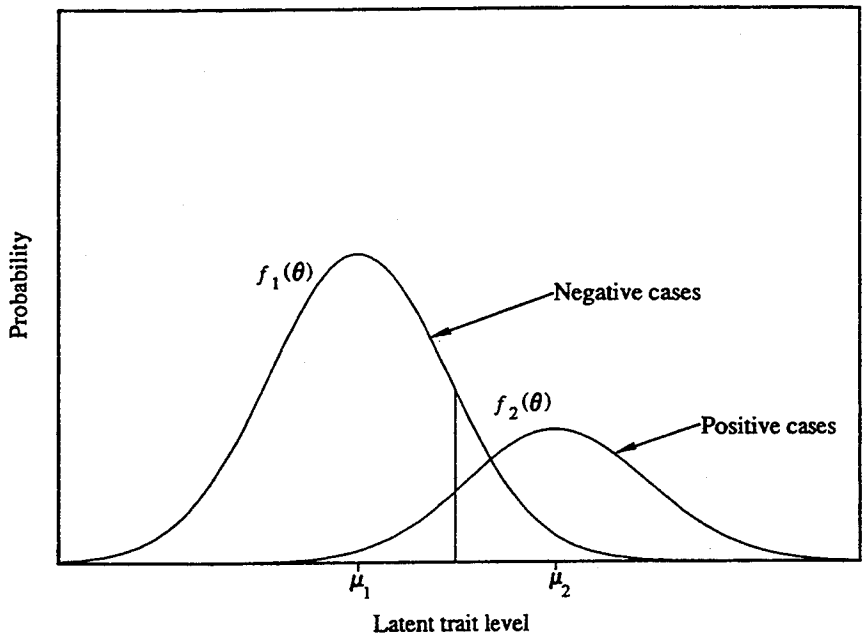

Figure 1. Model of discrete rating formation when the underlying trait is continuous. (The x-axis corresponds to the latent trait level, θ, e.g., the severity of symptoms associated with a disorder; $f_1(\theta)$ and $f_2(\theta)$ describe the relative proportions of negative and positive cases, respectively, at various levels of θ; the vertical line corresponds to the threshold of a hypothetical rater.)

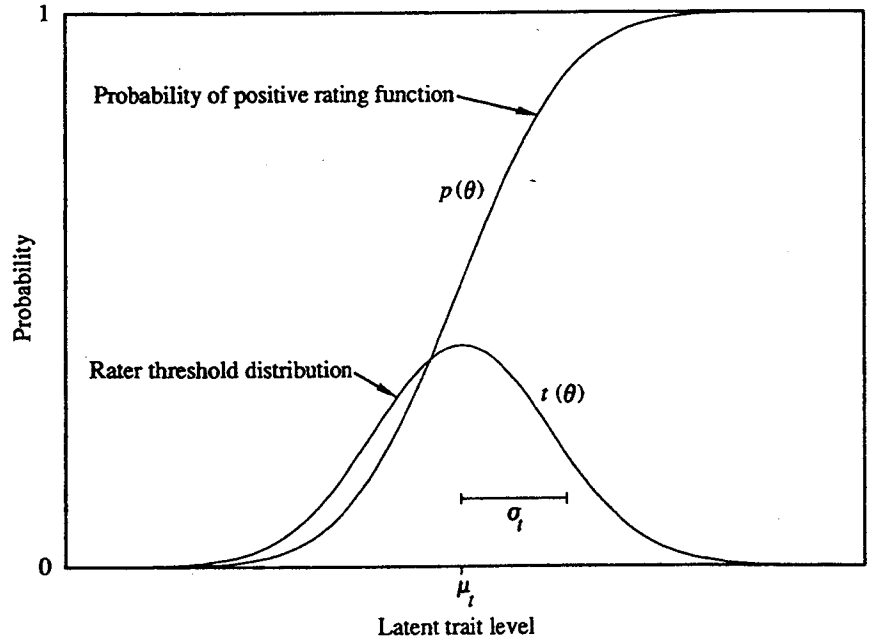

Figure 2. Probability density function of rater thresholds (normal distribution), $t(\theta)$, and probability of positive rating (ogive shaped) function, $p(\theta)$. (For a case at trait level θ, the probability of a positive rating is equal to the probability of sampling a rater threshold at or below that point or to the proportion of the rater threshold distribution less than or equal to θ. Thus the probability of positive rating function is equal to the cumulative distribution function of rater thresholds.)

Table 1
Glossary of Terms for Continuous Model

Term	Definition
θ	Latent trait level
$f_1(\theta), f_2(\theta)$	Relative frequency of negative and positive cases at θ
$f_3(\theta)$	$f_1(\theta) + f_2(\theta)$
N	Total number of cases rated
P	Prevalence of positive cases
μ_t	Mean threshold for making a positive rating
σ_t	Standard deviation of rater thresholds
$p(\theta)$	Probability of a positive rating given a case at θ; equal to cumulative distribution function of rater thresholds

of a threshold corresponds to its true negative rate, and the proportion to the right corresponds to its false positive rate. Raters are assumed to differ from one another in terms of the location of their thresholds. Specifically, let rater thresholds be assumed to be normally distributed along the trait continuum, with the probability of a rater threshold being located at each level of θ given by the probability density function $t(\theta)$, with mean μ_t and standard deviation σ_t.

Probability Model

In the discrete case, we considered the probability of a positive rating, p_s, given each latent category. We now wish to generalize this to provide a continuous function, $p(\theta)$, giving the probability of a positive rating for a case at any trait level, θ. Given a case at θ, the probability of the ith rating being positive is equal to the probability of sampling a rater whose threshold is less than or equal to θ, or the proportion of the probability density function of rater thresholds less than or equal to this value. Therefore

$$P\{u_i = 1 \mid \theta\} = p(\theta) = \int_{-\infty}^{\theta} t(\theta)\, d\theta, \tag{13}$$

with u_i denoting, as before, the outcome of the ith rating. We will call $p(\theta)$, the cumulative distribution function of $t(\theta)$, the *probability of positive rating* function. Readers familiar with item response theory will note that this is similar, both in terms of derivation and of function, to an item-characteristic curve. A difference is that whereas each test item is generally considered to have its own characteristic curve, $p(\theta)$ describes the probability of any rating's being positive. A better analogy, therefore, would be between this function and an aggregate item-characteristic curve describing the probability of answering correctly an item randomly selected from a pool of items with normally distributed difficulties. Fleiss (1965) has shown several interesting implications of item-characteristic curves for the measurement of rater agreement.

Again letting $u_.$ denote the number of positive ratings for a case, the conditional probability of exactly j ($j = 0, 1, \cdots, k$) positive ratings, given a case at trait level θ is

$$P\{u_. = j \mid \theta\} = \binom{k}{j} p^j(\theta)[1 - p(\theta)]^{k-j}. \tag{14}$$

The unconditional probability of exactly j positive ratings is therefore obtained by considering at each trait level the value of Equation 14 multiplied by the probability of selecting a case at that level, or

$$P\{u_. = j\} = \binom{k}{j} \int_{-\infty}^{\infty} f_3(\theta) p^j(\theta)[1 - p(\theta)]^{k-j}\, d\theta. \tag{15}$$

Thus, given N cases, the expected number with exactly j positive ratings is

$$E(n_j) = NP\{u_. = j\}, \tag{16}$$

and the likelihood of an observed set of frequencies $n_0, n_1, \cdots, n_k$, given a rating procedure characterized by $f_3(\theta)$ and $p(\theta)$, is

$$L = \prod_{j=0}^{k} P\{u_. = j\}^{n_j}. \tag{17}$$

Because $f_3(\theta) = f_1(\theta) + f_2(\theta)$, and because the definition of $p(\theta)$ is the cumulative distribution function of rater thresholds, we see that the likelihood of an observed pattern of ratings is therefore a function of $f_1(\theta)$, $f_2(\theta)$, and $t(\theta)$ or, assuming these to be normal, of the parameters $\mu_1, \sigma_1, \mu_2, \sigma_2, \mu_t, \sigma_t$, and P. Any one of the means and any one of the standard deviations may be specified arbitrarily, for example, $\mu_1 = 0$ and $\sigma_1 = 1$, reducing to five the number of parameters necessary to estimate. Provided that there are at least five ratings per case, maximum likelihood estimates may again be obtained numerically. The procedure is the same as in the discrete case, that is, the logarithm of Equation 17 taken to provide a more computationally tractable function and an estimation algorithm such as the Newton-Raphson method used to provide a convergent solution for the parameter values maximizing this function. The number of parameters necessary to estimate may be reduced if additional assumptions can be made, for example, specifying that $\sigma_1 = \sigma_2$, or if the prevalence of positive cases is known and model fit can be assessed by means of a likelihood ratio or goodness-of-fit χ^2 test, with df equal to k minus the number of parameters estimated.

Validity Coefficients

Once obtained, the parameters defining the distribution of cases on the latent trait continuum and probability of positive rating function can serve as a basis for inferences concerning the sensitivity and specificity of observer ratings. The formulas for estimating these values, shown in Table 2, follow directly from the rating model. For example, the probability of a case being observed at trait level θ, given that it is positive, is $(1/P)f_2(\theta)$, and the probability of a case at trait level θ receiving a positive rating is $p(\theta)$. Thus, the probability of a randomly selected positive case receiving a positive rating, that is, rating sensitivity, is given by the integral of $(1/P)f_2(\theta)p(\theta)$ over the range of θ. The formulas for rating specificity and the predictive value of positive and negative ratings follow similarly.

The interpretation of these values as sensitivity and specificity in the usual sense requires that the latent case types associated with the distributions $f_1(\theta)$ and $f_2(\theta)$ correspond exactly to positive and negative cases as they are generally defined. If the latent and actual types correspond only imperfectly, these results give upper-limit estimates for rating sensitivity and speci-

Table 2
Validity Coefficient Formulas for Continuous Model

Validity index	Computational formula
Sensitivity	$1/P \int_{-\infty}^{\infty} f_2(\theta)p(\theta)\,d\theta$
Specificity	$1/(1-P) \int_{-\infty}^{\infty} f_1(\theta)[1-p(\theta)]\,d\theta$
Percent correct	$\int_{-\infty}^{\infty} f_1(\theta)[(1-p(\theta)] + f_2(\theta)p(\theta)\,d\theta$
Predictive value of positive ratings	$\dfrac{\int_{-\infty}^{\infty} f_2(\theta)p(\theta)\,d\theta}{\int_{-\infty}^{\infty} f_3(\theta)p(\theta)\,d\theta}$
Predictive value of negative ratings	$\dfrac{\int_{-\infty}^{\infty} f_1(\theta)[1-p(\theta)]\,d\theta}{\int_{-\infty}^{\infty} f_3(\theta)[1-p(\theta)]\,d\theta}$

ficity, provided that the distribution of cases within each latent type is independent of true type. In the development of the following formulas, we shall assume an identity between latent type and actual category membership.

Example

A special case of the continuous model occurs when only one type of case exists and the latent trait can be viewed as following a single normal distribution. In this case, the rating model is described by two parameters, μ_t and σ_t; that is, rater accuracy and agreement are determined by the location of the mean rater threshold in relation to the distribution of cases and the variability of rater thresholds in relation to trait variability. This is especially useful because maximum likelihood estimates for these parameters can be obtained on the basis of as few as two ratings per case. The assumption of a trait characterized by a single normal distribution may, moreover, be more appropriate than the assumption of a mixture of two normal distributions in some cases.

For example, consider the reliability data concerning the diagnosis "depressive personality" reported in the field trials of the DSM-III classification system of psychiatric disorders (Williams & Spitzer, 1980). In this study, 662 patients were diagnosed by panels of $k = 2$ diagnosticians. From the value of the kappa coefficient reported, .157, it is possible to calculate the numbers of patients receiving the diagnosis once, twice, or not at all. Specifically, these values are $n_0 = 520$, $n_1 = 77$, and $n_2 = 65$. This diagnosis is not generally considered to have a specific genetic or biochemical etiology, and it would not be unreasonable to interpret it as an extreme form of characteristics that are normally distributed, making the special case of the continuous model described above plausible.

Necessary to estimate are μ_t and σ_t. To assure rapid convergence, starting values are selected by an ad hoc procedure, in this case, a simple multidimensional "grid search" algorithm that evaluates the log of the likelihood function for all combinations of parameter values, considering fixed increments of each parameter within a probable range. From this, initial estimates of $\mu_t = 1.1$, relative to a mean for the trait of 0, and $\sigma_t = .45$, relative to a trait standard deviation of 1, are obtained. Use of the Newton-Raphson algorithm then yields a convergent solution for the maximum likelihood estimates of these parameters of $\mu_t = 1.116$ and $\sigma_t = .456$, with standard errors of .065 and .056. In order to estimate rater sensitivity and specificity from the data, the formulas in Table 2 must be modified slightly to accommodate the provision of a single distribution of cases. If we let the probability density function of all cases be denoted by $f(\theta)$, rating sensitivity is estimated as the probability of a case that falls in the portion of the distribution generally considered to warrant a positive rating, which may be defined as the portion above the mean rater threshold, receiving a positive diagnosis by a randomly selected rater. This is given by the integral of $f(\theta)p(\theta)$ from μ_t to ∞, divided by the integral of $f(\theta)$ over the same range. Similarly, rating specificity is given by the integral of $f(\theta)[1 - p(\theta)]$ from $-\infty$ to μ_t, divided by the integral of $f(\theta)$ over the corresponding range. For the given data, values of approximately .79 and .94, respectively, are obtained. These values are much higher than what one might expect on the basis of the level of kappa reported. This may be seen as attributable to the tendency of the kappa coefficient to yield inordinately low values, given large differences in the proportions of positive and negative ratings, that is, the base rate problem discussed by Carey and Gottesman (1978) and others.

Multiple Opinions

If we follow the same reasoning as in the discrete model, the probability, given a positive case, of k consecutive positive ratings, or the sensitivity of a decision rule requiring unanimous positive ratings by k raters, is given by the integral of $(1/P)f_2(\theta)p^k(\theta)$ across θ. Similarly, the specificity of a rule requiring k unanimous negative ratings for a negative classification is equal to the integral of $[1/(1-P)]f_1(\theta)[1-p(\theta)]^k$ over θ. This leads to a generalization of Equation 9 for the probability of a case being positive, given an observed pattern of positive and negative ratings by a panel of observers. Let the function $p^*(\theta)$ be defined, giving the probability of some combination of positive and negative ratings for a case at trait level θ. Specifically, let

$$p^*(\theta) = \binom{k}{j} p^j(\theta)[1-p(\theta)]^{k-j}, \tag{18}$$

where the combination consists of j positive and $k - j$ negative ratings. The probability of a case being positive, given this combination of ratings, is then

$$\frac{\int_{-\infty}^{\infty} f_2(\theta)p^*(\theta)\,d\theta}{\int_{-\infty}^{\infty} f_3(\theta)p^*(\theta)\,d\theta},$$

and the probability of a case being negative is equal to 1 minus this amount.

Equations for estimating the accuracy of majority opinion decisions, similar to Equations 10 and 11, may again be derived. For example, the sensitivity of a positive majority rating is obtained by summing the numerator of the above expression across all values of j constituting a majority, multiplied by $1/P$. The predictive value of a majority positive rating is also obtained by dividing the sum of the numerator of the expression above across the appropriate values of j by the sum of the denominator over the same values. Equations for the specificity and predictive value of negative majority ratings can be constructed similarly.

Agreement Due to Error

It also follows from the continuous model that for a case at any trait level θ, the probability of two independent positive ratings is $p^2(\theta)$, and the probability of two negative ratings is $[1 - p(\theta)]^2$. The probability of two raters agreeing and being correct is therefore

$$\int_{-\infty}^{\infty} f_1(\theta)[1 - p(\theta)]^2 + f_2(\theta)p^2(\theta)\, d\theta,$$

and the probability of two raters agreeing and being in error is

$$\int_{-\infty}^{\infty} f_1(\theta)p^2(\theta) + f_2(\theta)[1 - p(\theta)]^2\, d\theta.$$

It is apparent that the former can be divided into separate components giving the probability of ratings being in agreement and correct for either a negative or a positive case. Dividing these by $1 - P$ and P, respectively, these terms provide separate indices of agreement not attributable to error that are independent of sample prevalences, that is, conditional upon a case being negative or positive.

Discussion

In summary, it has been shown that the use of probability modeling techniques related to latent class analysis and item response theory leads to many useful innovations in analyzing observer agreement. In certain cases, as when there are no systematic factors connecting the opinions of raters other than mutual accuracy, these methods lead directly to estimates for the accuracy of individual ratings and classifications based on the opinions of several raters. If factors other than mutual accuracy affect rater agreement, for example, sources of common error, these methods provide a basis for deriving upper-bounds estimates for rating sensitivity and specificity. They also lead to ways of expressing rater agreement that avoid many of the difficulties associated with previous approaches.

The probability modeling approach has many implications for how dichotomous ratings are treated in practice and suggests improvements in the ways these data are typically used. For example, Equation 9 and the corresponding continuous trait formula lead directly to the expectation that simple majority decision rules will result in suboptimal classification in certain instances. It is possible for cases receiving two negative ratings and three positive ratings to have a higher probability of belonging to the negative category. Thus, in order to make the best use of the information available in ratings by multiple observers, probability modeling may be necessary.

Limitations of these methods must also be considered. One is that the assumption of normal distributions in the continuous case may lead to lack of fit between the model and observed data. Other distributional forms for cases and rater thresholds may be considered, provided that they can be explicitly parameterized. For example, parameters reflecting the skew of the distributions of positive and negative cases may be added to the continuous model, potentially providing a more accurate fit, given that there are enough raters per case and associated degrees of freedom to estimate the additional parameters.

Although studies in which the rating panel is sampled separately for each case are common in psychological research, the alternative, for a fixed panel of raters to rate each case, is also widely used. In this case, the assumption of a constant probability, conditional on latent trait level, of any rating being positive is violated. If the raters are relatively similar, the effect of this would be expected to be small. However, in general, when data take the form of ratings by a fixed panel, such that each case is rated by each observer, other methods should be considered such as the latent class models discussed by Dawid and Skene (1979), Dillon and Mulani (1984), and Walter (1984). More complex fixed designs, such as those in which some or all of the raters repeat ratings at different points in time, can also be accommodated under the framework of Goodman's (1974) general latent class model.

The relationship between the continuous model and item response theory also becomes more apparent with data arising from a fixed-panel design. There, the situation of cases being rated positive or negative on some trait by each of a panel of raters is closely analogous to that of subjects answering correctly or incorrectly a set of test items. Thus, the item difficulty and subject ability estimates obtained by means of item response theory correspond directly to rater thresholds and case trait levels. Methods developed in conjunction with modern test theory may thus be used to provide estimates of these individual rating parameters. A paper describing this approach in more detail is currently in preparation (Uebersax & Grove, 1988).

References

Carey, G., & Gottesman, I. I. (1978). Reliability and validity in binary ratings: Areas of common misunderstanding in diagnosis and symptom ratings. *Archives of General Psychiatry, 35,* 1454–1459.

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

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics, 28,* 20–28.

Dillon, W. R., Madden, T. J., & Kumar, A. (1983). Analyzing sequential categorical data on dyadic interaction: A latent structure approach. *Psychological Bulletin, 94,* 564–583.

Dillon, W. R., & Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research, 19,* 438–458.

Fleiss, J. L. (1965). Estimating the accuracy of dichotomous judgements. *Psychometrika, 30,* 469–479.

Gelfand, A. E., & Solomon, H. (1974). Modeling jury verdicts in the American legal system. *Journal of the American Statistical Association, 69,* 32–37.

Gelfand, A. E., & Solomon, H. (1975). Analyzing the decision-making process of the American jury. *Journal of the American Statistical Association, 70,* 305–310.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable: Part I. A modified latent structure approach. *American Journal of Sociology, 79,* 1179–1259.

Grove, W. M., Andreason, N. C., McDonald-Scott, P., Keller, B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry, 38,* 408–411.

Haberman, S. J. (1978). *Qualitative data analysis* (vol. 1). New York: Academic Press.

Haberman, S. J. (1979). *Qualitative data analysis* (vol. 2). New York: Academic Press.

Hui, S. L., & Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics, 36,* 167–171.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory.* Homewood, IL: Irwin.

Janes, C. L. (1979). An extension of the random error coefficient of agreement to N × N tables. *British Journal of Psychiatry, 134,* 617–619.

Kaye, K. (1980). Estimating false alarms and missed events from interobserver agreement: A rationale. *Psychological Bulletin, 88,* 456–468.

Kraemer, H. C. (1982). Estimating false alarms and missed events from interobserver agreement: A reply to Kaye. *Psychological Bulletin, 92,* 749–754.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*(2), 99–120.

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

Schouten, H. J. A. (1985). *Statistical measurement of interobserver agreement.* Unpublished doctoral dissertation, Erasmus University, Rotterdam, The Netherlands.

Schouten, H. J. A. (1986). Statistical measurement of interobserver agreement. *Psychometrika, 51,* 453–466.

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement on psychiatric diagnosis revisited. *Archives of General Psychiatry, 44,* 172–177.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry, 42,* 725–728.

Sprott, D. A., & Vogel-Sprott, M. D. (1987). The use of the log-odds ratio to assess the reliability of dichotomous questionnaire data. *Applied Psychological Measurement, 11,* 307–316.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science, 182,* 990–1000.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin, 101,* 140–146.

Uebersax, J. S., & Grove, W. M. (1988). *New statistical methods for measuring diagnostic agreement.* Manuscript submitted for publication.

Uebersax, J. S. (in press). PANEL: Latent class modeling of interrater agreement. *Applied Psychological Measurement,* Computer Program Exchange, *48.*

Walter, S. D. (1984). Measuring the reliability of clinical data: The case for using three observers. *Revue d'Epidemiologie et Sante Publique, 32,* 206–211.

Williams, J. B. W., & Spitzer, R. L. (1980). DSM-III field trials: Interrater reliability and list of project staff participants. In American Psychiatric Association (Ed.), *Diagnostic and statistical manual of mental disorders* (3rd ed., pp. 467–481). Washington, DC: Editor.

Young, M. A. (1983). Evaluating diagnostic criteria: A latent class paradigm. *Journal of Psychiatric Research, 17,* 285–296.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374–378.

Received June 19, 1986
Revision received November 25, 1987
Accepted April 15, 1988 ■