A RAND NOTE

Latent Structure Agreement Analysis

John Uebersax, Will Grove

November 1989

RAND

The research described in this report was supported by The RAND Corporation using its own research funds.

The RAND Publication Series: The Report is the principal publication documenting and transmitting RAND's major research findings and final research results. The RAND Note reports other outputs of sponsored research for general distribution. Publications of The RAND Corporation do not necessarily reflect the opinions or policies of the sponsors of RAND research.

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

A RAND NOTE

N-3029-RC

Latent Structure Agreement Analysis

John Uebersax, Will Grove

November 1989

RAND

PREFACE

The probability modeling approach to analyzing rater agreement has emerged in the literature in a somewhat disjointed manner, with different models being proposed by various authors, complicating the task of the researcher who wishes to acquire familiarity with these methods and apply them in his or her research. The goal of this Note is to describe these approaches, emphasizing their basic similarities and viewing them as variants of a common methodology.

This Note should be of use both to the applied researcher who is interested in analyzing rater agreement data, and the technically oriented reader concerned with methods for analyzing agreement. Accordingly, not all sections are intended for all readers--the former group may find some sections to contain more detail than they require, and the latter may find some material redundant. Readers more concerned with substantive applications may want to concentrate on the introductory portions of each section and the computational examples.

Although an attempt has been made to be as comprehensive as possible in surveying previous work in this area, no doubt there are important contributions to this literature that have inadvertently been overlooked.

SUMMARY

How do we know how many opinions are required to make a diagnosis with necessary accuracy? One way is by examining how often physicians agree on the diagnosis. This Note discusses statistical techniques that can be used to analyze agreement data to address this and related questions. Specifically, these methods make it possible to determine from the opinions of panels of diagnosticians in an agreement study the following: (1) the probable accuracy of an individual diagnosis; (2) the probability of disease presence or absence given unanimous or conflicting opinions by several diagnosticians; and (3) how many opinions should be required to make the diagnosis. The methods discussed include two related techniques, which differ in assumptions about disease subtypes and associated differences among cases in their ability to be correctly diagnosed. These techniques have many applications in addition to that of medical diagnosis.

ACKNOWLEDGMENTS

We wish to thank David Kanouse and Barbara Williams, the current and former heads of the Behavioral Sciences Department at RAND, John Winkler, the associate head of the Behavioral Sciences Department, and Albert Williams, director of the RAND Health Sciences Program, without whose assistance and encouragement this work would not have been possible.

Ed Park kindly provided the data on physician ratings of treatment appropriateness considered in Sec. II.

We are also indebted to Patrick Shrout and Stephen Walter, who provided valuable comments concerning a previous version of this Note, and Daniel Relles for his careful review and helpful suggestions.

CONTENTS

PREFACE iii

SUMMARY v

ACKNOWLEDGMENTS vii

TABLES xi

Section
I. INTRODUCTION 1
Organization of This Note 4
Methods for Measuring Agreement 4

II. LATENT CLASS AGREEMENT ANALYSIS 8
Varying Rating Panel 8
Fixed Rating Panel 21
Directions for Future Research 29

III. LATENT TRAIT AGREEMENT ANALYSIS 30
Varying Rating Panel 30
Fixed Rating Panel 35
Directions for Future Research 39

Appendix
A. SAMPLE INPUT FILE FOR FIXED PANEL LATENT CLASS AGREEMENT MODEL 41

REFERENCES 43

TABLES

1.1. Example diagnostic agreement data 3
2.1. Minimum ratings per case required for model identifiability and assessment of fit: varying panel design 11
2.2. Observed and expected results for Yerushalmy rating data 18
2.3. Fit of alternative latent class models of Yerushalmy data 19
2.4. Parameter estimates for three-class model of Yerushalmy data 19
2.5. Minimum ratings per case required for model identifiability and assessment of fit: fixed panel design 23
2.6. Observed and expected results for physician ratings of treatment appropriateness 26
2.7. Fit of alternative latent class models of treatment appropriateness ratings 27
2.8. Parameter estimates for three-class model of treatment appropriateness ratings 27
3.1. Results of hypothetical diagnostic agreement study 38
3.2. Parameter estimates for fixed panel latent trait agreement model 38

I. INTRODUCTION

Powerful methods for measuring agreement on diagnosis and related forms of classification now exist. The origin of these methods can be traced as far back as Poisson's studies of juror agreement [1], and they are closely related to the well known statistical techniques of latent class and latent trait analysis [2-4]. That they are more computation-intensive than traditional approaches to measuring agreement has probably been a factor in their not yet having received widespread use. However, because of advances in microcomputer hardware and software, they are now well within reach of most researchers, and offer considerable promise for leading to better use of agreement data than has previously been possible.

To fully appreciate the usefulness of these methods and their advantages relative to other ways of measuring agreement, it is helpful to consider them in light of a hypothetical example.[1] Suppose that a patient is diagnosed as having a rare disease. Immediately there are several questions that come to mind, foremost among them being:

- How likely is it that this diagnosis is correct?

Suppose, though, that this question cannot be answered directly, since there is no definitive test for the disease. The questions then asked might be as follows:

---

[1]The hypothetical example, as well as much of our discussion, focuses on medical diagnosis as an instance of expert rating. However, it is understood that what is said applies equally well to other types of dichotomous classifications, e.g., the designation of a defendant as guilty or not guilty by jurors or the categorization of parts as operative or nonoperative by inspectors.

- To what extent do diagnoses by this diagnostician tend to reflect the judgment of other or most diagnosticians?

and

- Given that a single diagnosis is subject to error, how worthwhile would it be to obtain additional opinions, for example, a diagnosis by a second or even third independent source?

The latter question, in turn, gives rise to another:

- Given opinions by several diagnosticians, which may include both positive and negative diagnoses, what is the probability of having or not having the disorder?

The practical nature of these questions hardly needs to be emphasized. Data are routinely collected by means of inter-rater agreement studies for the specific purpose of answering them. In its basic form, such a study consists of a sample of N cases, each evaluated by two or more diagnosticians. The subset of such studies we are mainly concerned with here are those where (1) the number of opinions remains constant across cases, (2) diagnosticians formulate their opinions independently of one another, and (3) evaluations take the form of dichotomous ratings, for example, "disorder present" and "disorder absent." More complex models, such as those involving multiple or graded response categories, may be derived from this simplified model.

As an example of such a study, consider the data in Table 1.1. These data, originally presented by Yerushalmy [5], concern ratings of radiographic films as either indicative or not indicative of tuberculosis by eight physicians each. As shown, the majority of cases received eight negative ratings, with a smaller number receiving eight positive ratings. However, disagreement is also indicated by the cases receiving various combinations of positive and negative diagnoses.

Table 1.1

EXAMPLE DIAGNOSTIC AGREEMENT DATA

Number of Positive Diagnoses j	Observed Frequency f_j
0	13560
1	877
2	168
3	66
4	42
5	28
6	23
7	39
8	64

SOURCE: From Ref. 5.
NOTE: Each case diagnosed eight times. Total number of cases is 14,867.

Surprisingly, although studies such as this which collect multiple-rater agreement data are common, particularly in medical research [6-8], traditional methods for analyzing this information are not well suited to addressing the above-posed questions. Our purpose here is to describe and illustrate a statistical approach that makes it possible to answer these questions much more directly and precisely. Much of what we present is not new--rather, elements are to be found scattered throughout a diverse literature in statistics, medicine, psychology, sociology, and education. We attempt here to weave these elements into a coherent set of techniques that may be properly viewed as a *methodology*, rather than simply a set of methods.

ORGANIZATION OF THIS NOTE

In the remainder of this section, we review previous approaches for the measurement of agreement. Following this, we explain the basic rationale of the approach considered here, and present a taxonomy of specific techniques subsumed under the general model. In the next two sections, we present the two main variants of this approach.

METHODS FOR MEASURING AGREEMENT

Kappa and Other Agreement Indices

Several methods have previously been proposed for measuring agreement. The most common approach is the calculation of an *agreement index*. The most elementary such index consists of the proportion of times two ratings of the same case agree. Other agreement indices include Yule's Y [9], the odds ratio [10],[2] and the phi coefficient.

A widely known class of agreement indices is obtained by dividing the difference between observed pairwise agreement and the level expected by chance by 1 minus the level expected by chance [12]. Various indices of this class differ in how the expected proportion of chance agreement is calculated. Foremost among these indices is the kappa coefficient [13, 14], which estimates chance agreement based on the product of the marginal proportions of positive and negative ratings. Although kappa has been widely used, and although its usefulness for verifying that observed levels of agreement exceed chance levels is clearly established, concern has been expressed about its potential limitations. Several authors have discussed what has been termed the *base rate problem* [15-17], whereby a rating procedure with a high level of accuracy may yield low levels of agreement as measured by the kappa coefficient in samples where the proportion of positive and negative ratings (i.e., the base rates) are close to 1 and 0.[3]

---

[2]Darroch and McCloud [11] also develop a more extensive methodology for analyzing rater agreement based on the odds ratio.

[3]Shrout, Spitzer, and Fleiss [18], however, contended that this is in fact a desirable property.

A more fundamental limitation of agreement indices in general is that they summarize all of the information on agreement and disagreement in a single term. Thus, agreement on positive ratings and agreement on negative ratings are subsumed under one index, which results in loss of useful information. Further, because of the lack of an explicit probability model underlying their calculation, such measures do not readily permit agreement data to be used to answer the types of questions posed above.

Variance Components Approaches

An alternative is to express agreement on dichotomous classifications in a manner analogous to the intraclass correlation used to assess the reliability of interval or ratio scale measures [19-21]. By this approach, positive and negative ratings are coded 1 and 0, and the proportion of total variation among ratings that is attributable to between-case variability (i.e., not attributable to variation in ratings of the same case) is calculated as a measure of classification precision or reliability. Limitations characteristic of intraclass correlation approaches to expressing rating reliability in general, however, apply here as well. Specifically, raters with a given degree of consistency in the absolute sense of tending to agree or disagree on ratings of the same type of case will yield higher or lower intraclass correlations, depending upon the level of between-case variation, which is a function of the prevalence of positive and negative cases in the sample. This is directly analogous to the base rate problem of the kappa coefficient, just as the kappa coefficient itself is closely related to the intraclass correlation coefficient. Again, however, the more important limitation of variance-partitioning approaches is that they do not express agreement in a way that lends itself to answering the questions posed above.

Latent Structure Models

The methods we consider here fall under the general heading of what may be termed *Latent Structure Agreement Analysis*. These methods approach the problem of quantifying agreement from a much different perspective than agreement indices or variance-partitioning methods. Specifically, they develop a parameterized model, which entails an explicit characterization of the relationship between individual rater accuracy and inter-rater agreement. In essence, these methods may be understood as attempting to answer the question, What level of rater accuracy would be required to generate a pattern of agreements and disagreements such as that observed? This approach views the accuracy of raters and the prevalence of various types of cases as unobserved parameters, and estimates these parameters based on observed data. Once derived, these estimates can be applied in Bayesian calculations to provide answers to the kinds of questions posed above.

Four main variants of this approach have thus far been suggested in the literature. Agreement data may be collected using a research design where multiple opinions for each case come from either the same or differing sets of raters. We refer to these as *fixed* and *varying rating panel* designs, respectively.

In addition, a disorder may be viewed as *discrete* or *continuous*. By the former view, cases are seen as belonging to one of a relatively small set of categories or types, each of which corresponds to a certain trait level and has an associated probability of eliciting a positive rating. By the continuous view, cases are assumed to have trait levels and corresponding probabilities of eliciting positive ratings that may fall anywhere on a continuum. From the former assumption comes an approach to analyzing agreement that may be recognized as a special case of latent class analysis [2-4]; accordingly, we term the models in this category *Latent Class Agreement Analysis*. Methods based on the assumption of a continuously varying trait, in turn, may be seen as a special case of latent trait analysis [2, 22], and are therefore termed *Latent Trait Agreement Analysis*.

Either latent structure model may be combined with either rating design, leading to four main variants of the Latent Structure Agreement Analysis approach. These are termed the *varying panel/latent class*, *fixed panel/latent class*, *varying panel/latent trait*, and *fixed panel/latent trait* models, respectively.

II. LATENT CLASS AGREEMENT ANALYSIS

VARYING RATING PANEL

This approach corresponds to discussions of rater agreement by Gelfand and Solomon [1, 23], Kaye [24], Kraemer [25], and Uebersax [26].[1]

We first describe the basic approach and discuss the estimation and comparison of models, and then discuss the application of derived parameters to the estimation of rating accuracy and the interpretation of multiple opinions. Following this, we consider a computational example.

Model

Let N cases each be evaluated by randomly selected groups of k raters, and let each rater's evaluation take the form of a dichotomous rating, e.g., a positive or negative diagnosis. Recall (as illustrated in Table 1.1) that out of k ratings for each case, any number j (j = 0, 1, ..., k) may be positive. Considering outcomes across all cases, the frequencies of cases with each possible number of positive ratings may be obtained, denoted f_0, f_1, ..., f_k. In accordance with the assumptions of latent class analysis, these frequencies are assumed to be determined by two sets of parameters: the prevalences of c mutually exclusive and exhaustive latent classes to which cases belong (*latent class prevalences*) and the conditional probabilities of a positive rating, given a case belonging to each (*conditional rating probabilities*). Each latent class is assumed to correspond to a type of case with a specific probability of eliciting a positive rating. Thus, latent classes may represent genetically different subtypes of a disorder, or functional groupings of cases based on levels of symptom

[1] Stewart and Rey [27] and Fleiss and Shrout [28] also discuss methods that are similar, but require that estimates of the prevalence of positive and negative cases be available.

intensity or salience, for example, categories of "not symptomatic," "moderately symptomatic," and "highly symptomatic."

The prevalences of each latent class are denoted by π_1, π_2, ..., π_c, where π_s ($s = 1, 2, \ldots, c$) is the probability of a randomly sampled case belonging to latent class s.[2] For convenience, we indicate a positive rating by 1 and a negative rating by 0. We denote the conditional rating probabilities by $\pi_{1|1}$, $\pi_{1|2}$, ..., $\pi_{1|c}$, where $\pi_{1|s}$ is the probability of a positive rating given a member of latent class s.[3] By convention, we number latent classes in order of increasing probability of eliciting a positive rating, i.e., such that $\pi_{1|1} < \pi_{1|2} < \ldots < \pi_{1|c}$.

The expected number of cases receiving exactly j out of k positive ratings, denoted e_j, is equal to

$$e_j = N \binom{k}{j} \sum_{s=1}^{c} \pi_s \pi_{1|s}^{\,j} (1 - \pi_{1|s})^{k-j}. \qquad (1)$$

This leads to the set of expected frequencies of cases with various numbers of positive ratings, e_0, e_1, ..., e_k. The goal, then, is to obtain estimates of latent class prevalences and conditional rating

---

[2] A special case of these models occurs when it is assumed that there is only one class to which cases belong. This situation, which we shall be concerned with primarily in conjunction with the evaluation and comparison of latent class models, is related to the log-linear models for agreement analysis described by Tanner and Young [29].

[3] In the general case, this model requires that raters be randomly sampled for each case. Because each rating is thus a random sampling, the probability of a positive rating for a given case remains constant, even though raters themselves may differ in their tendency to make positive or negative ratings (in the fixed panel design discussed below, allowances are made for rater differences). The varying panel model, however, is also applicable under other circumstances--for example, if the same test is repeated on multiple occasions, or the same set of raters evaluate each case, but their probabilities of making positive ratings, conditional on latent class, are the same. The essential feature of the varying panel model is that, for a given case, the value of $\pi_{1|s}$ is always the same.

probabilities, i.e., π_s and $\pi_{1|s}$ parameters, that maximize the correspondence between observed and expected frequencies of cases with various numbers of positive ratings, that is, the f_j and e_j terms.

Estimation

Standard numerical estimation procedures can be used to find the parameter values that maximize this correspondence. Uebersax [26] described a procedure for obtaining maximum likelihood estimates based on the Newton-Raphson method, following the general approach of Lazarsfeld and Henry [2]. More recently, we have used an EM algorithm [30] related to that described by Goodman [3] and Dawid and Skene [31].[4] This algorithm is more flexible than the Newton-Raphson method, but tends to converge more slowly. A compromise is to initially apply the EM algorithm to obtain good approximations to maximum likelihood parameter estimates, and then to use these as starting values for the Newton-Raphson procedure, which converges more rapidly on final estimates. Approximate standard errors of parameter estimates are obtained by the standard method of inverting the information matrix [2].

Identifiability

Lazarsfeld and Henry [2], Goodman [3], and others discuss identifiability of latent class models. An unidentifiable model is analogous to a set of equations where there are more independent equations than variables, permitting an infinite number of solutions. A necessary condition for model identifiability is that the number of parameters requiring estimation be less than the number of degrees of freedom for the observed data. Given k ratings per case, there are k + 1 possible numbers of positive ratings, but only k degrees of freedom, since $f_0 + f_1 + \ldots + f_k = N$. Thus, for k ratings per case, there can be no more than k parameters requiring estimation. The parameters requiring estimation are c - 1 of the π_s terms (one need not be

[4] A description of this algorithm as applied to the varying panel model is presented in [32].

estimated, since they must sum to 1) and the c $\pi_{1|s}$ terms. Thus, for a model with c latent classes to be identifiable, it is necessary that $k \geq 2c - 1$, unless constraints are imposed on parameters. Table 2.1 shows, for varying panel models with various numbers of latent classes, the minimum number of ratings per case necessary for a model to be identifiable. It may be noted that for a model with two latent classes, at least three ratings per case are required to estimate model parameters (although four ratings would also permit a test of model fit). Further, it has been our experience (and also noted by Kraemer [25]) that two-class models are often not sufficient to characterize the complexity of a rating process. We have typically found models with three or four classes more suitable.

Satisfying the condition above is usually a *necessary* but not a *sufficient* condition for latent class model identifiability. The varying panel agreement model with dichotomous ratings, however, is a relatively simple application of the general latent class model. Experience thus far suggests that for this class of models the necessary condition above is also a sufficient condition, except in certain trivial cases, e.g., when all cases are unanimously rated positive or negative, or when data that are fit perfectly by a model with a smaller

Table 2.1

MINIMUM RATINGS PER CASE REQUIRED FOR MODEL IDENTIFIABILITY AND ASSESSMENT OF FIT: VARYING PANEL DESIGN

Number of Latent Classes	Ratings per Case Required for Model Identifiability	Ratings per Case Required for Chi-Square Test
1	1	2
2	3	4
3	5	6
4	7	8

NOTE: Assumes unconstrained model.

number of latent classes are analyzed using a model with a larger number of classes.

Constraints on Model Parameters

It is often possible to estimate a model not otherwise identifiable by imposing constraints on model parameters. For varying panel designs, the most common constraint involves setting one or more parameters to specified values. For example, one may set $\pi_{1|s}$ to 0 or 1 for a particular latent class. For examples of parameter constraints in the estimation of latent class agreement models, see Refs. 33 and 34.

Model Fit and Comparison

The fit of a latent class agreement model may be assessed by comparing observed results to what would be expected by the model, using either a Pearson or likelihood ratio chi-square statistic [3]. The Pearson chi-square is calculated by the formula $\chi^2 = [\Sigma_j(f_j - e_j)^2]/e_j$, and the likelihood ratio chi-square by the formula $L^2 = 2\Sigma_j f_j \log(f_j/e_j)$, where the values for e_j are calculated using estimates of π_s and $\pi_{1|s}$ parameters. The degrees of freedom associated with each is k - 1 minus the number of estimated parameters. For unconstrained models, this is equal to k - $2c$ + 1. Model fit is indicated by a low value relative to the degrees of freedom, i.e., a nonsignificant value. Statistical significance may be determined from standard tables of the χ^2 statistic.

An advantage of the likelihood ratio chi-square is that it permits comparison of alternative models of the same data. The statistical significance of the difference between two models is evaluated by subtracting their corresponding likelihood ratio chi-squares. The degrees of freedom for the resulting difference statistic is equal to the difference in the degrees of freedom for the individual chi-squares. This requires that the models compared be *nested*, i.e., that the parameters of one be a subset of those of the other. This is always the case for models that differ only in the number of latent classes.

In assessing model fit it is important to take sample size into account. Given a sufficiently large sample, even a small difference between observed and expected frequencies will likely result in significant chi-square values. Thus, it may also be useful to assess fit in terms of statistics such as the *normed fit index* [35], which are less sample size dependent. Clogg [36] recommends an equivalent index, calculated as $(L_0^2 - L_1^2)/L_0^2$, where L_1^2 is the likelihood ratio chi-square for a given latent class model, and L_0^2 is the corresponding statistic obtained using a one-class (independence) model. This is analogous to the proportion of variance unexplained by the one-class model that is explained by the multiple-class model

Having described the parameters of the varying rating panel latent class agreement model and discussed methods by which parameters are estimated and models evaluated, we now proceed to the subject of how these estimates can be used to address the questions concerning the accuracy and interpretation of ratings initially posed.

Estimation of Rating Accuracy

In the definition of latent classes we stated that each corresponds to a subset of cases with similar trait levels and probabilities of eliciting a positive rating. If each latent class can be interpreted as a variety of positive or negative case, model parameter estimates may be used to directly estimate rating accuracy.[5]

The accuracy of dichotomous ratings is commonly expressed in terms of four indices: *sensitivity* (*Se*), *specificity* (*Sp*), *positive predictive validity* (*Pv+*), and *negative predictive validity* (*Pv-*). Rating sensitivity is defined as the probability of a positive rating given a positive case. Rating specificity is the probability of a

[5]Alternatively, each latent class may be viewed as a specific mixture of positive and negative cases. In this situation, slightly more complex formulas than those presented here are required, but, as discussed in Uebersax [26], it is generally possible to derive at least upper bound estimates for rating accuracy using these methods.

negative rating given a negative case. Positive and negative predictive validity are defined as the reverse conditional probabilities of sensitivity and specificity. That is, positive predictive validity is the probability of a positive case given a positive rating and negative predictive validity is the probability of a negative case given a negative rating. By denoting a positive and negative case + and -, and a positive and negative rating '+' and '-', we may define *Se* = Pr['+'|+], *Sp* = Pr['-'|-], *Pv+* = Pr[+|'+'], and *Pv-* = Pr[-|'-'].

For a given model, let the numbers a and b be such that latent classes 1, 2, ..., a are subtypes of negative cases, and latent classes b, $b + 1$, ..., c are subtypes of positive cases. *Se*, Sp, *Pv+*, and *Pv-* are then obtained as follows:

$$Se = \frac{\sum_{s=b}^{c} \pi_s \pi_{1|s}}{\sum_{s=b}^{c} \pi_s}, \tag{2}$$

$$Sp = \frac{\sum_{s=1}^{a} \pi_s (1 - \pi_{1|s})}{\sum_{s=1}^{a} \pi_s}, \tag{3}$$

$$Pv+ = \frac{\sum_{s=b}^{c} \pi_s \pi_{1|s}}{\sum_{s=1}^{c} \pi_s \pi_{1|s}}, \tag{4}$$

and

$$Pv- = \frac{\sum_{s=1}^{a} \pi_s (1 - \pi_{1|s})}{\sum_{s=1}^{c} \pi_s (1 - \pi_{1|s})}. \tag{5}$$

Also of interest is the *false-negative error rate*, or the probability of a negative rating given a positive case, equal to 1 -

Se, and the *false-positive error rate*, or the probability of a positive rating given a negative case, equal to 1 - *Sp*.

Interpreting Multiple Opinions

One of its most useful features is that the latent class approach to analyzing agreement leads directly to methods for the interpretation and integration of opinions by multiple raters. Again, let latent classes be assumed to be varieties of negative and positive cases. Simple Bayesian calculations show that the probability of a case being positive, i.e., belonging to one of the positive latent classes, given exactly j out of k positive ratings, is equal to

$$\Pr[+|j' = j] = \frac{\sum_{s=b}^{c} \pi_s \pi_{1|s}^{j} (1 - \pi_{1|s})^{k-j}}{\sum_{s=1}^{c} \pi_s \pi_{1|s}^{j} (1 - \pi_{1|s})^{k-j}}, \qquad (6)$$

where j' is a variable to denote the number of positive ratings observed for a case. Subtracting this from one, the probability of a case being negative given j out of k positive ratings is obtained. This equation can be used to classify cases in the original rating study as positive or negative. By consideration of other values for k, it may also be used to derive classification rules for future cases based on different numbers of ratings.

Number of Opinions Necessary for Required Accuracy

The above formula is easily applied to determine the number of opinions necessary to insure a required degree of classification accuracy. Suppose, for example, that a sufficiently accurate classification is defined as one with a certain positive predictive validity. One may then, for example, ask what the minimum number of ratings is such that the probability of a case being positive, given unanimous positive ratings, is greater than or equal to this value. The situation of unanimous positive ratings may be seen as a special case of

the above where $j = k$. Thus this formula can be used to estimate the positive predictive validity of unanimous positive ratings by panels of 1 rater, 2 raters, etc. The minimum panel size necessary to classify a case positive with the required accuracy would then be the smallest number needed for (6) to exceed the criterion established. By extension of this reasoning, one may allow for non-unanimous panel outcomes or use other criteria for minimal required accuracy in determining panel size.

We have shown how parameter estimates obtained with the latent class agreement model can be used to estimate rater accuracy, probabilistically interpret opinions by multiple raters, and determine an appropriate number of opinions for a sufficiently accurate classification. Of necessity, we consider only some of the applications possible. Many others are implicit in the ability of these methods to provide direct or upper bound estimates of rating accuracy. For example, estimated rater accuracy can be used to determine the expected attenuation in statistical power of comparisons that involve groups whose members are assigned on the basis of fallible ratings [37], estimate the decrease in apparent accuracy of a diagnostic test compared to a criterion diagnosis that is itself unreliable [38], or correct for bias in estimation of disease prevalence due to misclassification error [39].

Software

Varying panel latent class agreement models can be estimated with the PANEL microcomputer program. We document this program in a companion RAND Note [32].

Varying Number of Ratings per Case

The varying panel latent class model may be generalized to designs where cases are rated different numbers of times. Examples of this occur when some ratings are lost or only some cases in a study are multiply rated. Let K be the maximum number of ratings any case receives. We may summarize results of an agreement study as the proportion of cases that are rated k ($k = 1, 2, \ldots, K$) times and

receive j (j = 0, 1, ..., k) positive ratings. The EM algorithm may again be used to obtain maximum likelihood estimates of π_s and $\pi_{1|s}$ parameters.

Pearson and likelihood ratio chi-square statistics can be used to test model fit. These may be viewed as the sum of separate chi-squares for cases with each number of ratings. Assessment of statistical significance for a test of model fit is complicated by the fact that outcomes for various values of k could not be interpreted as resulting from independent multinomials. To test statistical significance requires that a common multinomial be estimated. It is not difficult to see how this can be done. Given an underlying multinomial for results with K ratings, expectations of results with $k < K$ ratings are obtained using a formula related to the hypergeometric distribution (Uebersax [26], Equation 6). Thus, a likelihood function may be constructed for results across all values of k given probabilities for the K-way multinomial. These probabilities may then be estimated from observed data using a numerical procedure such as the Newton-Raphson method. An analytic method for estimating the common underlying multinomial may also be possible. Chi-square statistics are calculated by comparison of the proportions of cases with various combinations of k and j expected given the latent class model with those expected given the multinomial model. The degrees of freedom for this test are equal to K minus the number of estimated latent class model parameters.

The assumption of a common multinomial is not necessary, however, to use the difference likelihood ratio chi-square statistic for comparison of alternative latent class models. For nested models, this may be calculated and tested for significance as before, with degrees of freedom equal to the difference in the number of estimated parameters. The normed fit index may also be calculated and used as before.

Example

We illustrate these methods with the Yerushalmy data previously shown in Table 1.1. Three models, with two, three, and four latent classes, designated M_2, M_3, and M_4, are estimated.[6] Table 2.2 contains expected frequencies for each model. The correspondence of expected and observed frequencies is seen to increase with the number of latent classes. Fit indices are shown in Table 2.3. The two-class model does not fit well. Chi-square statistics for a three-class model are statistically significant, suggesting lack of fit, but this is partly due to the sample size. The likelihood ratio chi-square for a one-class independence model is 7160.808, resulting in a normed fit index for M_3 of 0.997, so that, by this criterion, M_3 does provide good fit. Model M_4 fits the data better than M_3 by a statistically significant degree, (difference L^2 of 21.897 - 0.099 = 21.798, with 3 - 1 = 2 df), but, again, this is virtually guaranteed by the large sample size.

Table 2.2

OBSERVED AND EXPECTED RESULTS FOR YERUSHALMY RATING DATA

Number of Positive Ratings j	Observed Frequency f_j	Expected Frequency e_j Model		
		M_2	M_3	M_4
0	13560	13452.90	13557.27	13559.99
1	877	1090.14	883.24	877.02
2	168	45.27	146.65	167.91
3	66	25.08	92.25	66.29
4	42	55.10	42.24	41.25
5	28	79.94	16.39	29.05
6	23	72.49	21.68	22.13
7	39	37.56	50.51	39.64
8	64	8.51	56.76	63.73

[6]We estimate these models using the PANEL program.

Table 2.3

FIT OF ALTERNATIVE LATENT CLASS MODELS
OF YERUSHALMY DATA

Model	df	Pearson Chi-Square χ^2	Likelihood Ratio Chi-Square L^2	Normed Fit Index
M_2	5	874.201	528.495	0.926
M_3	3	22.473	21.897	0.997
M_4	1	0.099	0.099	1.000

Table 2.4

PARAMETER ESTIMATES FOR THREE-CLASS MODEL
OF YERUSHALMY DATA

Latent Class s	Prevalence π_s	Conditional Positive Rating Probability $\pi_{1\|s}$
1	0.9636 (0.0027)	0.0072 (0.0003)
2	0.0275 (0.0024)	0.2660 (0.0177)
3	0.0088 (0.0008)	0.9003 (0.0134)

NOTE: Standard errors are shown below estimates in parentheses.

We accordingly focus our attention on M_3 (Table 2.4). For illustration, we assume that the three latent classes consist of two negative classes and one positive class with respect to tuberculosis, for example, (1) unaffected cases, (2) cases with less serious conditions that have an elevated probability of being diagnosed positive, and (3) cases with tuberculosis. Since there is only one

positive latent class, Equation (2) reduces to make rating sensitivity equal to $\pi_{1|3}$, estimated as 0.9003. From Equation (3), rating specificity is estimated as [(0.9636)(1 - 0.0072) + (0.0275)(1 - 0.2660)]/(0.9636 + 0.0275) = 0.986.[7] From Equations (4) and (5), the positive predictive validity of diagnosis is estimated as approximately 0.357, and negative predictive validity as 0.999.

We also use parameter estimates to determine probable diagnostic status given combinations of positive and negative ratings. From Equation (6) we estimate the probability of tuberculosis given one positive and one negative rating as 0.061. Since the probability of tuberculosis given one positive rating, $Pv+$, is estimated as 0.357, we see the difference that a second negative rating makes.[8] Similarly, given five positive and three negative ratings, Equation (6) results in an estimated probability of 0.263 of a positive case.

Finally, we consider how many opinions are necessary to make a diagnosis with required accuracy. Suppose that we define sufficient accuracy as a positive predictive validity of at least 0.90. Use of Equation (6) results in estimated predictive validities of 0.781, 0.925, and 0.977 for a positive diagnosis based on unanimous positive ratings by two, three, and four diagnosticians, respectively. We would therefore need a minimum of three ratings to obtain the necessary accuracy.

We have thus shown how, by the latent class approach, agreement data can be used to address the practical questions concerning ratings initially posed. We next consider a version of these methods applicable to fixed panel designs.

---

[7]We base calculated values on four-place accuracy of parameter estimates; rounding error may therefore occur.

[8]This illustrates the practical value of the latent class modeling approach. One would of course expect a lower probability of a disorder given that the second opinion is negative. But without such an approach it would not be possible to determine by how much it is reduced.

FIXED RATING PANEL

In a fixed panel design the same raters are used to rate each case, corresponding to what is also commonly called a *fully crossed* rating design. This design is useful in that it usually requires fewer raters, and provides information about the comparative accuracy of individual raters.

Discussions of fixed panel latent class agreement models may be found in Bergan [40], Clogg [41], Dawid and Skene [31], Dillon and Mulani [33], Espeland and Handelman [34], Uebersax and Grove [42], and Walter and Irwig [43]. The fixed panel agreement model corresponds closely to traditional latent class analysis applications as described by Lazarsfeld and Henry [2], Goodman [3], and Haberman [4].

Model

We again assume that each of a sample of N cases is rated by a panel consisting of k raters. We now assume, though, that the raters are the same for each case, and are numbered $j = 1, 2, \ldots, k$.

Let a positive rating again be represented by 1 and a negative rating by 0. Let the vector $\mathbf{u}_i$ be one of I ($I = 2^k$) unique patterns of positive and negative ratings (see Table 2.6), whose jth element, u_{ij}, corresponds to the rating of the jth rater. As before, let s denote one of c latent classes to which a case may belong, and let the prevalence of latent class s be π_s.

Again, latent classes are defined such that all cases belonging to the same latent class have the same probability of eliciting a positive rating. However, we now allow this probability to be different for each rater. To accommodate this, a slightly different notation is adopted. Specifically, let $\pi_{1|sj}$ ($s = 1, 2, \ldots, c$; $j = 1, 2, \ldots, k$) be the conditional probability of a case belonging to latent class s being rated positive by rater j.

Given π_s and $\pi_{1|sj}$ parameters, we may calculate the joint probability of a case being a member of latent class s and receiving rating pattern $\mathbf{u}_i$. This, denoted by π_{is}, is calculated as

$$\pi_{is} = \pi_s \prod_{j=1}^{k} \pi_{1|sj}^{u_{ij}} (1 - \pi_{1|sj})^{1-u_{ij}}. \tag{7}$$

The exponents u_{ij} and $1 - u_{ij}$ function such that either $\pi_{1|sj}$ or $1 - \pi_{1|sj}$ are counted in calculating the joint probability, depending on whether the jth rater's rating is positive or negative. The expected frequency of each rating pattern, e_i ($i = 1, 2, \ldots, I$), is then given by

$$e_i = N \sum_{s=1}^{c} \pi_{is}. \tag{8}$$

Estimation, Identifiability, and Assessment of Model Fit

The results of ratings by k raters across N cases may be summarized by the number of times each rating pattern occurs, i.e., a set of observed frequencies, f_i ($i = 1, 2, \ldots, I$). The purpose of estimation is to obtain estimates for π_s and $\pi_{1|sj}$ parameters that lead to expected frequencies as close as possible to observed frequencies. Again, the EM algorithm can be used to obtain maximum likelihood estimates.

The subject of identifiability for this class of models is fully discussed by Goodman [3] in the context of the general latent class model. As in the varying panel case, there are $c - 1$ prevalence parameters, but there are ck conditional rating probability parameters (one for each combination of rater and latent class), making the total number requiring estimation $c(k + 1) - 1$. For unconstrained models, a unique solution therefore requires that $I \geq c(k + 1)$. Again, this is a necessary but not a sufficient condition. In the case of two latent classes (see Table 2.5), three raters are required, which is consistent with this formula. For a three-class model, however, five raters are required, even though this formula suggests that four would be enough. The general method for establishing model identifiability is by evaluating the rank of the matrix of derivatives of pattern probabilities with respect to model parameters [3], or the rank of the

Table 2.5

MINIMUM RATINGS PER CASE REQUIRED FOR MODEL IDENTIFIABILITY AND ASSESSMENT OF FIT: FIXED PANEL DESIGN

Number of Latent Classes	Number of Raters Required for Model Identifiability	Number of Raters Required for Chi-Square Test	df
1	1	2	1
2	3	4	6
3	5	5	14
4	5	5	8

NOTE: Assumes unconstrained model; degrees of freedom are for χ^2 or L^2 test with minimum required number of raters.

matrix of second derivatives of the log-likelihood function with respect to model parameters. This test is automatically performed by standard latent class analysis programs. If a model is found to be not identifiable, the number of estimated parameters must be reduced, either by decreasing the number of latent classes or by imposing constraints on parameters.

As with the varying panel model, one useful type of constraint is to require certain parameters to be equal to fixed values. Another useful constraint for fixed panel models is to require that certain conditional rating probabilities be equal, e.g., the values of $\pi_{1|sj}$ be the same across raters for a particular latent class.

Model fit is again assessed with the χ^2 or L^2 chi-square statistic. The formulas are the same, except that the number of observed and expected frequencies now equal I, and the degrees of freedom for the statistics now equal I - 1 minus the number of estimated parameters. The normed fit index may also be calculated as before.

Applications

Parameter estimates can again be used to estimate rater accuracy. We again assume that latent classes are interpretable as varieties of negative and positive cases, latent classes 1 through *a* corresponding to negative cases and latent classes *b* through *c* to positive cases, and understand that when such a simple differentiation of latent classes is not possible the procedures described below may be suitably modified.

The accuracy of individual raters may be expressed using the indices discussed earlier, *Se*, *Sp*, *Pv+*, and *Pv-*, subscripts being added to denote values for each rater. These are obtained from Equations (2) through (5), with estimates of $\pi_{1|sj}$ used in place of those of $\pi_{1|s}$. Resulting values may also be averaged across raters, providing mean accuracy indices.

One may again use parameter estimates to classify cases based on multiple ratings. Recalling the definition of π_{is} as the joint probability of a case belonging to latent class *s* and receiving rating pattern $\mathbf{u}_i$, the probability of a case being positive given this pattern is

$$\Pr[+|\mathbf{u}_i] = \frac{\sum_{s=b}^{c} \pi_{is}}{\sum_{s=1}^{c} \pi_{is}}. \qquad (9)$$

An important aspect of Equations (7) and (9) is that they lead to different probabilities of a case being positive depending on which raters make positive and negative ratings. We discuss implications of this in the example below.

Software

The fixed panel latent class agreement model can be implemented using standard latent class analysis programs such as Clogg's MLLSA [36] and Haberman's LAT [4]. The PANMARK program of van de Pol, Langeheine, and de Jong [44], though primarily intended for Markov model analysis,

can also be used for these models. All of these programs are available in microcomputer form.

Example

It is useful to consider the fixed panel model in an application other than diagnosis, since, in fact, the applicability of these methods extends far beyond that context. We consider ratings on the appropriateness of 859 possible indications for performing the procedure carotid endarterectomy by a panel of medical experts, gathered in a study described by Park et al. [45]. For present purposes, we recode ratings, originally made on a nine-point Likert-type scale (1 = extremely inappropriate indication; 9 = extremely appropriate indication), to dichotomies, a positive rating corresponding to a judged indication and a negative rating to a nonindication. The observed frequencies of all possible rating patterns among five raters are shown in Table 2.6. We consider models with two, three, and four latent classes, designated M_2, M_3, and M_4. The expected pattern frequencies given each model are also shown in Table 2.6. Fit indices for each model are shown in Table 2.7. The χ^2 and L^2 statistics for both M_3 and M_4 are nonsignificant, indicating good fit.

A one-class independence model yields a value of 1433.925 for L_0^2. Using this to calculate the normed fit index, we see that M_3 and M_4 also provide good fit by this criterion. With a difference likelihood ratio chi-square of 23.059 - 7.534 = 15.525 (16 - 14 = 2 df), the fit of M_4 is better than that of M_3 by a statistically significant amount, but this must be weighed against the greater parsimony of M_3.

Parameter estimates for M_3 are shown in Table 2.8.[9] To see how these might be used, suppose that the three latent classes are (1) nonindications, (2) equivocal indications, and (3) valid indications for treatment, and that of interest is, for each rater, the probability of a positive rating given a valid indication, or each rater's sensitivity.

---

[9]Parameter estimates shown are from the MLLSA program. The input file used to generate these results is shown in the Appendix. Standard errors shown are from the PANMARK program.

Table 2.6

OBSERVED AND EXPECTED RESULTS FOR PHYSICIAN RATINGS OF TREATMENT APPROPRIATENESS

Rating Pattern i	Rater 1	2	3	4	5	Observed Frequency f_i	Expected Frequency e_i Model M_2	M_3	M_4
1	+	+	+	+	+	69	35.52	69.25	68.94
2	+	+	+	+	-	2	3.42	1.85	2.47
3	+	+	+	-	+	4	9.14	4.36	4.43
4	+	+	+	-	-	1	0.88	0.17	0.16
5	+	+	-	+	+	2	20.20	2.11	2.24
6	+	+	-	+	-	1	1.96	0.25	0.76
7	+	+	-	-	+	0	5.23	0.59	0.00
8	+	+	-	-	-	0	0.51	0.14	0.00
9	+	-	+	+	+	82	102.93	80.75	81.13
10	+	-	+	+	-	4	10.11	9.90	4.70
11	+	-	+	-	+	23	26.91	23.69	25.29
12	+	-	+	-	-	8	4.34	6.52	6.92
13	+	-	-	+	+	67	59.59	63.80	66.71
14	+	-	-	+	-	24	10.33	19.50	24.31
15	+	-	-	-	+	42	23.65	45.72	40.81
16	+	-	-	-	-	41	55.48	41.41	41.14
17	-	+	+	+	+	0	1.08	0.04	0.00
18	-	+	+	+	-	0	0.10	0.01	0.00
19	-	+	+	-	+	0	0.28	0.03	0.00
20	-	+	+	-	-	0	0.03	0.01	0.00
21	-	+	-	+	+	0	0.62	0.09	0.00
22	-	+	-	+	-	0	0.06	0.02	0.00
23	-	+	-	-	+	0	0.16	0.06	0.00
24	-	+	-	-	-	0	0.02	0.02	0.00
25	-	-	+	+	+	5	3.30	3.56	2.61
26	-	-	+	+	-	0	1.34	1.51	1.14
27	-	-	+	-	+	8	2.69	3.32	8.66
28	-	-	+	-	-	8	12.13	9.04	7.57
29	-	-	-	+	+	5	6.74	9.95	7.02
30	-	-	-	+	-	28	31.92	26.41	26.98
31	-	-	-	-	+	49	58.16	48.69	48.17
32	-	-	-	-	-	386	370.39	386.25	386.86

SOURCE: Park et al. [45].

NOTE: Total N of 859. Columns may not sum to total due to rounding.

Table 2.7

FIT OF ALTERNATIVE LATENT CLASS MODELS OF TREATMENT APPROPRIATENESS RATINGS

Model	df	Pearson Chi-Square χ^2	Likelihood Ratio Chi-Square L^2	Normed Fit Index
M_2	21	126.347	130.496	0.909
M_3	16	24.085	23.059	0.984
M_4	14	9.248	7.534	0.995

NOTE: Degrees of freedom shown are obtained from the MLLSA program, which treats parameter estimates of 0 or 1 as constrained, reducing the number considered estimated.

Table 2.8

PARAMETER ESTIMATES FOR THREE-CLASS MODEL OF TREATMENT APPROPRIATENESS RATINGS

Latent Class s	Prevalence π_s	Conditional Positive Rating Probability				
		$\pi_{1\|s1}$	$\pi_{1\|s2}$	$\pi_{1\|s3}$	$\pi_{1\|s4}$	$\pi_{1\|s5}$
1	0.5838	0.0712	0.0000	0.0213	0.0596	0.1023
	(0.0219)	(0.0183)	--	(0.0081)	(0.0121)	(0.0165)
2	0.2625	0.8972	0.0118	0.3277	0.5967	0.7805
	(0.0224)	(0.0341)	(0.0154)	(0.0565)	(0.0497)	(0.0440)
3	0.1537	1.0000	0.5783	0.9806	0.9437	0.9752
	(0.0212)	--	(0.0710)	(0.0274)	(0.0285)	(0.0183)

NOTE: Standard errors are shown in parentheses. Estimates of 1 or 0 indicate convergence to a boundary value [3] (see also NOTE for previous table); for these, standard errors are not calculated.

These are equal to the estimates of $\pi_{1|3j}$ shown in Table 2.8. Thus, estimated rater sensitivities, Se_1, Se_2, Se_3, Se_4, and Se_5, of 1.0000, 0.5783, 0.9806, 0.9437, and 0.9752, are obtained. Following the

procedure for calculating positive predictive validity, we obtain estimates of 0.357, 0.966, 0.605, 0.431, and 0.362 for $Pv+_1$, $Pv+_2$, $Pv+_3$, $Pv+_4$, and $Pv+_5$.

From Equations (7) and (9) we see that the probability of a possible indication being a true indication given five positive ratings is 0.995. Suppose, however, that of the five ratings, four are positive and one is negative. The probability of a true indication now depends on which rater makes the negative rating: if it is Rater 4, for example, we obtain an estimate of 0.943; however, if it is Rater 2, we estimate the probability as only 0.622.

It is by its ability to combine opinions in an explicit and probabilistically correct way that the fixed panel latent class agreement model demonstrates perhaps its greatest value relative to traditional ways of interpreting panel ratings. For example, the non-Bayesian view might hold a rating pattern of {+, -, +, +, +} to just as strongly indicate a positive case as a pattern of {+, +, +, -, +}. However, this is neither probabilistically correct, nor necessarily the way we really interpret multiple opinions. If one rater tends to make positive or negative ratings more often than others, we are likely to take this information into account. All other things being equal, a positive rating by a conservative rater gives us greater cause to believe that a case is positive than one by a nonconservative rater.[10] An important limitation of traditional methods for interpreting multiple rater opinions is that they do not take this into account.

This also suggests why it may be useful to include in panels both conservative and nonconservative raters. If the need arises to identify a positive case with a high degree of certainty, one may be selected that even conservative raters rate as positive. Conversely, negative ratings by nonconservative raters may be useful when there is a need to identify a case as negative with a high degree of certainty.

---

[10]An interesting consequence of this is the opportunity it provides for "gaming" by raters. For example, if a rater wanted to ensure that a positive rating carried the most weight, it would be advantageous to make positive ratings sparingly up to that point--thus appearing conservative. A positive rating would then be interpreted as stronger evidence of a positive case.

Again, we have considered only some of the applications that the fixed panel latent class agreement model permits.

DIRECTIONS FOR FUTURE RESEARCH

We believe that the methods described in this section offer many advantages for the analysis and interpretation of rater agreement, and recommend their use. One possible concern is the assumption of cases as falling into only a small number of latent classes. Although it would naturally be more appealing to think of disorders as displaying instead continuously varying levels of a latent trait, latent class models appear to provide a suitable approximation for a large number of applications.

There are several areas where additional research would be helpful. The extent to which sample size affects the accuracy of parameter estimation needs to be investigated; simulation studies may prove helpful in determining this. Generalizations of these methods may also increase their range of application. It should be possible to adapt the fixed panel model to allow for missing observations, or rotating panel designs where the raters rating each case are systematically varied. We have considered only dichotomous ratings here, but latent class models for polytomous ratings have also been discussed [31, 33, 40, 41, 46]. Latent class models can also be used to analyze agreement on ordered response category or Likert-type ratings [47].

III. LATENT TRAIT AGREEMENT ANALYSIS

The methods in this section are related to the statistical techniques of item response theory [22, 48] and Rasch modeling [49], which together may be subsumed under the more general heading of *latent trait analysis* [2]. We term the use of these methods in the analysis of agreement *Latent Trait Agreement Analysis*. Related discussions may be found in Fleiss [50] and Kraemer [37], and Quinn [51] has recently shown that equivalent models may be derived from signal detection theory [52].

VARYING RATING PANEL

Model

We begin by assuming a continuous dimension of trait intensity or severity. The location of a case on this continuum we term its *latent trait level*, and denote by θ. The word "trait" is used broadly, and it is understood that the continuum may also be an aggregate dimension based on several traits or symptoms.

The latent trait agreement model may be understood in terms of two functions (Fig. 3.1). The first, $f(\theta)$, describes the probability of encountering a case at each latent trait level θ. The second, $p(\theta)$, describes the probability, given a case at level θ, of a positive rating. We term $f(\theta)$ the *trait probability density function*, and $p(\theta)$ the *probability of positive rating* (or *diagnosis*) *function*.

The probability of a randomly selected case being rated positive is equal to a weighted average of $p(\theta)$ over all levels of θ, where the weight is the probability of a case having trait level θ, i.e., $f(\theta)$. Thus, it is equal to the product of $f(\theta)$ times $p(\theta)$ summed across the range of θ, or the integral of $f(\theta)p(\theta)$ over all levels of θ.

If $f(\theta)$ and $p(\theta)$ were known, they would lead directly to estimates for the probability of various patterns of agreement and disagreement by multiple raters. For example, given a case at level θ, the probability of two positive ratings is $p(\theta)^2$; for a randomly selected case, this

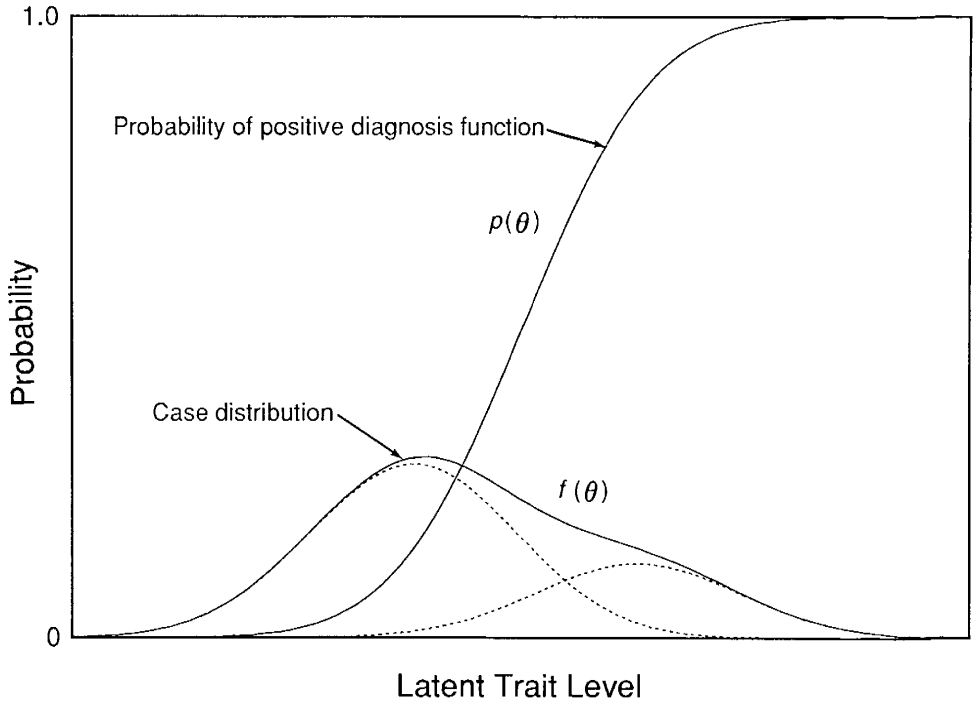

Fig. 3.1--Varying panel latent trait agreement model: trait probability density function, $f(\theta)$, and probability of positive rating function, $p(\theta)$, given a continuum of latent trait intensity or severity, θ. Dotted lines correspond to weighted (by prevalence) probability functions of negative (left) and positive (right) cases.

probability is thus equal to the integral of $f(\theta)p(\theta)^2$ over all levels of θ. Similarly, the probability of two positive ratings and one negative rating is equal to the integral of $f(\theta)p(\theta)^2[1 - p(\theta)]$ over all levels of θ. Generalizing this, the probability of exactly j positive ratings by k randomly selected raters is

$$\Pr[j' = j] = \binom{k}{j} \int f(\theta)p(\theta)^j[1 - p(\theta)]^{k-j}\, d\theta, \qquad (10)$$

where j' is a variable to denote the number of positive ratings. Multiplying $\Pr[j' = j]$ times the number of cases in an agreement study, N, gives the expected number of cases with exactly j positive ratings, e_j $(j = 0, 1, \ldots, k)$.

Knowing only $f(\theta)$ and $p(\theta)$, therefore, it is possible to predict the results of a rater agreement study. Conversely, given certain assumptions about their general forms, one can use the results of a study to estimate these functions. What we propose, therefore, is as follows: first, agreement data are used to estimate $f(\theta)$ and $p(\theta)$; then these functions are used to estimate rater accuracy and provide a basis for combining multiple ratings.

As a plausible way of approaching the initial estimation problem, we begin by assuming that there are two types of cases constituting a population, positive and negative cases, each normally distributed with respect to the trait continuum. Specifically, let $f_1(\theta)$ and $f_2(\theta)$ be normal distributions describing the unconditional probabilities of negative and positive cases, respectively, occurring at each level of θ, with $f_1(\theta)$ being defined by mean μ_1 and standard deviation σ_1, and $f_2(\theta)$ by μ_2 and σ_2. That is, $f_1(\theta)$ is the probability of sampling a case that is both negative and at trait level θ, and $f_2(\theta)$ is the probability of sampling a case that is both positive and at trait level θ. These functions are not probability density functions *per se*, since their integrals do not equal 1. Rather, they are the product of the probability density functions for negative and positive cases multiplied by their corresponding prevalences. The sum of these functions, $f(\theta) = f_1(\theta) + f_2(\theta)$, provides the latent trait probability density function.

Derivation of the Probability of Positive Rating Function

We now consider the function $p(\theta)$. Let each rater be assumed to have a rating *threshold*, or some point along θ such that cases with a trait level at or above this point are rated positive, and those below rated negative. In the varying panel case, let the thresholds of a population of raters be assumed normally distributed and described by the probability density function $t(\theta)$, with mean μ_t and standard deviation σ_t. The cumulative distribution function of $t(\theta)$ gives the probability that the threshold of a randomly selected rater is at or below each level of θ. This is equivalent to the probability of a case

with trait level θ equaling or exceeding the threshold of a randomly selected rater, and therefore being rated positive. Thus, this cumulative distribution function is equal to the probability of positive rating function, $p(\theta)$.

Estimation

We have therefore developed a model which provides the general form for $f(\theta)$ and $p(\theta)$. According to this model, $f(\theta)$ and $p(\theta)$ depend only on the means and standard deviations of $f_1(\theta)$ and $f_2(\theta)$ (μ_1, σ_1, μ_2, and σ_2), the mean and standard deviation of $t(\theta)$ (μ_t and σ_t), and the prevalences of positive and negative cases (which we designate P and $1 - P$, respectively). Only one prevalence must be estimated, since they sum to 1. Also, either of the means and either of the standard deviations for $f_1(\theta)$ and $f_2(\theta)$ can be chosen arbitrarily. Knowledge of as few as five parameters, therefore, allows estimation of $f(\theta)$ and $p(\theta)$.

For a set of parameter values, we may determine the probability of each number of positive ratings given k ratings per case with Equation (10). Given observed frequencies f_j ($j = 0, 1, \ldots, k$) for the number of cases with each number of positive ratings, we then calculate the log-likelihood of the joint outcome as

$$\log L = \sum_{j=0}^{k} f_j \log \Pr[j' = j]. \tag{11}$$

From the results of a rater agreement study, therefore, we may use numerical procedures to obtain maximum likelihood estimates for model parameters. Specifically, the maximum likelihood estimates of model parameters are those that maximize Equation (11). Uebersax [26] described the use of the Newton-Raphson method to obtain estimates for this model. For the Newton-Raphson procedure to converge effectively, however, it is usually necessary to apply an initial grid-search algorithm, which tests all combinations of parameter values using a

relatively coarse resolution, to find starting values in the vicinity of maximum likelihood estimates.[1]

As in the previous section, model fit may be assessed by comparison of observed and expected outcome frequencies using a χ^2 or L^2 test, with degrees of freedom equal to k minus the number of estimated parameters.

Estimating Rater Accuracy and Related Applications

Knowledge of $f(\theta)$ and $p(\theta)$ and their component parameters permits inferences concerning rating accuracy. For example, since the probability of a case at trait level θ being rated positive is $p(\theta)$, the conditional probability of a randomly selected positive case being rated positive, i.e., *Se*, is equal to the integral of $f_2(\theta)p(\theta)$ over all levels of θ, divided by P. Similarly, *Sp* is equal to the integral of $f_1(\theta)[1 - p(\theta)]$ over all levels of θ, divided by $1 - P$. Uebersax [26] shows similar formulas for positive and negative predictive validity.

Combining Multiple Opinions

Once estimated, $f(\theta)$ and $p(\theta)$ can also be used to classify cases based on multiple ratings. The probability of a case being positive, given j out of k positive ratings is

$$\Pr[+|j' = j] = \frac{\int f_2(\theta)p(\theta)^j[1 - p(\theta)]^{k-j}\, d\theta}{\int f(\theta)p(\theta)^j[1 - p(\theta)]^{k-j}\, d\theta}. \qquad (12)$$

We may also use Equation (12) with different values of k to derive classification rules for futures cases.

Uebersax [26] considers a computational example of the varying panel latent trait agreement model, so we do not present one here.

---

[1]Preliminary research suggests that it may be possible to eliminate the grid-search algorithm by the use of a "hybrid" estimation algorithm that combines the EM and Newton-Raphson methods.

FIXED RATING PANEL

Model

For this model, each rater is taken to have a characteristic threshold for making a positive rating. This threshold, however, is assumed subject to random variation, described by a normal probability distribution of values around a mean. The cumulative distribution function of this probability distribution gives the probability of a case at each level of θ equaling or exceeding the threshold of that rater, and a positive rating being made. Thus, associated with each rater j is a probability of positive rating function $p_j(\theta)$ having the shape of a normal cumulative distribution function and centered at the point on θ corresponding to that rater's mean threshold (Fig. 3.2). Following a standard technique in item response theory, we assume probability of positive rating functions to have the shapes of logistic ogives, which closely approximate normal cumulative distribution

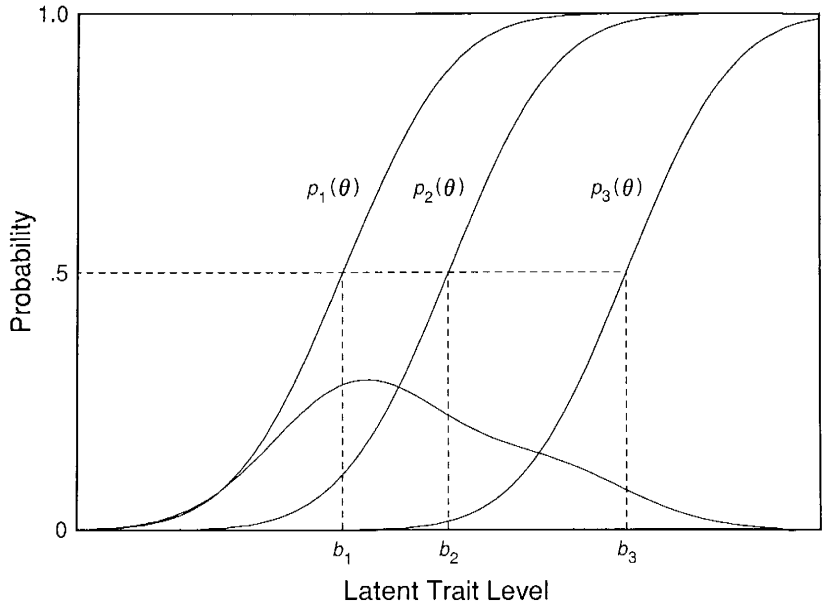

Fig. 3.2--Fixed panel latent trait agreement model: the probability of positive rating functions of three hypothetical raters are superimposed on the trait probability density function, $f(\theta)$; the b values correspond to each rater's mean threshold.

functions. The logistic function for each rater j depends on two parameters, a_j and b_j, which correspond to the variability and mean value, respectively, of that rater's threshold. Specifically, this function is defined as $p_j(\theta) = 1/\{1 + \exp[-1.7a_j(\theta - b_j)]\}$.

We define $f(\theta)$, $f_1(\theta)$, and $f_2(\theta)$ in the same way as for the varying panel model.

Estimation

As with the latent class fixed panel model, positive and negative ratings by k raters may correspond to one of $I = 2^k$ patterns. The probability of the ith such pattern, $\mathbf{u}_i$, occurring is

$$\Pr[\mathbf{v} = \mathbf{u}_i] = \int f(\theta) \prod_{j=1}^{k} p_j(\theta)^{u_{ij}} [1 - p_j(\theta)]^{1-u_{ij}} \, d\theta, \tag{13}$$

where $\mathbf{v}$ is the vector of observed ratings and u_{ij} again corresponds to the rating of the jth rater, coded 1 or 0.

The expected frequency of pattern i, e_i, is obtained by multiplying results of Equation (13) times the number of cases rated, N. The log-likelihood for the joint outcome of an agreement study is therefore

$$\log L = \sum_{i=1}^{I} f_i \log \Pr[\mathbf{v} = \mathbf{u}_i], \tag{14}$$

where f_i is defined as in the fixed panel latent class model.

Maximum likelihood estimates of parameters are again those that maximize log L and may be numerically obtained. If threshold variability is assumed constant across raters (though this is not an assumption one would make in all applications), the number of parameters necessary to estimate may be reduced to $k + 4$: a mean threshold (b_j) for each rater, within-rater threshold variability (a), the mean of either positive or negative cases on the latent trait continuum (μ_1 or μ_2), the standard deviation of either positive or negative cases (σ_1 or

σ_2), and the prevalence of positive cases (P). For unique estimates, this number must be less than or equal to I - 1. Model fit is assessed by comparison of observed and expected pattern frequencies with a χ^2 or L^2 test with degrees of freedom equal to I - 1 minus the number of estimated parameters.

Applications

The sensitivity, specificity, and positive and negative predictive validity of each rater's ratings can be estimated in the same way as with the varying panel model, using the individual probability of positive rating functions $p_j(\theta)$ in place of $p(\theta)$. These can also be averaged across raters to provide mean accuracy indices.

Parameter estimates can again be used to classify a case as positive or negative based on its ratings. For example, the joint probability of a case being positive and receiving pattern $\mathbf{u}_i$, which we denote $\Pr[\mathbf{u}_i, +]$, is obtained from Equation (13), using $f_2(\theta)$ in place of $f(\theta)$. The probability of a positive case given $\mathbf{u}_i$ is then equal to $\Pr[\mathbf{u}_i, +]$ divided by $\Pr[\mathbf{v} = \mathbf{u}_i]$.

Example

We illustrate this model with the hypothetical data in Table 3.1. These correspond to a study in which four diagnosticians rate 497 cases for presence or absence of a disorder. To reduce the number of estimated parameters, we assume $\sigma_1 = \sigma_2 = \sigma = 1$. We also assume $a_1 = a_2 = a_3 = a_4 = a$, i.e., that threshold variability is constant across raters. An arbitrary value of 0 is taken for μ_1. Thus, the parameters requiring estimation are μ_2, P, a, b_1, b_2, b_3, and b_4. Initial estimates are obtained by a grid-search algorithm. Using these as starting values, a Newton-Raphson algorithm provides the maximum likelihood estimates shown in Table 3.2.

Expected frequencies for each rating pattern given these estimates are shown in Table 3.1. Comparison of these with the observed frequencies results in values of 6.42 and 6.75 for χ^2 and L^2,

Table 3.1

RESULTS OF HYPOTHETICAL DIAGNOSTIC AGREEMENT STUDY

Rating Pattern i	Diagnostician 1	2	3	4	Observed Frequency f_i	Expected Frequency e_i
1	+	+	+	+	38	38.0
2	+	+	+	-	38	36.9
3	+	+	-	+	21	23.9
4	+	+	-	-	65	63.4
5	+	-	+	+	7	6.7
6	+	-	+	-	17	18.0
7	+	-	-	+	11	11.5
8	+	-	-	-	120	119.1
9	-	+	+	+	3	1.3
10	-	+	+	-	4	3.5
11	-	+	-	+	1	2.3
12	-	+	-	-	22	23.3
13	-	-	+	+	0	.6
14	-	-	+	-	5	6.7
15	-	-	-	+	7	4.2
16	-	-	-	-	138	137.9

Table 3.2

PARAMETER ESTIMATES FOR FIXED PANEL
LATENT TRAIT AGREEMENT MODEL

Parameter	Estimate	Standard Error
μ_2	2.92	1.17
P	0.35	0.08
a	1.65	0.55
b_1	0.08	0.26
b_2	1.66	0.67
b_3	2.88	1.01
b_4	3.32	1.19

respectively. With 15 - 7 = 8 df, these are both nonsignificant at the 0.5 level, indicating good fit.

From these parameter values, sensitivities for Raters 1 through 4 are estimated as 0.92, 0.74, 0.51, and 0.41, and specificities as 0.52, 0.81, 0.92, and 0.95. Estimated mean sensitivity and specificity across raters are 0.65 and 0.80, respectively.

DIRECTIONS FOR FUTURE RESEARCH

We have considered two latent distributions, one corresponding to negative and one to positive cases. However, it is possible to generalize this approach. For example, positive cases may consist of two subtypes, each normally distributed on the latent trait continuum. In some applications it might make sense to consider cases as following a single distribution [26]. There is also no need to require normal distributions; different parameterized distributional forms may also be considered.

We believe that significant improvements are possible for the estimation of these models. For example, marginal maximum likelihood estimation [53] may prove useful.

The question naturally arises of whether latent class or latent trait agreement models would be better for a given set of data. Ideally, both approaches could be used and a selection made on the basis of which provides better fit. However, although formal statistical methods for comparing the fit of nested models exist, there is no generally accepted method for comparing qualitatively different models; research in this area, though, is proceeding (see, for example, Ref. 54).

Because the estimation procedures and software are better developed for the latent class agreement model, we would generally recommend that investigators pursue that approach first.

Appendix A

SAMPLE INPUT FILE FOR FIXED PANEL LATENT CLASS AGREEMENT MODEL

The following shows an input file for estimating model M_3 of the Park et al. [45] treatment appropriateness rating data using the MLLSA latent class analysis program [36]:

```
Park et al. rating data--three latent classes
 5 3   859   150-.1      1 1              1   1
 2 2 2 2 2
FREE
  69 2 4 1 2 1 0 0 82 4 23 8 67 24 42 41
   0 0 0 0 0 0 0 0 0 5 0 8 8 5 28 49 386
.33      .33      .34
.90      .10      .60      .40      .10      .90
.90      .10      .60      .40      .10      .90
.90      .10      .60      .40      .10      .90
.90      .10      .60      .40      .10      .90
.90      .10      .60      .40      .10      .90
```

REFERENCES

1. Gelfand, A. E., and H. Solomon, "A Study of Poisson's Models for Jury Verdicts in Criminal and Civil Trials," *Journal of the American Statistical Association*, Vol. 68, pp. 271-278, 1973.

2. Lazarsfeld, P. F., and N. W. Henry, *Latent Structure Analysis*, Houghton-Mifflin, New York, 1968.

3. Goodman, L. A., "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika*, Vol. 61, pp. 215-231, 1974.

4. Haberman, S. J., *Qualitative Data Analysis: Vol. 2. Recent Developments*, Academic Press, New York, 1979.

5. Yerushalmy, J., "The Importance of Observer Error in the Interpretation of Photofluorograms and the Value of Multiple Ratings," *International Tuberculosis Yearbook*, Vol. 26, pp. 110-124, 1956.

6. Koran, L. M., "The Reliability of Clinical Methods, Data and Judgments: Part I," *New England Journal of Medicine*, Vol. 293, pp. 642-646, 1975.

7. Koran, L. M., "The Reliability of Clinical Methods, Data and Judgments: Part II," *New England Journal of Medicine*, Vol. 293, pp. 695-701, 1975.

8. Feinstein, A. R., "A Bibliography of Publications on Observer Variability," *Journal of Chronic Diseases*, Vol. 38, pp. 619-632, 1985.

9. Spitznagel, E. L., and J. E. Helzer, "A Proposed Solution to the Base Rate Problem in the Kappa Statistic," *Archives of General Psychiatry*, Vol. 42, pp. 725-728, 1985.

10. Sprott, D. A., and M. D. Vogel-Sprott, "The Use of the Log-Odds Ratio to Assess the Reliability of Dichotomous Questionnaire Data," *Applied Psychological Measurement*, Vol. 11, pp. 307-316, 1987.

11. Darroch, J. N., and P. I. McCloud, "Category Distinguishability and Observer Agreement," *Australian Journal of Statistics*, Vol. 28, pp. 371-388, 1986.

12. Zwick, R., "Another Look at Inter-Rater Agreement," *Psychological Bulletin*, Vol. 103, pp. 374-378, 1988.

13. Cohen, J., "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, Vol. 20, pp. 37-46, 1960.

14. Fleiss, J. L., "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin*, Vol. 76, pp. 378-381, 1971.

15. Carey, G., and I. I. Gottesman, "Reliability and Validity in Binary Ratings: Areas of Common Misunderstanding in Diagnosis and Symptom Ratings," *Archives of General Psychiatry*, Vol. 35, pp. 1454-1459, 1978.

16. Grove, W. M., N. C. Andreasen, P. McDonald-Scott, M. B. Keller, and R. W. Shapiro, "Reliability Studies in Psychiatric Diagnosis: Theory and Practice," *Archives of General Psychiatry*, Vol. 38, pp. 408-413, 1981.

17. Uebersax, J. S., "Diversity of Decision-Making Models and the Measurement of Interrater Agreement," *Psychological Bulletin*, Vol. 101, pp. 140-146, 1987.

18. Shrout, P. E., R. L. Spitzer, and J. L. Fleiss, "Quantification of Agreement on Psychiatric Diagnosis Revisited," *Archives of General Psychiatry*, Vol. 44, pp. 172-177, 1987.

19. Armitage, P., L. M. Blendis, and H. C. Smylie, "The Measurement of Observer Disagreement in the Recording of Signs," *Journal of the Royal Statistical Society*, Vol. 129, Series A, pp. 98-109, 1966.

20. Landis, J. R., and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, Vol. 33, pp. 159-174, 1977.

21. Landis, J. R., and G. G. Koch, "One-Way Components of Variance Model for Categorical Data," *Biometrics*, Vol. 33, pp. 671-679, 1977.

22. Lord, F. M., and M. R. Novick, *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, Massachusetts, 1968.

23. Gelfand, A. E., and H. Solomon, "Modeling Jury Verdicts in the American Legal System," *Journal of the American Statistical Association*, Vol. 69, pp. 32-37, 1974.

24. Kaye, K., "Estimating False Alarms and Missed Events from Interobserver Agreement: A Rationale," *Psychological Bulletin*, Vol. 88, pp. 456-468, 1980.

25. Kraemer, H. C., "Estimating False Alarms and Missed Events from Interobserver Agreement: Comment on Kaye," *Psychological Bulletin*, Vol. 92, pp. 749-754, 1983.

26. Uebersax, J. S., "Validity Inferences from Interobserver Agreement," *Psychological Bulletin*, Vol. 104, No. 3, pp. 405-416, 1988.

27. Stewart, G. W, and J. M. Rey, "A Partial Solution to the Base Rate Problem of the κ Statistic," *Archives of General Psychiatry*, Vol. 45, pp. 504-505, 1988.

28. Fleiss, J. L., and P. E. Shrout, "Reliability Considerations in Planning Diagnostic Validity Studies," in L. N. Robins and J. E. Barrett (Eds.), *The Validity of Psychiatric Diagnosis*, pp. 279-290, Raven Press, New York, 1989.

29. Tanner, M. A., and M. A. Young, "Modelling Agreement Among Raters," *Journal of the American Statistical Association*, Vol. 80, pp. 175-180, 1985.

30. Dempster, A. P., N. M. Laird, and D. M. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, Vol. 39, pp. 1-22, 1977.

31. Dawid, A. P., and A. M. Skene, "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm," *Applied Statistics*, Vol. 28, pp. 20-28, 1979.

32. Uebersax, J. S., *PANEL User's Manual*, The RAND Corporation, N-3030-RC, November 1989.

33. Dillon, W. R., and N. Mulani, "A Probabilistic Latent Class Model for Assessing Inter-Judge Reliability," *Multivariate Behavioral Research*, Vol. 19, pp. 438-458, 1984.

34. Espeland, M. A., and S. L. Handelman, "Using Latent Class Models to Characterize and Assess Relative Error in Discrete Measurements," *Biometrics*, Vol. 45, pp. 587-599, 1989.

35. Bentler, P. M., and D. G. Bonett, "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures," *Psychological Bulletin*, Vol. 88, No. 3, pp. 588-606, 1980.

36. Clogg, C. C., "Unrestricted and Restricted Maximum Likelihood Latent Structure Analysis: A Manual for Users," Working Paper 1977-09, Population Issues Research Office, Pennsylvania State University, 1977.

37. Kraemer, H. C., "Ramifications of a Population Model for κ as a Coefficient of Reliability," *Psychometrika*, Vol. 44, pp. 461-472, 1979.

38. Carroll, B. J., "Diagnostic Validity and Laboratory Studies: Rules of the Game," in L. N. Robins and J. E. Barrett (Eds.), *The Validity of Psychiatric Diagnosis*, pp. 229-244, Raven Press, New York, 1988.

39. Quade, D., P. A. Lachenbruch, F. S. Whaley, D. K. McClish, and R. W. Haley, "Effects of Misclassifications on Statistical Inferences in Epidemiology," *American Journal of Epidemiology*, Vol. 111, No. 5, pp. 503-515, 1980.

40. Bergan, J. R., "Latent-Class Models in Educational Research," in E. W. Gordon (Ed.), *Review of Research in Education 10*, pp. 305-360, American Educational Research Association, Washington, D.C., 1983.

41. Clogg, C. C., "Some Latent Structure Models for the Analysis of Likert-Type Data," *Social Science Research*, Vol. 8, pp. 287-301, 1979.

42. Uebersax, J. S., and W. M. Grove, "Latent Class Analysis of Diagnostic Agreement, *Statistics in Medicine*, in press.

43. Walter, S. D., and L. M. Irwig, "Estimation of Test Error Rates, Disease Prevalence and Relative Risk from Misclassified Data: A Review," *Journal of Clinical Epidemiology*, Vol. 41, No. 9, pp. 923-937, 1988.

44. van de Pol, F., R. Langeheine, and W. de Jong, *PANMARK User Manual*, Netherlands Central Bureau of Statistics, Voorburg, The Netherlands, 1989.

45. Park, R. E., A. Fink, R. H. Brook, M. R. Chassin, K. L. Kahn, N. J. Merrick, J. Kosecoff, and D. H. Solomon, *Physician Ratings of Appropriate Indications for Six Medical and Surgical Procedures*, The RAND Corporation, R-3280-CWF/HF/PMT/RWJ, July 1986.

46. Uebersax, J. S., "Latent Class Agreement Analysis with Varying Panels and Polytomous Ratings," unpublished manuscript, 1990.

47. Uebersax, J. S., "Latent Structure Modeling of Ordered Category Rating Agreement," paper presented at the annual meeting of the Psychometric Society, UCLA, Los Angeles, 1989 (The RAND Corporation, P-7597, November 1989).

48. Hulin, C. L., F. Drasgow, and C. K. Parsons, *Item Response Theory*, Dow Jones-Irwin, Homewood, Illinois, 1983.

49. Rasch, G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Second Edition, University of Chicago Press, Chicago, 1980.

50. Fleiss, J. L., "Estimating the Accuracy of Dichotomous Judgments," *Psychometrika*, Vol. 30, pp. 469-479, 1965.

51. Quinn, M. F., "Relation of Observer Agreement to Accuracy According to a Two-Receiver Signal Detection Model of Diagnosis," *Medical Decision Making*, Vol. 9, No. 3, pp. 196-206, 1989.

52. Swets, J. A., "Measuring the Accuracy of Diagnostic Systems," *Science*, Vol. 240, pp. 1285-1293, 1988.

53. Bock, R. D., and M. Aitkin, "Marginal Maximum Likelihood Estimation of Item Parameters: Application of the EM Algorithm," *Psychometrika*, Vol. 46, pp. 443-459, 1981.

54. Gibbons, R. D., and M. A. Young, "Comparison of Discrete and Continuous Latent Structures," Paper submitted for publication, 1989.