Chapter 3
Interdiagnostician Agreement and Diagnostic Validity

First Draft, July 30, 1992
©1992 William M. Grove

In classical psychometrics (see, for example, Lord & Novick, 1968, chapters 2-16), reliability and validity are treated as distinct topics, though they are admittedly related. However, modern test theory (known as latent trait theory) employs models of test scores that unify these concepts. Similarly, statisticians occupied with problems of diagnosis originally thought of diagnostic reliability as purely a problem of interobserver agreement. Nowadays, however, they are more likely to consider more complex, but more realistic models, in which validity considerations are explicitly incorporated.

In this chapter, we will begin by considering traditional interobserver agreement measures, because they build a foundation for what comes later on. We begin by considering the situation in which there are only two diagnostic judges and a single, dichotomous diagnosis. This will motivate the development of methods for using data on interrater agreement as way to infer the accuracy of diagnostic procedures.
We discuss Cohen’s coefficient kappa, the most commonly used measure of interrater agreement in diagnostic studies. We develop the rationale for this measure, discuss its properties, and relate it to epidemiological concepts of disease prevalence and diagnostic accuracy.

From this we proceed to consideration of models explicitly incorporating diagnostic accuracy statistics, including latent class and latent trait models. We show how agreement data can allow one to estimate diagnostic validity.

Finally, we give a worked example. We use data from a twin study of schizophrenia.

Measuring Interobserver Agreement

The most obvious measure of the agreement between two diagnosticians their frequency of agreement, usually expressed as a percentage of the total number of cases jointly examined. Figures over 90% have historically been regarded as signifying good to excellent agreement.

However, there is a serious problem with this procedure. Suppose that two diagnosticians are presented with a relatively rare condition in their clinic, say disorganized schizophrenia. If only 5% of the patients in the clinic actually suffer from this condition, then 95% diagnostic accuracy, and 100% interdiagnostician agreement, could be achieved
by the simple method of having both judges *never* make the diagnosis
in question. It is also the case that if diagnosticians were to assign
diagnoses at random, with each diagnostician giving 95% of cases a
negative diagnosis, then the diagnosticians would agree 90.5% of the time,
just by coincidence. Obviously, then observed agreement of 90% might not
indicate good diagnostic precision at all, under such circumstances.

There are two settings in which medical or psychiatric or
psychological conditions tend to have low point prevalences (or base rates;
we shall generally use these two terms interchangeably). The first is in
epidemiological studies, where diagnostic categories may be broad (e.g.,
affective disorders, schizophrenia, alcohol abuse/dependence) but most
individuals in the population are free of *any* disorder. In such studies,
reliabilities represented as percent agreements may be misleading even
when high, since most of the agreement stems from consensual judgments
that the disorder in question is absent.

The second situation in which prevalences can be low is when
diagnoses are made in a clinic, but the diagnostic categories are narrowly
drawn, or when intrinsically rare conditions are studied. An example of
narrowly drawn diagnostic categories would be subtypes of schizophrenia,
or types of specific eye disease on an ophthalmologic unit. An example of
an intrinsically rare condition would be 'pure paranoia' in psychiatry, or most recessive genetic conditions seen on medical genetics units.

Coefficient Kappa

It was to obviate problems like this that Cohen (1960) devised coefficient kappa. This is defined as

\[ \kappa = \frac{p_o - p_c}{1 - p_c} \]  

where

\( p_o \) is the observed percentage agreement, expressed as a decimal fraction, and

\( p_c \) is a correction for the degree of agreement to be expected by chance (see below).

The division by \( 1 - p_c \) serves to normalize the coefficient so that its maximal value is unity, representing perfect agreement. \( p_c \) can exceed \( p_o \) so that negative estimated values of \( \kappa \) can occur.

Note that \( p_o \) is an easily calculated quantity. However, \( p_c \) is "the agreement expected by chance." It turns out that many quantities could be substituted for \( p_c \), depending on how one imagines "chance" to be operating. From each different conception of "chance", a different \( p_c \) will of course be derived. In turn, a different \( \kappa \) will be obtained. Upon
Mathematics of Diagnosis—Chapter 3

reflection, it becomes clear that one’s notion of “chance” implies a schematic account of how the diagnostic process is carried out.

By far the most popular choice for $p_c$ in Equation 1 is

$$p_c = \hat{P}_1\hat{P}_2 + \hat{Q}_1\hat{Q}_2$$

[2]

where

$\hat{P}_1$ is the observed prevalence of the diagnosis in question for Observer 1,

$\hat{P}_2$ is the observed prevalence of the diagnosis in question for Observer 2,

$\hat{Q}_1 \equiv 1 - \hat{P}_1$, and

$\hat{Q}_2 \equiv 1 - \hat{P}_2$.

Here we introduce the notation, used below throughout, that $\hat{\theta}$ is an estimator of $\theta$, whether unbiased, consistent, maximum likelihood, or what not, and $X \equiv Y$ means “‘X’ is defined as Y”.

This choice of $p_c$ is simple and has a strong intuitive justification. It says that chance agreement occurs when judges each assign diagnoses independently and at random (i.e., without regard for the characteristics of the subject), and where each judge has his or her own characteristic predeliction ($P_i, i = 1, 2$) for making the diagnosis in question. Then, clearly, $p_c$ is the proportion of diagnoses which would be in agreement under such a chance set-up.
Other choices of $p_c$ could be defended as well. One might take notice of the fact that in many reliability studies judges are given stacks of cases to look over, half of which are presumed to have the diagnosis in question, and half not. Alternatively, judges are told (truly or falsely) that the cases are $1/2$ $D$ and $1/2$ not-$D$. In that case, it would seem plausible that the judges would assign diagnosis, if they did so by blind chance, with

$$p_c = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$$

so that

$$\kappa = \frac{p_o}{1 - 1/2} = 2p_o. \quad [3]$$

regardless of the empirical frequencies with which judges actually assign diagnoses. However, it is obvious that the employment of a value of $p_c$ other than $(P_1 \times P_2)(Q_1 \times Q_2)$ is easiest to defend when judges are forced to match their marginal relative frequencies of the diagnosis in question to a specified frequency.

There is a moral in the ambiguity of $p_c$. Its ambiguity directly reflects the lack of a coherent theoretical model of how diagnoses are actually made. The concept of “chance agreement” is really a vaguely articulated recognition that diagnosticians can agree on account of either of two different operating mechanisms: either they agree because they are both
correct, or they agree because they have both made the same mistake. Unfortunately, $p_c$ and hence $\kappa$ does not separate these two sources of agreement. This vagueness will be eliminated in the more complex models to be presented below.

Relation of Kappa to Other Corrected Agreement Indices

Fleiss (1975) pointed out that, when corrected for chance agreement, a wide variety of normalized agreement indices take the form

$$I_x = \frac{I_o - I_c}{\text{max}(I_o) - I_c} \quad [4]$$

where

$I_o$ is an observed agreement statistic of some kind and

$I_c$ is a chance correction,

(Actually, Fleiss assumes $\text{max}(I_o) = 1$ whereas it may happen that the true bound is lower for some statistics.) For wide variety of agreement indices defined on $2 \times 2$ tables, Fleiss (1975) then showed that, after being corrected for chance agreement, they reduce either to $\kappa$ or to the Pearson product-moment correlation coefficient defined for dichotomous variables ($\phi$). This is a remarkable and useful result.

Conger and Ward (1984) extend this result by noting that $\phi$ has a maximum value less than unity whenever the $2 \times 2$ table marginals are unequal, i.e., when $\hat{P}_1 \neq \hat{P}_2$. Similarly, the maximum value of $\kappa$ is
likewise less than unity under the same circumstances. For a fixed set of table marginals one can compute the maximum possible values of $\kappa$ and $\phi$, namely $\kappa_{\text{max}}$ and $\phi_{\text{max}}$ respectively. Conger and Ward then showed that, for a fixed set of table marginals, it is always the case that $\kappa/\kappa_{\text{max}} = \phi/\phi_{\text{max}}$.

Indeed, it is further the case that when the observed margins of the $2 \times 2$ table are identical (i.e., the table is symmetric, $\hat{P}_1 = \hat{P}_2$), then $\phi = \kappa$. Hence a broad class of agreement statistics are numerically identical to $\kappa$. This sort of numerical convergence is one powerful justification for the popularity of coefficient $\kappa$.

Variations on Kappa:

More Than Two Judges, More Than Two Categories

We have considered only the simple case of a single dichotomous diagnosis and two judges. More complex versions of kappa, generalizing to more diagnoses and/or more judges, have also been developed. These situations are discussed by Fleiss (1971, 1979, 1981). Most of these complications do not affect the developments here, so they are not further discussed in this book. However, it is important to deal with the situation in which multiple judges diagnose the same cases.
Kappa for Multiple Judges

Under Different Conceptions of “Agreement”

When more than two diagnosticians judge individuals' illnesses, the formula given in Equation 1 dealing with observed agreement \( (p_o) \) becomes ambiguous. When two raters give the same diagnosis, this is obviously an agreement and hence contributes to \( p_o \). But consider the following alternative scenario for a study with seven diagnosticians. Suppose that for a randomly chosen individual, five diagnosticians give a positive diagnosis and two a negative one. Is this an agreement or a disagreement?

It is possible to consider such agreements (or disagreements) in several ways. These are conveniently distinguished by their usual labels “complete” agreement (or DeMoivre agreement), “majority” agreement, and pairwise agreement. In the first, raters in our sample scenario are deemed not to have agreed, since they did not all agree. Under the second conception of agreement, there are five agreements and two disagreements, since five agreed with the majority opinion while two did not. Under the third conception of agreement, there are \( \binom{7}{2} = 21 \) pairwise comparisons possible. \( \binom{5}{2} + \binom{2}{2} = 11 \) of these are agreements and the remaining 10 are disagreements.
Different agreement coefficients and in particular different $\kappa$ formulas can be worked out for each of these definitions of agreement. We illustrate by providing an extension of $\kappa$ to cover majority agreement.

To deal with majority agreement, first formulate the case of DeMoivre (consensus) agreement. Re-express $\kappa$ in terms of disagreements rather than percent agreements (Light, 1971). One then obtains a general expression for any multiple-judge $\kappa_m$ as

$$\kappa_m = 1 - \frac{d_o}{d_c}$$

[5]

where

$d_o$ is the observed number of disagreements and

$d_c$ is the number of disagreements expected by chance.

Conger (1980) gives an elegant formula for $\kappa_m$ for agreement defined in terms of what he calls $g$-wise agreement, which encompasses all definitions of agreement we have listed. When $g = J$, i.e. agreement occurs only when the number of agreeing judges equals the total number of judges, then we have consensus, complete, or DeMoivre agreement. When $g = 2$ we have the usual pairwise agreement. When $g = J/2$ (rounded up if necessary) we have majority agreement. First, Conger establishes that, for
pairwise agreement,

\[ d_o = N \sum_{j < f} N p_{X_{ij} \neq X_{ij'}} \prod_{j' < j''} (N^2 p_{j' \cdot p_{j''}} + N^2 q_{j' \cdot q_{j''}}) \]

\[ d_e = \sum_{\theta < \theta'} (N^2 p_{j' \cdot p_{j''}} + N^2 q_{j' \cdot q_{j''}}) \prod_{j' < j''} (N^2 p_{j' \cdot p_{j''}} + N^2 q_{j' \cdot q_{j''}}) \]  \[ \text{[6]} \]

where

\( p_{X_{ij} \neq X_{ij'}} \) is the observed proportion of pairwise disagreements (i.e., instances in which the diagnostic scores \( X_{ij} \) and \( X_{ij'} \) for the \( i \)th patient do not coincide,

\( p_j \) is the proportion of positive diagnoses given by the \( j \)th diagnostician, and

\( q_j \) is the proportion of negative diagnoses given by the \( j \)th diagnostician.

This rather imposing-looking equation simplifies remarkably (Conger, 1980, last equation on p. 323):

\[ \kappa_m = \frac{\sum_{j < j'} K_{j j'}}{\binom{f}{2}} \]  \[ \text{[7]} \]

i.e., \( \kappa_m \) for pairwise agreement equals the average of all possible ordinary \( \kappa_s \) computed between pairs of judges.
This was then extended to cover general $g$-wise agreement as follows.

$$g_{p_o} = \sum_j n_{(1j)(2j)\ldots(gj)}/n$$

$$g_{p_c} = \sum_j n_{(1j)n_{(2j)}\ldots n_{(gj)}}/n^g$$

$$g_\kappa = \frac{g_{p_o} - g_{p_c}}{1 - g_{p_c}}$$

$$g_{\kappa_m} = \sum \frac{g_\kappa}{\binom{J}{g}}$$  [8]

The leading letter $g$ in all these symbols emphasizes the dependence of the definition of the statistic on the requirement of $g$-way agreement.

The term $n_{(1j)(2j)\ldots(gj)}$ is the number of $g$-way agreements involving rater $j$, while the term $(n_{(1j)n_{(2j)}\ldots n_{(gj)}})/n^g$ computes the expected number of such agreements, under the assumption of stochastic rater independence.

The summation in the last line of Equation 8 is over all possible $g$-way combinations of $J$ diagnosticians. The last line shows that the summary $g_{\kappa_m}$ is an average $\kappa$, just as it was for pairwise agreements, except this time it is for $g$-wise agreements.

It would seem that this approach would be computationally cumbersome, as it involves tabulating all possible $g$-sets of rater agreements and disagreements. However, Tanner and Young (1985) give a simpler approach using generalized linear models for contingency table
data, in which the specification of scoring for $g$-tuples is much more convenient.

Landis and Koch (1977b) give an alternative approach to majority agreement which is even less complex than the Conger (1980) approach. One simply creates an indicator score for each patient. The indicator score is 1 if and only if the patient is diagnosed positively by a majority of the diagnosticians, and 0 otherwise. ("Majority" can be defined in terms of plurality or some specific kind of supermajority, as circumstances warrant. Landis and Koch give an example where 5 of 7 diagnosticians constitutes a "majority.") Then the simple pairwise $\kappa$ for each diagnostician's score with the indicator score is computed over all patients. Since the data are not reduced to a single summary $\kappa$, it becomes obvious which raters, if any, frequently disagree with the majority opinion.

There is a major conceptual problem with the idea of majority agreement. If diagnosticians are fallible, then a simple majority or even supermajority diagnosis can be wrong. We would then be testing the "ability" of individual diagnosticians to pick the wrong answer to the diagnostic question. Certainly, if the diagnosticians in question are reasonably skilled and a sufficiently large supermajority is required, the wrong diagnosis is unlikely to be the majority one very often.
On the other hand, when a large supermajority is required for a majority diagnosis, another pressing problem emerges. If a supermajority is required both for positive and for negative diagnoses, then a number of subjects (perhaps many of them) will be given no majority diagnosis. What are we to do with them? If we throw them out of the reliability study *post hoc*, we will bias the agreement statistics upward. This is because we are dropping precisely the most difficult-to-classify individuals, on whom we already know agreement was least.

On the other hand, if we leave in subjects with no assigned majority diagnosis, we include a subgroup of subjects of uncertain scoring. When is a diagnostician to be considered to have agreed with the majority diagnosis of a subject, when that subject has no such majority diagnosis? Something arbitrary must be done about scoring “correctness” of diagnoses for these subjects. One obvious possibility would be to score all diagnosticians’ choices as correct for these subjects. But this too would bias the agreement statistics upward. Another possibility would be to score no diagnostican’s choices as correct; but this biases the statistics downward.

It seems that the study of majority agreement offers little of advantage. We consider pairwise agreement to be the basic form of Interrater agreement of interest, with one single exception.
Here is the exception. Suppose that one diagnostician is considered a master in the craft and is taken, for the purposes of certain studies, to be infallible. Unlike the (super)majority agreement approach, we assume that the master diagnostican can provide a definitive diagnosis for each and every patient. Study of agreement with such a diagnostican is called “agreement with a known standard” (Wackerly, McClave & Rao, 1978; Tanner & Young, 1985). This master diagnostican may not be an individual examiner at all—it may be the result of culling all available information on the case to arrive at a diagnosis which is presumed, on best available evidence, to be correct in all cases. Spitzer (1983) refer to this situation in psychiatry as the L.E.A.D. system, in punning reference to the wished for but non-existent “gold standard” diagnosis in psychiatric research. The L.E.A.D. system incorporates Longitudinal data, Expert opinion, and All available information into the Diagnosis. When such a diagnosis is available, it may be rationally presumed for some purposes to be tantamount to having a known-to-be-correct diagnosis.

In a situation like this, the study of interdiagnostician agreement becomes the study of diagnostician accuracy, albeit by proxy (since the L.E.A.D. diagnosis is not known to be truly infallible). Tanner and Young (1985) give a generalized linear model treatment of this problem, for two important schemes. In one, a standard diagnosis is available for all $N$
subjects, but different raters example different subsets of the total $N$. In the other, all raters provide diagnoses for all $N$ subjects, on whom a standard diagnosis is also available.

However, in any of these situations involving a standard diagnosis, agreement between raters (except when one is the "master" or standard rater) is irrelevant except insofar as it reflects agreement with the standard. Therefore, the formulation of agreement statistics is typically of little interest. Instead, one would ordinarily compute statistics that directly reflect the accuracy of diagnosticians in assigning positive and negative diagnoses to individuals. This problem is covered in detail below in the section on signal detection theory and diagnosis. Therefore, we postpone further discussion of it for now.

For further developments relating to interrater agreement as opposed to rater accuracy, it proves to be exceedingly inconvenient to consider any other definition of agreement than pairwise. Therefore, from now on we make the following assumptions. First, no rater is considered to be privileged in accuracy; there is no "gold standard" of diagnosis available in our studies. In fact, even the L.E.A.D. diagnosis of Spitzer et al. is generally unavailable. Second, no special properties are accorded to majority, supermajority, or absolutely complete agreement between raters.
More Complicated Interrater Reliability Study Designs

In almost all of the interrater agreement study designs above, it has been assumed that all judges diagnose all cases. This makes the computations and the theory simple. However, it is not always very realistic. Instead, we can imagine a number of possible interrater reliability study designs, such as ones in which

- Raters are paired with each other at random;
- Raters are paired with each other in some systematic fashion, such as a balanced incomplete blocks design (BIBD) (i.e., all judge pairs occur an equal number of times);
- The number of judges per subject varies, with each ratee being rated by different numbers of judges (typically, no two subjects are rated by the same judges);
- and so on through an infinitude of potential designs.

While \( \kappa \) has been extended to cover a few of the above cases, such as varying numbers of randomly chosen judges per subject (Fleiss & Cuzick, 1979), in general such complex designs have no directly defined \( \kappa \) statistic. In particular, choosing a reasonable formula for \( p_c \) may give great difficulty.
One way around this difficulty is by analogizing $\kappa$ to a better-understood statistic that is defined over a broad range of study designs. One such statistic that turns out to be very convenient is the intraclass correlation coefficient (ICC). This statistic, which at first bears little resemblance to $\kappa$, can be defined for our purpose as follows. Let $X$ be the assigned diagnostic variable (1 if and only if the diagnosis is positive, 0 otherwise) and let $D$ correspondingly denote the value of the true diagnosis (likewise scored 0 or 1). Then we can define

$$\rho \equiv \frac{\sigma_X^2}{\sigma_D^2}$$  \[9\]

where

$\rho$ is the population value of the intraclass correlation coefficient;

$\sigma_X^2$ is the variance of the assigned diagnostic variable; and

$\sigma_D^2$ is the variance of the true diagnostic variable, namely

$$\sigma_D^2 = P(1 - P).$$  \[10\]

$\rho$ therefore has the qualities of a “percentage of variance accounted for” statistic. As such, it can only (in the population, absent sampling variation), take on values between zero and unity. (ICCs estimated from sample data can assume values less than zero.) Zero for an ICC means chance-level agreement, while one means perfect agreement.
A Generalizability Study Approach to Kappa

Through Intraclass Correlation Coefficients

Lindquist (1953) and later on, Cronbach, Rajaratnam, and Glaser (1963) showed that psychometric reliability statistics have an intimate connection with the analysis of variance. In this approach, judges (or raters) form one factor of an analysis of variance (ANOVA) design, and subjects or patients (often called “targets” in the articles and book of Cronbach et al.) form the other crossed factor, thus yielding a two-way Judges × Targets ANOVA.

For balanced ANOVAs, it turns out to be quite simple to compute an appropriate ICC. Bartko (1976) discusses various ICCs that may be useful to compute from such designs. We confine our attention to just one of these, namely the reliability of a single randomly chosen judgment from a single randomly chosen judge, which Bartko calls “ICC(1)”.

Suppose one has conducted an ANOVA in which every one of $J$ judges rates every one of $N$ targets. Then this $J \times N$ two-way ANOVA yields three mean squares, corresponding to each source of variation: targets (subjects or patients), judges, and error. Then the ICC is estimated
Mathematics of Diagnosis—Chapter 3

consistently (albeit not unbiasedly) by the computational formula

\[
\hat{\rho} = \text{ICC}(1) = \frac{(MS_T - MS_E)/J}{(MS_T - MS_E)/J + (MS_J - MS_E)/N + MS_E}
\]

\[
= \frac{(J - 1)(SS_T - SS_E)}{(J - 1)SS_T + J \frac{N-1}{N}(SS_T + SS_E) - SS_E}
\]  \[11\]

where

\(MS_T = SS_T/(N - 1)\) is the mean square for targets (subjects),

\(MS_J = SS_J/(J - 1)\) is the mean square for judges, and

\(MS_E = SS_E/(J - 1)(N - 1)\) is the mean square for residuals from the model (error).

In addition to yielding easy computational formulae for ICCs, this work paves the way for a remarkable mathematical convergence with \(\kappa\). In fact, Fleiss and Cohen (1973) proved that, for two raters \((J = 2\) in the above),

\[
\kappa = \frac{SS_T - SS_E}{SS_T + 2SS_J + SS_E} = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}_J^2 + \hat{\sigma}_E^2 + \frac{1}{N-1}(\hat{\sigma}_J^2 + \hat{\sigma}_E^2)} \approx \hat{\rho} = \text{ICC}(1) \]  \[12\]

where \(\hat{\sigma}^2\) denotes an unbiased estimator of the corresponding variance component \(\sigma^2\). That is, as \(N\), the number of subjects, becomes large, \(\kappa\) and \(\text{ICC}(1)\) for the same design become numerically indistinguishable.

Since \(N\) is usually pretty large, say over 50, in most reliability studies the
slight difference between $\kappa$ and ICC can usually be neglected. In the sequel we rely on the approximation $\kappa \approx \tilde{\rho}$.

What about designs in which not both $\kappa$ and the ICC can even be computed? Consider for example the following actual study design employed by Andreasen, Grove, Shapiro et al. (1981) to estimate inter-rater reliability in the NIMH Clinical Research Branch–Collaborative Studies on the Psychobiology of Depression–Clinical (Karz, Secunda, Hirschfeld et al., 1979). This study of approximately 1000 patients with affective disorder included samples collected at five centers. It was essential, prior to pooling samples for data analysis, to know whether diagnoses were being made to similar standards at all five sites.

A reliability study was performed as follows. At each of the five sites, a master rater was responsible for training diagnosticians, hence this person was regarded as having particular skill at interviewing and diagnosis (based on extensive training and experience, of course). Each master rater (one from each site) was sent to a common site, located in a major urban center, where each interviewed 10 locally recruited study subjects. Interviewers were kept blind to subjects' previously obtained diagnoses. $\binom{5}{2}$ rater pairs were possible, and in fact each such rater pair occurred twice.
This design was then repeated at another, less urban site, as a check on whether urban-nonurban origin of subjects made a difference to diagnoses. As it did not seem to, data from the two sites were pooled for analysis. Therefore, the study was to contain four occurrences of each possible rater pair. This layout constitutes a balanced incomplete blocks design (BIBD). In such a design each subject need only be interviewed twice, and yet data on all possible interrater differences in diagnostic habits are obtained. It was considered impractical to have each subject interviewed separately by five different diagnosticians.

There is no computational formula for $\kappa$ for such a BIBD. However, the ICC is well defined, and can be written schematically as

$$\rho \equiv \frac{\sigma_T^2}{\sigma_T^2 + \sigma_j^2 + \sigma_E^2}$$  \hfill [13]

where

$\rho$ is the population value of the ICC,

$\sigma_T^2$ is the variance among target

$\sigma_j^2$ is the variance among judges,

$\sigma_T^2$ is the variance among targets, and

$\sigma_E^2$ is the variance of residuals from this model.

This quantity is consistently (but not unbiasedly) estimated by plugging in unbiased estimates of the variance components:

$$\hat{\rho} \equiv \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}_j^2 + \hat{\sigma}_E^2}$$  \hfill [14]
which estimates can themselves be obtained in a variety of ways. (Estimation of variance components is itself a complex subject. See Searle, Casella and McCulloch, 1992, for a complete discussion.) Relying on Equation 12 for the essential equivalence between \( \kappa \) and ICC, we can use the obtained ICC as a \( \kappa \) even for such a design as this.

The ICC approach lends itself to “broken” designs. It so happens that the design described above was literally broken, in that one judge was incapacitated in mid-study by a skiing-induced leg fracture. Substituting for the disabled rater led to imperfections in the design as actually carried out. Yet the computational formulae used above in Equation 10 still applied without change. (Computational advise is given for such models in Grove and Hollon, 1991). Further computational methods for assessing reliability are given by Uebersax (1982) for arbitrarily complex designs.

Problems with Kappa

And Controversial Alternatives To It

Kappa has an, according to some, inconvenient feature. The proficiency of the diagnosticians can remain constant, and yet \( \kappa \) varies with the base rate of the condition being diagnosed (Grove, Andreasen, McDonald-Scott et al., 1981). This is shown by the following example: suppose two diagnosticians are each 95% accurate. Then when \( P = .5 \) (the most favorable case), their \( \kappa \) is 0.81. However, if these two diagnosticians
move to a clinic where \( P = 1/4 \), retaining their 95% accuracy, then \( \kappa = 0.76 \).

Should they then move to another clinic in which \( P = .01 \), then \( \kappa = 0.14 \)!

To some statisticians, this state of affairs is quite unobjectionable. They point out that the trait variance (on the “trait” of diagnosis, scored 0 or 1) is smaller in a clinic where \( P \) is small, and hence the restriction of range reduces the interrater correlation indexed by \( \kappa \). To other workers in this area, however, this variation of \( \kappa \) with changing \( P \) represents an undesirable feature.

In fact, the dependence of \( \kappa \) on base rates sometimes creates considerable practical problems. First, despite warnings from such authorities as Shrout, Spitzer and Fleiss (1987), \( \kappa \)'s tend to acquire a patina of diagnostic accuracy instead of merely one of diagnostic agreement. Second, results reported for epidemiologic studies tend to have \( P \) in the range 0.1% – 5% for individual disorders. When such studies are published in the same journals as hospital-based studies (where often \( .25 \leq P \leq .75 \)), readers are likely to look askance at the reliabilities obtained in the epidemiologic studies. In fact, major authors in this area have given guidelines for the interpretation of \( \kappa \) without mention of the dependence of \( \kappa \) on prevalence. For example, Fleiss (1981) advises that \( \kappa \)s between .40 and .75 are “fair,” while those over .75 are “good”. Landis and Koch (1977a) label .21-.4 as “fair”, .41-.6 as “moderate”, .61-.8 as
“substantial” and .81-1.0 as “almost perfect”. This kind of advice fosters this tendency to think of $\kappa$s in isolation from base rates.

Spitznagel and Helzer (1985) have documented the dependence of $\kappa$, for fixed diagnostic accuracies, on base rates. They show that as $P$ approaches or dips below about 10%, the dependence of $\kappa$ on $P$ becomes obvious, and below about 5% it becomes quite problematic. A debate ensued as to whether $\kappa$ was being blamed as the messenger of bad tidings, i.e. simply informing the researcher that, in this population with these base rates, reliability was not very high (a position taken by Shrout et al., 1987) and by Kraemer), or whether instead $\kappa$ ought properly to be indicted for this (mis)behavior.

Spitznagel and Helzer, who were among those troubled by this feature of $\kappa$, proposed another agreement statistic. This is Yule’s coefficient of concordance, commonly called $Y$. Yule’s $Y$ is defined simply as $(\sqrt{ad} - \sqrt{bc})/(\sqrt{ad} + \sqrt{bc})$ where $a, b, c,$ and $d$ are defined in Table 1; $a = p_{++}$ is, for example, the proportion of individuals on whom both judges agree on a positive diagnosis. A graph (Figure 1) shows the relative behavior of the $\kappa$ and Yule’s $Y$ statistics, for the interval from $P = .5$ to $P = .001$ for a typical set of values for diagnostic accuracy in psychiatric research. Obviously, the $Y$ statistic is less dependent on $P$ than is $\kappa$ until $P$ is lower than seen even in typical epidemiological studies. In typical
clinical studies, the two statistics would yield nearly numerically identical values and so there would be little to choose between them.

Table 1
Agreement Between Two Judges

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ +</td>
<td>a = p_{++} b = p_{-+}</td>
</tr>
<tr>
<td>+ -</td>
<td>c = p_{+-} d = p_{--}</td>
</tr>
</tbody>
</table>

Insert Figure 1 about here

What are we to make of this controversy over the behavior of $\kappa$?

Because of the many mathematical ties between $\kappa$ and ICCs and also because of Equation 4, it seems that $\kappa$ is the agreement measure of choice for diagnostic studies. However, Spitznagel and Helzer (1985) still have a point. To the extent that one wishes to think of $\kappa$ as a diagnostic accuracy measure, then base rates must be considered.

The basic problem is that $\kappa$ depends on base rates and on diagnostic accuracy. In fact, no measure that condenses four cells' worth of agreement information (rater 1 positive vs. negative $\times$ rater 2 positive vs. negative) can do anything but confound base rates and accuracy. There is
simply no way to reduce all these data to one number (whether $\kappa$ or Yule's $Y$) without loss of information. It would seem that a wiser course is to use the data to look at accuracy and base rates separately.

Recall that, intuitively, two judges can reach agreement in either of two ways: by both being right, or by both making the same mistake. This means that one can decompose the probability $p_{++}$ (the probability of the judges agreeing on the presence of the disorder $D$) as follows:

$$p_{++} = P \Pr\{X_1 = 1, X_2 = 1 \mid D = 1\} + Q \Pr\{X_1 = 1, X_2 = 1 \mid D = 0\} \quad [15]$$

where

$X_j$ is the diagnosis assigned by judge $j$,

$X_j \mid D = d$ is the diagnosis assigned by judge $j$, conditional on the subject being a member of true disease class $d$, and

$\Pr\{\cdot\}$ denotes a probability. The variable $D$ is defined as in Equation 6 above.

The first term in Equation 15 above quantifies agreement when the judges are both right, and the latter when they are both wrong. The other four cells of the $2 \times 2$ agreement table, denoted by $p_{+-}, p_{-+},$ and $p_{--}$ are defined in parallel fashion, and were given explicitly in Table 1.
Tying Kappa to a Model of Diagnostic Validity

These conceptions tie directly into a relatively simple model of diagnostic validity. Either a diagnosis is right or it is wrong, and this is not a matter of degree. When Diagnostician 1 makes a mistake, this doesn't in any way influence (or even predict) the probability that Diagnostician 2 will also make a mistake. Judges make their diagnoses not only experimentally independently (i.e., they don't peek at each others' ratings), but also stochastically independently.

Let us now make this intuitive idea a quantitative model. Think of a “true score” $D$ assigned by Nature to the patient's illness, for example 0 for disease absent and 1 for disease present, as before. There are also assigned diagnoses, called “observed scores” in the psychometric terminology, assigned by judge 1 and judge 2 in a reliability situation, which will also be 0 and 1. However, in classical psychometric theory, it is the case that, by definition,

$$X_T = E[X] \quad \text{(16)}$$

so that

$$X = X_T + \epsilon \quad \text{(17)}$$

from which it follows that

$$E[\epsilon] = 0.$$
where $E[\cdot]$ denotes mathematical expectation and $X_T$ is the true score.

This simple and useful relation cannot hold for the true and assigned diagnostic scores as defined above. Therefore, $D \neq X_T$. This is so because diagnostic errors of omission and commission do not generally cancel. Therefore, a person's true diagnosis is not the average of an infinite number of repeated diagnoses by clinicians (Klein & Cleary, 1967).

If we cannot rely on the large body of psychometric theory developed for mental test scores, it might seem that we are at a loss. However, epidemiologists define a number of concepts which are very helpful in specifying the relations between the variables $D$ and $X$. Epidemiologists refer to individuals who truly have the disorder in question (i.e., individuals with $D = 1$) as "cases" and those with $D = 0$ as "noncases". "Noncases" may be ill but they do not have the diagnosis of interest.

Based on this fundamental dichotomy, epidemiologists define three important quantities. First, there is the true prevalence of a diagnosis in a given population, (in our notation $P$, with complement $Q = 1 - P$). Next, there is the rate at which cases are correctly diagnosed — this is called the sensitivity of a diagnostic process, or of a diagnostician, according to context. We will denote this by the Greek letter $\alpha$. There is a corresponding probability for noncases; the rate at which they are correctly called noncases is the specificity, which we denote by $\beta$. 
Useful consequences directly follow. For example,

\[
\hat{P} = P\alpha + Q(1 - \beta)
\]

where

\(\hat{P}\) means “an estimator of \(P\).”

That is, a group of individuals diagnosed as having a certain disorder comprise two kinds: individuals who are correctly called cases, and noncases who are misdiagnosed as cases. Now, in general,

\[
P \neq P\alpha + Q(1 - \beta)
\]

and so \(\hat{P}\) is a biased estimator of \(P\). That is, \(E[\hat{P}] \neq P\). This is one reason why we choose to surmount the estimator \(\hat{P}\) with a tilde. We do not use the usual notation for an estimator, which would be \(\hat{P}\), or \(P\) surmounted by a “hat.” The hat notation is understood by some readers to denote an unbiased estimator. \(\hat{P}\) is merely a very simple estimator of \(P\), often the best one can do.
Kappa for Equally Proficient Raters
As a Function of Base Rates and Validities

Kraemer (1979) proved a very enlightening theorem tying Cohen’s \( \kappa \) to the epidemiologists’ concepts of base rate, sensitivity, and specificity. When there is a latent diagnostic dichotomy like the one we have been considering and when equally accurate judges independently assign diagnoses dichotomously as well, she showed that

\[
\kappa = \frac{PQ}{\hat{P}\hat{Q}}(\alpha + \beta - 1)^2
\]

[19]

Again, note that

\( P \) is the epidemiologist’s prevalence,
\( \alpha \) is what they call sensitivity, and
\( \beta \) is what they call specificity.

Inspection of this formula is instructive. The leading term \( PQ/\hat{P}\hat{Q} \) is always less than one (unless diagnoses are perfectly accurate), because \( \hat{P} \) is biased away from \( P \) and toward 1/2. Thus, this term penalizes \( \kappa \) for the fact that, as shown implicitly in Equation 18 above, the judges misestimate the base rate of the disorder in question.

The quantity \( (\alpha + \beta - 1) \) can be understood as follows. Recall the dummy variable \( X \) we introduced before; it took on the value 1 if and
only if a case was diagnosed as having the illness in question, and zero otherwise. It turns out to be the case that

\[ \alpha + \beta - 1 = (\mathbb{E}[X] \mid D = 1) - (\mathbb{E}[X] \mid D = 0) \equiv \Delta_X \]  \hspace{1cm} [20]

which means that the quantity \((\alpha + \beta - 1)\) measures the validity of the diagnostic procedure. That is, \(\Delta_X\) gives the expected difference in \(X\) between the two disease classes “case” and “noncase” when \(P = 1/2\).

We see then that \(\kappa\) is roughly proportional to the squared validity of the diagnostic procedure. The squaring of the validity term occurs because there are two judges, invalidity of either of whose diagnoses attenuates interrater agreement.

Kappa for Unequally Proficient Judges

As a Function of Base Rates and Validities

A parallel formula can be easily derived for two judges who are unequally accurate. This is

\[ \kappa = \frac{PQ}{\sqrt{\hat{P}_1 \hat{Q}_1 \hat{P}_2 \hat{Q}_2}} \frac{PQ}{(\alpha_1 + \beta_1 - 1)(\alpha_2 + \beta_2 - 1)} \]  \hspace{1cm} [21]

However, in psychiatric diagnostic studies conducted since the advent of specified diagnostic criteria (e.g., Research Diagnostic Criteria of Spitzer, Endicott, and Robins, 1978), judges operate from a rigid set of diagnostic rules and are generally carefully trained and supervised. This presumably
minimizes differences in diagnostic accuracy between them. Later we will look at an empirical example where it is quite reasonable to posit wide disparities between judges' performances.

The problem with Equation 19 is that it gives $\kappa$ in terms of $P, \alpha,$ and $\beta$. However, we already know how to obtain $\kappa$ from the observed and the chance-expected levels of agreement. What we would like better is to obtain, from a knowledge of $\kappa$, the values of the epidemiologists' parameters $P, \alpha,$ and $\beta$. Unfortunately, while the latter parameters determine $\kappa$, the converse is not true; there is no way to solve Equation 19 backwards.

Elimination of Parameters for Special Situations

There are special situations in which one may be justified in effectively eliminating a parameter from the set \{P, $\alpha$, $\beta$\}. In such a situation, it may be possible uniquely to estimate the remaining parameters. For example, in a study by Rice, McDonald-Scott, Endicott, Coryell, Grove, Keller and Altis (1986), relatives of patients with affective disorder were studied at two points in time five years apart. Suppose now that the threshold for diagnosing the disorder of interest (bipolar II disorder, which is depression plus hypomanic episodes) is set very high. For example, we might require five positive symptoms to make the diagnosis, whereas DSM-III-R requires only three. Or we might require
that there be three documented episodes or more, rather than only one. In any case, if the threshold is set high enough, it may be the case that we are then entitled to assume that, effectively, $\beta = 1$, i.e. there are no false positive diagnoses. Rice et al. pointed out that in this special case, a consistent estimator of the true base rate $P$ is given by

$$\hat{P} = \frac{\hat{P}^2}{\hat{p}_{++}} \quad [22]$$

and the specificity is consistently estimated by

$$\hat{\alpha} = \frac{\hat{p}_{++}}{\hat{P}} \quad [23]$$

The denominator (numerator) of the right hand side of Equation 22 (Equation 23), $\hat{p}_{++}$ is just the observed relative frequency with which individuals obtain positive diagnoses on both interviews.

To justify both of these estimators 22 and 23, we must also assume that the prevalence of the diagnosis is stable from interview 1 to interview 2 five years later, i.e. $\hat{P}_1 = \hat{P}_2$. We also assume that the sensitivity of the diagnostic process is likewise identical at times 1 and 2.

Rice et al. give parallel estimators for the special situations in which one can assume that $\alpha = 1$ or that $\alpha = \beta$. However, the above is the most empirically reasonable situation arising in psychiatric research, where
diagnostic criteria tend to be narrowly drawn, favoring specificity at the expense of sensitivity.

Alas, we are not generally justified in assuming either that sensitivity or specificity is perfect, or that these two parameters are equal. Nevertheless, it turns out to be the case that this difficulty is not insurmountable. To estimate sensitivities, specificities, and prevalence in general situations, we need some more elaborate models of interrater agreement to estimate prevalence, sensitivity, and specificity, however. We now turn to such models.

**Formal Models of Diagnosis**

**Pure Category Models:**

**Latent Class Theory with Local Independence**

The equations above (Equations 18-23) also can be justified by deriving them from a psychometric model for the diagnostic process. Doing this paves the way for still more sophisticated diagnostic models, to be presented below.

Suppose that there exist two classes of individuals as before, cases and noncases. Case/noncase status is not directly observed but must be inferred from fallible indicators like signs and symptoms. Assigned diagnoses resulting from a diagnostic process are therefore
only statistically rather than deterministically related to latent class membership. Let us (for the moment) assume the following.

- Each individual who is truly a case is equally likely to be mislabelled as a noncase.
- Each individual who is truly a noncase is equally likely to be mislabelled as a case.
- Judges are all equally accurate.
- Judges make their diagnoses with experimental independence. This means that judges do not confer or directly influence one another.

Under these assumptions, the following will hold:

- There is a constant probability of misclassification for cases, namely $1 - \alpha$,
- There is a constant probability of misclassification for noncases, namely $1 - \beta$, and
- Judges’ diagnoses will be stochastically independent, conditional upon true (latent) case/noncase status. (Stochastic independence refers to the property that one judge’s diagnosis does not predict the others, among those truly from a common latent disease class.)

This set of assumptions defines a special case of what is called a latent class model with local independence (Lazarsfeld & Henry, 1968;
Goodman, 1974; Titterington, Smith & Makov, 1985). “Local independence” refers to the third bulleted item, viz. the stochastic independence of diagnoses conditional upon latent case/noncase status. These models are used in psychology, sociology, and other fields such as marketing research.

One of the principal advantages of such models is that they allow one to formulate agreement between pairs of judges in terms of the judges' underlying (but unobserved) diagnostic accuracies, much as in Equation 19. In principle, we can then reason backwards from the observed agreements to the judges' accuracies, which we could not do using Equation 19 alone. One could then use such parameter estimates for a number of purposes. For example, one could decide whether a given diagnostic procedure is sufficiently accurate. Under some such latent class models one could decide whether certain judges were particularly accurate, and if so, identify outstanding (or dismal) raters. Finally, one could potentially use the parameter estimates to reshape the diagnostic process. This would produce a revised procedure better than the initial one, a process which could in principle be repeated. We will give examples of such uses below.
One can not only estimate the parameters of such latent class models. One can also test hypotheses about them, or put confidence intervals around them, as well. There are several methods for doing this.

We consider here only the problem of estimating the model parameters for an interrater agreement study in which all individuals are diagnosed by all raters. Lazarsfeld and Henry (1968) suggested a computational method that relied on the sample moments of the data, i.e. the first-order marginals of the interrater agreement contingency table plus the two-way tables, three-way tables, ..., on up to the full $J$-way table obtained for a set of $J$ judges. However, their computational method was inefficient (in the statistical sense: it did not use all the information in the data), and is no longer used.

We now derive estimators of latent class model parameters. We use the method of maximum likelihood which is explained below.

If one provisionally assumes such a latent class model to be correct, then one can write out a probability model for an individual obtaining a positive diagnosis. Take for a start the situation in which there is just one judge:

\[
\Pr \{X = 1 \} = \alpha^D + (1 - \beta)^{1-D}
\]
when $D = 1$ and
\[
\Pr\{X = 1\} = (1 - \alpha)^{1-D} + \beta^D
\]
when $D = 0$. For a randomly chosen subject, the probability of a positive diagnosis is the weighted mixture of these probabilities
\[
\Pr\{X = 1\} = P[\alpha^D + (1 - \beta)^{1-D}] + Q[(1 - \alpha)^{1-D} + \beta^D] \tag{24}
\]

For $J > 2$ judges we need some more notation. Let $X_{ij}$ be the diagnosis given by the $j$th judge ($j = 1, \ldots, J$) to the $i$th subject ($i = 1, \ldots, N$). Let $\alpha_j$ and $\beta_j$ be the sensitivity and specificity for the $j$th judge, respectively. Then
\[
\Pr\{X = x\} = P \prod_{j=1}^{J} \alpha_j^{x_{ij}}(1 - \alpha_j)^{1-x_{ij}} + Q \prod_{j=1}^{J} (1 - \beta_j)^{x_{ij}}\beta_j^{1-x_{ij}} \tag{25}
\]

For the important special case in which all judges have sensitivities equal to a common constant $\alpha$ and specificities equal to a common constant $\beta$ (the equal-accuracy case) this reduces to
\[
\Pr\{X = x\} = \binom{N}{r_i} P\alpha^{r_i}(1 - \alpha)^{J-r_i} + Q(1 - \beta)^{r_i}\beta^{J-r_i} \tag{26}
\]

where
\[r_i = \sum_{j=1}^{J} x_{ij} \text{ is the number of positive diagnoses given to the } i\text{th subject.}

Now we consider all $N$ subjects in a diagnostic study. Let $X = (X_{ij})$ be an $N \times J$ matrix of diagnoses with its $(i,j)$th element equal to the
diagnostic dummy variable assigned to the \( i \)th subject by the \( j \)th judge. The logarithm of the likelihood function for this data set is given by

\[
\log L = \text{const.} + \sum_{i=1}^{N} (P \prod_{j=1}^{J} \alpha_j^{X_i^j} (1 - \alpha_j)^{1 - X_i^j} + Q \prod_{j=1}^{J} (1 - \beta_j)^{X_i^j} \beta_j^{1 - X_i^j})
\]  [27]

since, for a fixed set of data, the leading term (the multinomial coefficient corresponding to the binomial coefficient in Equation 26) is a constant. Such a constant can be ignored in maximizing \( L \) (or equivalently \( \log L \)). That part of the likelihood which omits the constant term is called the \textit{kernel} of the likelihood.

We can use this likelihood for parameter estimation via the method of maximum likelihood. For such estimation it is most convenient to work with the log likelihood rather than the likelihood itself. Since one is a monotonic function of the other, when the maximum of the log likelihood has been achieved then the likelihood itself is also maximized. In the Appendix we briefly explain maximum likelihood estimation from basic principles to computational details, for the benefit of readers unfamiliar with this subject.

In the models in this section, we have assumed that all diagnosticians examine all patients. This assumption is not always correct. However, the models proposed here can be extended to varying rater panel data. Computationally, these are handled by a microcomputer program called
PANEL (Uebersax, 1989) whose use is described by Uebersax and Grove (1990).

At the end of this chapter we give a worked example of latent class analysis applied to diagnostic agreement. The example concerns the diagnosis of schizophrenia in a famous twin study and relies on a fixed panel of five renowned diagnosticians.

**Pure Category Models:**

**Latent Class Models with Intermediate Classes**

In the foregoing we have assumed that there are just two latent categories, cases and non-cases. It is possible to imagine more complex situations. In the most obvious, consider a typical clinical setting in which many different diagnoses are made at least occasionally, where one of the diagnoses can of course be “no diagnosis” or “well.” In such a situation, one might wish to model the multiple-category agreement of diagnosticians with a view to estimating not only their abilities to tell cases from well individuals, but also their skill at differential diagnosis.

Kappa coefficients have been described for such situations (Fleiss, 1981). Latent class models can be constructed in which there are three or more latent classes corresponding to two or more diagnostic categories plus the “well” category. However, it seems that there is little advantage to doing so. Such models can always be mimicked by having a separate
two-category model (disordered with a particular disorder $X$, versus not so disordered) for each disorder of interest. Therefore, we do not treat such models here.

However, there is another situation in which multiple latent categories have been postulated that is of considerable interest, since it motivates much of the development to follow. Many clinical disorders share the characteristic of possessing so-called “sub-clinical” states or mild cases, as well as the more typical “textbook” or full-blown case. It is reasonable to suppose that diagnosticians have less trouble agreeing on the labelling of a “textbook” case than a “gray” or intermediate case. This point was first made quantitatively by Yerushalmy (1947) and developed more fully by Neyman (1947). However, this idea lay unnoticed in the rater agreement literature until rediscovered by Maxwell (1977).

Maxwell proposed a new coefficient of agreement for such diagnostic situations, which as he pointed out are the rule rather than the exception in psychiatry. His coefficient explicitly assumes that gray cases are assigned to the “well” class and to the textbook class with conditional probabilities equal to $1/2$ each, while textbook cases are correctly diagnosed with $\alpha = 1$. With this simplifying assumption, and $\beta$ can be uniquely estimated. This coefficient was criticized by (Dewey, 1983) because of its arbitrary underlying assumptions.
Latent class models can deal with intermediate categories in a more general and flexible manner. If this intermediate-class situation is instead modelled by a latent class analysis with only two categories, then the assumption of local independence, referred to above, is violated. This is because of the mixture of gray and textbook cases. The mixture generates a correlation between ratings (see Figure 2). As a result, all the parameter estimates for base rates, sensitivities, and specificities can be biased, sometimes strongly so. Figure 2 shows a cluster of textbook cases which are nearly certain, on average, to be correctly diagnosed and another group of gray cases which have only an average 1/2 probability of being detected. The line on the graph represents the least-squares line through the data, and clearly demonstrates that judgments are *not* independent, since the mixture of the two types of cases generates a covariance between judges.

Insert Figure 2 about here

One can at least approximately restore the validity of the local independence assumption by revising the two-category model to incorporate a third, intermediate category. However, to do this requires special care. Two-category models describe a nominally scaled latent variable. A model having an intermediate class explicitly assumes that the
propensity of gray cases to be diagnosed as ill is intermediate between
that of non-cases and textbook cases, and conversely that the probability
of intermediate cases to be diagnosed as not ill is likewise intermediate
between textbook cases and non-cases. That is, the latent variable is
ordinally, not nominally, scaled. Standard latent class models, on the
other hand, assume that the classes are only nominally scaled (i.e., possess
category labels with only arbitrary ordering). Therefore, standard methods
of estimating latent class model parameters have to appropriately
modified to correctly handle this situation.

The simplest way to modify the latent class model parameter
estimation algorithms is to introduce constraints. While this method only
works for studies in which at least four diagnosticians have participated, it
may nonetheless prove useful.

Consider for example the situation in which all diagnosticians are
equally skilled. We identify two sensitivities, one for gray cases ($\alpha_g$)
and the other for textbook cases ($\alpha_t$), and likewise two corresponding
specificities. At each iteration $i$ of the estimation procedure, we examine
$\alpha^i_g$ in relation to $\alpha^i_t$. If $\alpha^i_g > \alpha^i_t$ (i.e., if a gray case is estimated to be more
likely to be called a case than is a textbook case) then we set $\alpha^{i+1}_g = \alpha^{i+1}_t$
(or $\alpha^{i+1}_g = \alpha^{i+1}_t - \epsilon$ for some arbitrary small constant $\epsilon$). From that
point on in the iteration process, $\alpha_g$ is not treated as an independently
estimable parameter but as an alias for \( \alpha_t \) (or for \( \alpha_t - \epsilon \)). The specificity parameters are treated in a parallel fashion. Fixing parameters during iteration in maximum likelihood estimation is done in any case, e.g., when the estimated value of a probability would fall outside the interval [0, 1]. Here we are simply fixing one parameter equal not to a constant (e.g., 0 or 1) but to another, freely estimable parameter.

It is important to note that the fixing of \( \alpha \) or \( \beta \) during estimation affects the degrees of freedom used in statistical tests. Consonant with the discussion of maximum likelihood estimation above, for each such parameter that is fixed, degrees of freedom are generally reduced by 1.

There is another way to think about the “gray case” situation. Upon reflection, the hypothesized situation seems not quite realistic. Is it really plausible that clinical cases come in two grades and two grades only—gray and textbook? It would seem that it is instead more realistic to assume the existence of a continuum of “textbook-ness”, extending from diseased individuals in whom signs of illness are nonetheless indetectable, through individuals so severely and typically affected that their illnesses are reliably detected by all but the most obtuse diagnosticians. We now consider such continuum models.
Pure Continuum Models:

Latent Trait Theory

To motivate the work in the next sections, we now briefly explain a model from modern psychometrics called item characteristic curve theory or logistic latent trait theory (see Lord & Novick [1968] for a comprehensive discussion of this theory). This theory was developed to account for the behavior of individual psychological test items, which take the place of individual diagnosticians as we have been considering here. There is an latent ability measured by the test (corresponding to the concept of "caseness" above) usually denoted in this literature by $\theta$. The receiver operating characteristic curve for a diagnostician from signal detection theory is precisely paralleled by an item characteristic curve, one for each item $j = 1, \ldots, J$ on the test. It is generally assumed at the outset that the shape of this curve is known. Typically it is assumed that the item characteristic curve has the form of the cumulative distribution function of the logistic distribution:

$$
\Pr\{X_j = 1 \mid \theta\} = \psi(\theta) = \frac{\exp(1.7[b_j + a_j \theta])}{1 + \exp(1.7[b_j + a_j \theta])}, \quad j = 1, \ldots, J
$$

[28]

where $b_j$ represents a measure of the probability of getting item $X_j$ correct
(i.e. obtaining a positive diagnosis) for an individual of average ability (caseness),

\(a_j\) represents the capacity of item \(X_j\) to discriminate grades of the latent ability (caseness) \(\theta\), and

the constant 1.7 is chosen to make the numerical details agree with a similar model in which the logistic cdf is replaced by the Gaussian cdf. (If the constant 1.6 had been used then the mean of the logistic cdf would have been the same as a Gaussian cdf [cumulative density function]. If the constant 1.8 had been used then the variance would have been equated. 1.7 is a compromise that produces at most 0.01 difference between the logistic and Gaussian pdfs [probability density functions].)

In this model it is implicitly assumed in typical applications that the latent distribution of an ability (corresponding to our “caseness” variable) is unimodally distributed. This is quite unlike signal detection theory’s bimodal latent distribution; hence the choice of a logistic cdf.

In fact, in the Rasch-type model, if even one diagnostician has a logistic item characteristic curve, then every other item characteristic curve is thereby forced to be logistic in shape as well. Lord and Novick (1968, chapter 8) show for the Rasch model that any distribution of the latent ability or “caseness” distribution desired can then be obtained by suitable nonlinear transformation of that key diagnostician’s (item’s)
characteristic curve. Therefore, the shape of the latent ability or caseness distribution is to a large extent arbitrary once the form of the item characteristic curve has been determined. This implies that there may be only a loose relationship between the latent caseness distribution and the observed counts of number of diagnoses positive.

It is theoretically possible to obtain observed distributions (number of diagnoses positive) which are multimodal or even bathtub-shaped. Such results can theoretically be obtained by choosing extremely high values for all of the parameters $a_j$, and simultaneously clustering the $b_j$ closely about some central value. In the mental testing literature, this amounts to asking the examinee (almost) the exact question, over and over again. In such a situation, if examinees are consistent in their responses, their score distribution will end up looking bathtub-shaped—some examinees will have said “no” again and again while others will have said “yes” consistently, with few examinees in the middle of the score distribution (Lord & Novick, 1968, chapter 16, Figure 16.14.3).

However, such situations are extremely unlikely to occur in the field of medical and psychiatric diagnosis, because our diagnostic tools are simply not that discriminating. It therefore turns out to be the case that the latent continuum of caseness, in logistic latent trait theory, is more
or less forced to be unimodal by the specification of a logistic cdf for the
diagnostician’s item characteristic curve.

If it so happens that \( a_j = a, j = 1, \ldots, J \), i.e. that all items
diagnosticians) are equally discriminating, then this is called a Rasch
model (Rasch, 1960). In this case, all the information about the judges’
accuracy is contained in the number of subjects garnering a given number
of positive diagnoses \((0, 1, \ldots, J)\) as in Equation 26 above. That is, the
number of diagnoses positive is a sufficient statistic for \( \theta \). Uebersax and
Grove (1990) give the maximum likelihood estimates for this situation.

How is one to choose between latent variable models which assume
a nominal scale, like latent class analysis, and ones which assume a
continuous scale, like logistic latent trait analysis. Gibbons and Young
(1992) have taken up this important problem. They show that latent class
analysis is identical to logistic latent trait theory if the latent classes are
two in number and certain restrictions are imposed. This allows one to
decide empirically which provides a better account of rating data. The
reader is referred to their paper for a fuller exposition.

A major problem with Equation 28 is that it does not allow for
variation among diagnosticians in their diagnostic thresholds nor in their
abilities to discriminate cases from non-cases. Moreover, in assuming
a logistic cdf, a unimodal underlying caseness distribution is also more
or less assumed (see above). This is reasonable as a distribution of ability for many psychological tests (e.g., intelligence tests, tests of subject area knowledge), but is unrealistic for many medical applications. Therefore, we now turn to models based on the item characteristic curve but explicitly incorporating multimodal latent caseness distributions.

Maximum likelihood of the parameters of the Rasch and other logistic latent trait models is discussed in Lord (1980). However, we are using these models only as a stepping-stone for more realistic approaches to the diagnostic situation. Therefore, we do not pursue quantitative issues with the logistic latent trait model here. Instead, we turn to another class of models that is a still closer approximation to what we seek. This approximation is provided by signal detection theory.

Pure Continuum Models:

Signal Detection Theory

There are two ways to think about multimodal continuum models: a signal detection theory approach, and a multimensional parametric distributional approach. The former makes fewer strong assumptions about the distributional properties of the underlying continuum, and so for generality’s sake this is the approach followed here. A very useful reference covering signal detection theory as applied to diagnosis is Swets and Picket (1982).
The name “signal detection theory” suggests the transmission of information, and early applications in this area concerned problems involving the ability of listeners to distinguish different consonant sounds (e.g., “buh” versus “duh”), which were used as signal and noise events, sent over a telephone line.

The terminology of signal detection theory therefore differs from the epidemiologist's, though they really consider precisely the same quantities. What we have called a case, signal detection theorists call a “signal”; our non-case becomes their “noise”. Instead of considering just two probabilities of making positive or negative diagnoses (our sensitivity and specificity) they have separate names for all four cells of an hypothetical table in which the columns are the true case/non-case status and the rows are the assigned diagnosis.

A “hit” is an event in which a case is correctly diagnosed, a “miss” is an event in which a case is incorrectly diagnosed, a “false alarm” is an event in which a non-case is incorrectly called a case, and a “correct rejection” is an event in which a non-case is correctly diagnosed. These are, of course, just $p_{++}$, $p_{+-}$, $p_{-+}$, and $p_{--}$ in Table 2. The leading subscript denotes the true diagnosis known to the experimenter, while the other subscript denotes the diagnosis according to the judge under test.
Thus far, we have considered true disease status to be an unknown. However, in signal detection experiments, the experimenter knows to which of two classes a given stimulus belongs. In their terminology, signal/noise status is known. For example, the experimenter knows whether “duh” (signal) or “buh” (noise) was sent over the line. This allows the experimenter to directly estimate hit, miss, false alarm, and correction rejection rates. Therefore, in applying signal detection theory to diagnosis, we assume *ad argumentum* that the true disease status of each diagnosed subject is known.

Now, this assumption is rarely met in research in psychiatry or clinical psychology, though it may be quite tenable for a number of diseases that occur in internal medicine, for which quasi-definitive laboratory tests are available. In the next section, we will generalize signal
detection theory to deal with situations in which true disease status is unknown.

As was stated above, in signal detection theory a latent continuum is assumed. This continuum is assumed to have a bimodal distribution, with the upper mode corresponding to signals (cases) and the lower one to noise events (non-cases). (See Figure 3 for an example.) Signal detection or diagnosis is presumed to operate as follows: each judge has a characteristic point on the latent continuum at which their probability of making a positive diagnosis is $1/2$; this point is called $\beta$ in the signal detection literature. This is (not our $\beta$, by which we have denoted the specificity of the diagnostic process. To keep the notation clear, we will refer to signal detection $\beta$ as $T$ in this book, since it represents a diagnostic threshold. Each judge (or all judges in common) also has a characteristic ability to separate signal from noise, or cases from non-cases. This discriminating ability is measured by $d'$. Since we have already noted this parameter above by $\Delta$ we will continue to do so here for consistency.

Insert Figure 3 about here
For a given shape of (bimodal) latent continuum distribution, the specification of signal detection $T$ and $A$ parameters uniquely specifies the expected frequencies with which judges will make accurate and inaccurate diagnoses of positive and negative kinds. It also implicitly specifies the expected rates of diagnostic agreement between judges.

In the signal detection literature, it is commonly taken as a convenient auxiliary assumption that the latent distributions making up the modes for signal and noise are each Gaussian. However, it is essential to note that this assumption is not an integral part of the theory, and estimates of signal detection parameters have been proposed which do not make an assumption of normal distributions. Figure 4 gives the probability of a positive diagnosis as a function of caseness in the form of a normal ogive. This would be approximately the form expected in the logistic latent trait theory of the previous section and would look slightly different for a signal detection model. However, this curve can actually have any one of a number of functional forms. In typical applications, the diagnostician’s item characteristic curve will proceed more or less smoothly from the lower left to the upper right corner of the graph, monotonically increasing and ordinarily showing asymptotes in the corners.
In signal detection theory, a very useful plot summarizing all the information in a diagnosticians’s performance is given by the receiver operating characteristic (ROC) curve. An example curve is given in Figure 5. Instead of assuming a fixed $T$ for the diagnostician, this graph displays the whole range of possible diagnosticians performances as his/her diagnostic threshold varies from quite strict to quite lenient. Since, in the signal detection model, the case and non-case distributions overlap all along the continuum of caseness (see Figure 3), there is a natural trade-off between sensitivity and specificity.

At the lower left corner of the graph, the threshold is very strict, leading to a low rate both of false positives and of true positives. As the threshold moves along the continuum of caseness, note that at first the sensitivity is increased considerably without much injury to the specificity. As caseness increases further, it is no longer possible to boost sensitivity without appreciable cost in declining specificity. Finally, at the top right of the graph, almost everyone gets a positive diagnosis, and so sensitivity is quite high but only at the cost of very low specificity.
The diagonal line in Figure 5 represents a diagnostician who does no better than chance. The area under the ROC succinctly summarizes the total performance of a diagnostician at all possible thresholds $T$ and is denoted by $A_z$ (Swets, 1986). $A_z$ is the most commonly used measure of diagnostic proficiency in the signal detection literature. It is calculated as follows. Figure 6 shows Figure 5, replotted on binominal axes. If the ROC in Figure 5 is a normal ogive, the replotted ROC in Figure 6 will then be a straight line. In that case, one can compute

$$
\mu_{sn} = \mathbb{E}[X \mid X > T]
$$

$$
\Delta' = \frac{\mu_{sn} - \mu_n}{\sigma_n}
$$

$$
z(A) = \frac{\text{slope } \Delta'}{\sqrt{1 + \text{slope}}} n
$$

$$
A_z = \Phi(z(A))
$$

[29]

where

$\mu_{sn}$ is the mean of all observations lying above the decision threshold $T$ on the latent continuum (including “signal” + “noise”),

$\mu_n$ is the mean of the latent distribution of “noise” (non-case) observations,

$\Delta'$ is the difference between these two means,

$\sigma_n$ is the standard deviation of the latent distribution of non-cases,
“slope” is the slope of the ROC on the binormal graph, and
\( \Phi(\cdot) \) is the cumulative normal distribution function.

Note that we distinguish \( \Delta' \) above from our parameter \( \Delta \). Our \( \Delta \) is the quantity \( (\mu_s - \mu_n) \) whereas \( \Delta' \) is the mean between all observations above a cutoff (which includes some “noise” or non-case observations) and the non-case distribution mean.

Insert Figure 6 about here

\( A_x \) is the area between the ROC curve and the diagonal line in Figure 5. As it goes up, the ROC curve more closely approaches a step function at specificity 1, i.e. perfect sensitivity without any sacrifice of sensitivity. This measure \( A_x \) is independent of the decision criterion (diagnostic threshold) actually employed by diagnosticians, and is also unaffected by disorder prevalence. These are all reasons why \( A_x \) is the preferred measure of diagnostic excellence in signal detection theory.

Unfortunately, we cannot directly compute \( A_x \) since we lack an essential piece of information available to the signal detection experimenter. Unlike such experimenters, who can present “buh” and “duh” over a noisy telephone line and know what they are presenting, we are unable to present diagnosticians with cases and non-cases about whose class assignments we are literally certain.
However, all is not lost. In the next section, we will finally arrive at our Grail: models in which, under favorable circumstances, we can start with data on interrater agreement and infer our way to high-confidence assignments of individuals to case and non-case categories. We will also be able to infer some accuracy data for individual diagnosticians along the way, though we will not be able to infer the whole ROC curve for a diagnostician.

However, before we do this, it is interesting tie signal detection theory to the $\kappa$ coefficient, from which we started. It turns out that there are noteworthy relationships between signal detection theory parameters and $\kappa$-type agreement coefficients.

Suppose that the ROC for diagnosticians has the double stair-step shape shown in Figure 7. On the lowest step, sensitivity is nil but specificity is perfect. On the highest step, sensitivity is perfect but specificity is nil. In the middle, diagnoses are assigned at random, without respect to where they lie on the caseness continuum. It turns out that this is exactly the sort of ROC that yields a coefficient of agreement of $\kappa$ between a diagnostician and the true diagnosis (Swets, 1986). It is therefore obviously the case that the appeal of $\kappa$ (as an index of diagnostic
accuracy against the true diagnosis, not as an index of interobserver agreement) depends on the plausibility of this stair-step ROC.

Swets (1986) shows convincingly that in many medical decision making problems, an ROC much more like the one in Figure 5 generally obtains. A stair-step ROC would be extremely unusual and indeed has apparently never been reported to date with empirical data. Therefore, \( \kappa \) by implication has little to offer as an index of diagnostic accuracy. However, this is not to say that it is an unreasonable index of diagnostic agreement, when a definitive diagnosis for accuracy determination is unavailable.

We pointed out above that signal detection theory assumes something we may not assume, namely that the true diagnosis is known to the experimenter in a diagnostic study. It would seem that the applicability of signal detection theory to interdiagnostician agreement problems would be nil.

However, recall that it is ordinarily the case that the diagnosticians under study have appreciable ability not only to agree with one another, but also to correctly discriminate cases from non-cases. When this is true, functions of the number of positive diagnoses a patient obtains offer a (perhaps crude) measure of where on the caseness continuum a patient’s
condition lies. Alternatively, recall that diagnostic criteria sets like DSM-III-R often rely on symptom counts. It seems eminently reasonable to suppose that the number of symptoms an individual patient has positive shows a (perhaps crude) relationship to that individual’s position on the latent caseness continuum.

These considerations establish the feasibility, in principle, of applying signal detection theory-like models of diagnosis even to situations where a definitive diagnosis is unavailable. The next two sections of this chapter do just this.

Hybrid Continuum-Category Models

Latent Trait Finite Mixture Model

For Fixed Rater Panel

Examine Figure 3 again. Suppose that the latent distribution of caseness is as portrayed, namely a mixture of symmetric unimodal distributions, with support along the real line, mixed in some proportion \( P : Q \). Then we can write the density of the left component as \( g_1(\theta) \) and the right-hand component as \( g_2(\theta) \). (We will assume throughout that the right-hand component is the one containing cases, and that it scores higher on any dimension we may be considering.) Then the unconditional
density in Figure 3 is given by

\[ f(\theta) = Qg_1(\theta) + Pg_2(\theta) \]  \hspace{1cm} [30]

If \( g_1(\theta) \) and \( g_2(\theta) \) come from the same functional family (e.g. the exponential family as with unit normal distributions) then we can write

\[ f(\theta) = Qg_1(\theta) + Pg_1(\theta - \Delta) \]  \hspace{1cm} [31]

when the two components have a common scale and

\[ f(\theta) = Qg_1(\theta) + Pg_1\left(\frac{\theta - \Delta}{\sigma_2/\sigma_1}\right) \]  \hspace{1cm} [32]

when the scales differ, since for normal distributions with common variance \( g_2(\theta) = g_1(\theta - \Delta) \).

In fact, the above can be generalized to \( K \) component densities by

\[ f(\theta) = \sum_{k=1}^{K} P_k g_k(\theta; \mu_k, \sigma_k) \]  \hspace{1cm} [33]

where

\( g_k(\theta; \mu_k, \sigma_k) \) is the \( k \)th density, depending on

\( \mu_k \), a location parameter for the \( k \)th component density, and

\( \sigma_k \), a scale parameter for that density.

This is a multicomponent finite mixture model (Titterington, Smith, & Makov, 1985). However, for our current problem there is no motivation
to extend the treatment beyond $K = 2$ component densities, so we do not pursue this idea here.

For simplicity’s sake, let us assume that $g_k(\theta), k = 1, 2$ are normal distributions with common scale $\sigma$ and mean difference $\Delta$, mixed in proportions $P : Q$. We arbitrarily set $\mu_1 = 0$ and $\sigma = 1$ without loss of generality, since there always exists a linear transformation of the caseness continuum that achieves this scaling.

Now consider a randomly chosen diagnostician yielding diagnostic scores $X_j$ with characteristic and unvarying diagnostic threshold $T_j$ along the caseness continuum. The probability that this diagnostician gives the $i$th (randomly chosen) case a positive diagnosis is given by

$$
\Pr\{X_{ij} = 1\} = \int_{T_j}^{\infty} f(\theta_i) d\theta_i \\
= Q \int_{T_j}^{\infty} g_1(\theta_i) d\theta_i + P \int_{T_j}^{\infty} g_1(\theta_i - \Delta) d\theta_i \\
= Q[1 - \Phi(T)] + P[1 - \Phi(T - \Delta)] \tag{34}
$$

where

$\Phi(\cdot)$ is as usual the unit normal cdf.

For $J$ randomly chosen raters with a common diagnostic threshold $T$ and common ability to discriminate cases from non-cases $\Delta$, the likelihood
of a patient's receiving exactly $r$ out of $J$ diagnoses positive is

$$\Pr\left\{\sum_{j=1}^{J} X_j = r \right\} = \binom{J}{r} \{Q[1 - \Phi(T)] + P[1 - \Phi(T - \Delta)]\} \tag{35}$$

which despite explaining the agreement between $J$ judges requires integration only in one dimension for its evaluation. Maximum likelihood estimators for this situation, and for Equation 30, were given by Uebersax and Grove (1992).

However, it is unrealistic to imagine that diagnosticians have the decision function implied by Equation 35, which is simply

$$X_{ij} = \begin{cases} 1, & \text{if } \theta_i > T_j \\ 0, & \text{if } \theta_i \leq T_j \end{cases} \tag{36}$$

Let us develop a more realistic model.

Suppose now that different diagnosticians have not only different diagnostic thresholds $T_j$, but also different characteristic abilities to discriminate cases from non-cases. Then the probability that an individual will garner, for example, all positive diagnoses is

$$\Pr\{X_{ij} = 1, \; j = 1, \ldots, J\} =$$

$$= Q \int_{T_1}^{\infty} \cdots \int_{T_j}^{\infty} g_1(\theta_i) \cdots g_1(\theta_i) d\theta_i +$$

$$P \int_{T_1}^{\infty} \cdots \int_{T_j}^{\infty} g_1(\theta_i - \Delta_j) \cdots g_1(\theta_i - \Delta_j) d\theta_i \tag{37}$$
where $T_j$ are the diagnostic thresholds and the $\Delta_j$ parameters index the discriminating abilities of the diagnosticians. For combinations of diagnoses other than all positive, simply substitute integration from $-\infty$ to $T_j$ for every negative diagnosis in the iterated integral in Equation 37.

Alas, Equation 37 requires integration of rectangular regions of the unit multivariate normal distribution. For three dimensions, Henkelman, Kay, and Bronskill (1990) have used this model to estimate ROC curves for imaging techniques in the assessment of liver metastases. In this case the three “diagnosticians” are three tests: radionuclide scintigraphy, magnetic resonance imaging, and computed tomography. These data are reanalyzed below, using the model next to be presented.

The major problem with the Henkelmen et al. approach is that integration of a multivariate normal mixture distribution is a computationally intensive problem, difficult above three dimensions (i.e., three diagnosticians or tests) and extremely unpleasant above five dimensions. For many problems, we need a more tractable method of analysis.

Luckily, a simpler method is at hand. For such complicated situations, it is possible to borrow from latent trait theory as follows. (The following treatment is based on Uebersax and Grove, 1992.) First, treat each diagnostician as if they were an item on a test governed by logistic
latent trait theory. The rating probability function gives the probability
with which a positive diagnosis is given by rater $j$ for a randomly selected
case with latent trait level $\theta_i$. Let $\mathbf{x}_i = \{X_{ij}, \ldots, X_{ij}\} \quad j = 1, \ldots, J$ be
the vector of diagnoses $(X_{ij}) \in \{0, 1\}, j = 1, \ldots, J$ assigned by all $J$
diagnosticians to the $i$th case. Let $\pi_i$ be the probability of $\mathbf{x}_i = x_i$ for a
randomly sampled case. Then

$$\pi_i = \int_{-\infty}^{\infty} f(\theta_i) \prod_{j=1}^{J} p_j(X_{ij} \mid \theta_i) d\theta_i$$

[38]

where

$f(\theta_i)$ is defined in Equation 33.

Now assume that each rater has a characteristic diagnostic threshold
$T_j$ as before. However, instead of assuming the unrealistic decision
function in Equation 32, now we assume that the case's apparent trait
level varies randomly about its true trait level. We will further assume
that this variation about the true trait level, which we assume is due to
measurement error, follows a normal distribution. Then the probability
that a case is assigned a positive diagnosis is given by the normal ogive as
in Figure 4. Since this normal ogive is closely approximated by the logistic
cdf, we employ the logistic curve. Define

$$\psi_j(\theta_i, T_j, a_j) \equiv [1 + \exp \{1.7a_j(T_j - \theta_i)\}]^{-1}$$

[39]
much as in Equation 28.

Using Equations 38 and 39, for this variable-threshold model we immediately derive

$$\pi_i(T, a) = \int_{-\infty}^{\infty} f(\theta, T) \prod_{j=1}^{J} \psi_j(\theta_i, T_j, a_j)^{X_{ij}} [1 - \psi_j(\theta_i, T_j, a_j)]^{1 - X_{ij}} d\theta_i \quad [40]$$

as a function of the vector of thresholds $T$ and the vector of discriminating ability parameters $a$.

Since $f(\theta, T)$ is itself a mixture distribution, substitution gives

$$\pi_i(T, a) = \int_{-\infty}^{\infty} \{Qg_1(\theta_i) + Pg_1(\theta_i - \Delta)\} \times \prod_{j=1}^{J} \psi_j(\theta_i, T_j, a_j)^{X_{ij}} [1 - \psi_j(\theta_i, T_j, a_j)]^{1 - X_{ij}} d\theta_i \quad [41]$$

If we add the restriction for the Rasch model, namely that the diagnosticians are equally discriminating (though their thresholds may differ), $\forall a_j, a_j = a, j = 1, \ldots, J$, then the common variability of apparent variability of caseness about the truecaseness $\theta_i$ for the $i$th case is given approximately by $1/a^2$. Otherwise, the measurement error variance for rater $j$ is given approximately by $1/a_j^2$.

The kernel of the likelihood for the general model, incorporating different rater discriminating abilities, is then given by substitution into Equation 41:

$$\log L = \sum_{i=1}^{N} \sum_{j=1}^{J} \pi_i(g_1(\cdot), \Delta, T, a)^{X_{ij}} [1 - \pi_i(g_1(\cdot), \Delta, T, a)]^{1 - X_{ij}} \quad [42]$$
For parameter estimation, this can then be maximized by the EM method as above. We give an example worked in this fashion below, involving detection of metastatic cancer by three imaging methods.

**Latent Trait Finite Mixture Model**

**For Varying Rater Panel**

The above model assumes that each diagnostican examines each subject. This may not always be the case. When diagnosticians vary from patient to patient, then diagnostican variability is confounded with patient variability. This can be dealt with in the model embodied in the preceding section on latent class analysis. However, the extension to multimodal signal detection models has yet to be made.

**Relation Between Diagnostic Validity and Diagnostic Agreement**

One of the most important implications of the work presented in this chapter is that it undercuts the traditional distinction between reliability and validity. Just as with the taxometric models in the preceding chapter, the models presented here allow one to infer an unobserved but conjectured class membership, starting with nothing more than diagnostic agreement data. This will become clear in the worked examples below. However, here it is essential to remark that, once one thinks of interdiagnostician agreement as having a cause other than simple
chance co-occurrences of positive diagnoses, one has already implicitly abandoned the reliability-validity distinction.

However, one has only done so in a limited sense. Uebersax (personal communication) has pointed out that one can conceive of diagnostic categories on as many as four levels.

1. First, there is the level of a diagnosis accorded by a given diagnostician. This diagnosis may differ from that given by another diagnostician, even if both agree to employ the same technical terms and the same diagnostic criteria. Disagreement occurs for two reasons:
   (a) disagreement about presence or absence of individual signs and symptoms, and
   (b) disagreement about the correct application of diagnostic rules.

2. Second, one can think of a level at which various disorders are “defined,” for example by lists of criteria ensconced in manuals like DSM-III-R. (I prefer not to think of these as definitions proper, but rather as crude but hopefully robust actuarial rules for inferring latent diagnostic group membership.) Diagnostic disagreements can occur between clinicians because they employ different “definitions” of the illnesses in question.
3. Third, one can think of latent diagnostic constructs underlying DSM-III-R rubrics. These are what the criteria lists aim at, though they may only be crudely targeted by such lists. Diseases united in nature may be split in the nomenclature, and disorders united in the nomenclature may in fact have fundamentally different natures.

4. Fourth, different pathologies and etiologies may underly similar clinical disorders, even after the uncertainties in (3) have been taken into account. An example would be Becker and Duchenne muscular dystrophies, which result from different mutations in the dystrophin supergene. Conversely, different disorders may result from the same fundamental lesion. Acquired immune deficiency syndrome may lead to Kaposi’s sarcoma or pneumocystis pneumonia. Tuberculosis may lead to radically different disorders depending on which target organ is attacked.

All the elaborate mathematical analyses of interdiagnostician agreement can hope to do is penetrate to the second or perhaps third level in the above scheme. The “latent class” delivered by latent class analysis is in effect a consensus concept of the disorder in question, among the raters actually employed in a given study.

Let us now proceed to see some examples of the analyses developed in this chapter put to work. We consider three data sets. The first is from
five diagnosticians asked to diagnose the co-twins (mono- and dizygotic) of schizophrenic hospital registry patients. The diagnosticians were explicitly invited to employ their own clinical concepts of schizophrenia, rather than being asked to employ a common nomenclature. Though the data set is small, the results are nonetheless interesting.

The second example concerns the reading of X-rays by physicians, looking for signs of tuberculosis. The data set is very large and so allows statistically powerful model tests.

The third example concerns the detection of metastases using magnetic resonance imaging, computed axial tomography, and radionuclide labelled scans. Much argument has ensued about the relative superiority of these scanning methods, which differ substantially in cost. Since a mistake could cost a patient their life, we are particularly interested in any objective evidence that would lead us to prefer one scanning method to another.

**Worked Examples**

**Example 1:**

Analysis of Schizophrenia Twin Diagnoses

With a Latent Class Model
These data were gathered at the Maudsley Institute of Psychiatry in London, which is the psychiatric facility of the Royal Bethlehem Hospital. On admission, each patient is asked whether they are a twin. Those who said they were had their co-twins (mono- or dizygotic) investigated psychiatrically by lengthy personal interview. Six well-recognized clinicians (psychiatrists and clinical psychologists) then diagnosed the cases from interview and hospital records. Here we only deal with the co-twins, since the condition for admission to the study was that the lead twin was diagnosed schizophrenic. The diagnosticians were not asked to employ any particular concept of schizophrenia. They ranged from very liberal concepts (e.g., judge PM) to quite strict concepts (e.g., judge JB). Table 3 gives the diagnoses for the co-twins for each diagnostician alone, as well as for a "consensus" diagnosis arrived at after the fact by conference.

The resulting six-way agreement table is sparse but nonetheless analyzable using latent class analysis. If we use MLLSA to estimate the latent class parameters under a model which assumes that each diagnostician has their own sensitivity and specificity, we have to estimate five $\alpha$ and five $\beta$ parameters as well as $P$. Since we have $2^5 = 32$ cells in the agreement table and only 11 parameters, the model is identified and we can obtain unique estimates.
Table 3

Agreement Data for Schizophrenia Project

<table>
<thead>
<tr>
<th>Diagnostian</th>
<th>PM</th>
<th>ES</th>
<th>JP</th>
<th>JB</th>
<th>LM</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>N N N N N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>N N N N N S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>N N N S N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>N N N S S S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>N N S N N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>N N S N S S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>N N S S N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>N N S S S N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S S N N N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S N N N N S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S N N S N N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S N S N S N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S N S S S N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S S N N N S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S S N S N S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S S S N N S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>S S S S S S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11</td>
</tr>
</tbody>
</table>
For our analyses, we collapse the diagnostic distinctions into the
table into a dichotomy, S+?S as “schizophrenic” and “O + N” as “not
schizophrenic.” MLLSA’s estimates are given in Table 4, omitting of course
the consensus diagnoses. For these purposes, it is legitimate to ignore the
difference between MZ and DZ twins. Model $M_1$ assumes all raters to have
distinct sensitivities $\alpha_j$ and specificities $\beta_j$, $M_2$ assumes sensitivities equal,
$M_3$ assumes specificities equal, and $M_4$ assumes all raters equally accurate
(i.e., equal sensitivities and equal specificities). In this and all succeeding
tables, $X^2$ is Pearson $\chi^2$ while $G^2$ is likelihood ratio $\chi^2$.

From these parameter estimates we can conclude that the model
having the least number of free parameters, which still fits the data
acceptably, is $M_2$, the model in which diagnosticians have equal
sensitivities. The hypothesis of equal accuracy ($M_4$) can be decisively
rejected, as can the model of equal specificities ($M_3$). From inspection of
the estimated parameters, it is clear that diagnostician ES has a bit lower
specificity than the others. Otherwise, the raters are indistinguishable in
accuracy.

Interestingly, another set of diagnoses (untabled) was also given by
the psychiatrist Erik Essen-Möller. His frequency of giving the diagnosis
of schizophrenia is intermediate compared to the tabled diagnosticians.
Based on the above parameter estimates for model $M_2$, he would have an
Table 4

Latent Class Models for Schizophrenia Data

<table>
<thead>
<tr>
<th>Model</th>
<th>( P )</th>
<th>Rater</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>(-2 \log L)</th>
<th>( G^2 )</th>
<th>( df )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_1 )</td>
<td>.2502</td>
<td>PM</td>
<td>.9997</td>
<td>1.000</td>
<td>4.744</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES</td>
<td>.9995</td>
<td>.8095</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP</td>
<td>.9281</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JB</td>
<td>.9281</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LM</td>
<td>.8568</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_2 )</td>
<td>.2503</td>
<td>PM</td>
<td>.9422</td>
<td>1.000</td>
<td>9.367</td>
<td>4.623</td>
<td>4</td>
<td>.328</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES</td>
<td>&quot;</td>
<td>.8095</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP</td>
<td>&quot;</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JB</td>
<td>&quot;</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LM</td>
<td>&quot;</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_3 )</td>
<td>.2497</td>
<td>PM</td>
<td>.9999</td>
<td>.9615</td>
<td>31.572</td>
<td>26.828</td>
<td>4</td>
<td>.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ES</td>
<td>.9999</td>
<td>&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JP</td>
<td>.9295</td>
<td>&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JB</td>
<td>.9283</td>
<td>&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>LM</td>
<td>.8580</td>
<td>&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_4 )</td>
<td>.2491</td>
<td>All</td>
<td>.9440</td>
<td>.9610</td>
<td>36.197</td>
<td>31.453</td>
<td>8</td>
<td>.0001</td>
</tr>
</tbody>
</table>

Estimated sensitivity and an estimated sensitivity both relatively high (over .975 each). Overall, he would be the best judge among the group in terms of total overall accuracy. Interestingly, his diagnoses have the highest MZ:DZ twin concordance ratio of any judge (Gottesman & Shields, 1972, p. 346). Since the great weight of the evidence indicates that schizophrenia is substantially heritable, this property of Essen-Möller's diagnoses adds to their construct validity.
This information could be used to focus attention on Essen-Möller’s diagnostic deliberations. One could try to trace his diagnostic process in order to produce improved diagnostic criteria for schizophrenia (improved at least in the sense that they would presumably show a closer tie to genetic causes of schizophrenia).

Example 2:

Analysis of Radiographic Data

With Two- to Four-Class Latent Class Models

Yerushalmy (1956) presented data on eight physicians who each interpreted 14,867 chest radiographic films. The task was to diagnose the patients, on whom the films were taken, as having or not having tuberculosis. Table 5 gives the observed frequencies of positive diagnoses, and the expected frequencies under several competing latent class models.

The models fit here assume that all physicians are equally proficient at reading the X-rays. The model $M_2$ in Table 5 is a two-class latent class model. It is obvious that this simple model does not fit these data well; the likelihood ratio $X^2$ statistic is 528.5 on 5 $df$. If we expand the model to include a “gray” category of (probably non-tubercular) cases having various lung ailments, who nonetheless have an elevated probability of being misdiagnosed (model $M_3$), the fit improves but is still completely unacceptable; likelihood ratio chi-square $G^2 = 22.5$. Only a four-class
model \((M_4)\) fits acceptably \((X_1^2 = 0.1)\). For this model, the probabilities of being called tubercular by a randomly chosen physician are .005, .115, .492, and .934 as one moves from class 1 to class 4.

One use of these numbers, pointed out by Uebersax and Grove (1990) and others, is that one can construct prediction tables for future examinees. For example, if 7 of 8 diagnoses are positive, then the chances that the patient has tuberculosis exceeds 95%, while for just 50% confidence only 6 of 8 diagnoses need be positive.

Table 5

<table>
<thead>
<tr>
<th>Number of positive ratings</th>
<th>Observed frequency</th>
<th>Expected frequency for model (M_2)</th>
<th>(M_3)</th>
<th>(M_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13,560</td>
<td>13,452.90</td>
<td>13,557.27</td>
<td>13,559.99</td>
</tr>
<tr>
<td>1</td>
<td>887</td>
<td>1,090.14</td>
<td>883.24</td>
<td>877.02</td>
</tr>
<tr>
<td>2</td>
<td>168</td>
<td>45.27</td>
<td>146.65</td>
<td>167.91</td>
</tr>
<tr>
<td>3</td>
<td>66</td>
<td>25.08</td>
<td>92.25</td>
<td>66.29</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>55.10</td>
<td>42.24</td>
<td>41.25</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>79.94</td>
<td>16.39</td>
<td>29.05</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>72.49</td>
<td>21.68</td>
<td>22.13</td>
</tr>
<tr>
<td>7</td>
<td>39</td>
<td>37.56</td>
<td>50.51</td>
<td>39.64</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
<td>8.51</td>
<td>56.76</td>
<td>63.73</td>
</tr>
</tbody>
</table>
Example 3:

With A Latent Trait Finite Mixture Model

The final example is based on Uebersax and Grove (1992). Henkelman, Kay, and Bronskill (1990) presented data for 298 cases examined for liver metastases using three imaging techniques: magnetic resonance imaging, radionuclide scintigraphy, and computed axial tomography. (Their approach to these data was explained above in the section entitled “Latent Trait Finite Mixture Model For Fixed Rater Panel.”) Each imaging technique yielded, upon expert interpretation, an ordered category score from 1 to 5, with 5 representing highest probability of metastases. Table 6 presents the data.

Since the data in this table are sparse (with many observed zero counts), we are concerned that the distribution of the likelihood ratio chi-square $G^2$ statistic may not be adequately approximated by $\chi^2$. Therefore, in preliminary analyses (not detailed here) we collapsed the data into three rating levels, so that categories 2-4 were joined for each examination method. Then we fit the latent trait mixture model with two normal distributions ($\mu_1 = -\Delta/2$ and $\mu_2 = \Delta/2$) and assumed that $\sigma_1 = \sigma_2$. Values of $G_1^2 = 11.93$ result, so that the model appears to fit the data.

This nonsignificant likelihood ratio chi square justifies us in returning to an analysis of the full five-level rating data. To these rating
Table 6
Crossclassification of Results of
Three Diagnostic Tests for Metastases

<table>
<thead>
<tr>
<th>Radionuclide rating level</th>
<th>MRI rating level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>CT rating level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>CT rating level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36 20 2 0 0</td>
<td>10 7 2 0 0</td>
<td>1 3 0 0 0</td>
<td>22 14 3 1 0</td>
<td>7 7 1 4 0</td>
<td>1 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3 2 1 0 0</td>
<td>0 0 0 0 1</td>
<td>0 0 0 0 0</td>
<td>3 0 1 0 1</td>
<td>0 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1 0 0 1 0</td>
<td>3 0 1 0 1</td>
<td>0 0 0 0 0</td>
<td>3 0 0 0 1</td>
<td>1 0 1 0 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1 0 0 1 0</td>
<td>3 0 1 0 1</td>
<td>0 1 1 1 1</td>
<td>3 0 0 0 1</td>
<td>1 0 1 0 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1 0 0 1 0</td>
<td>3 0 1 0 1</td>
<td>0 1 1 1 1</td>
<td>3 0 0 0 1</td>
<td>1 0 1 0 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Radionuclide rating level</th>
<th>MRI rating level</th>
<th>4</th>
<th>5</th>
<th>CT rating level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>CT rating level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 1 0 1 0</td>
<td>1 1 0 0 1</td>
<td>1 1 0 0 1</td>
<td>0 3 1 2 0</td>
<td>0 2 0 1 1</td>
<td>2 1 0 2 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0 3 1 2 0</td>
<td>0 2 0 1 1</td>
<td>1 1 0 3 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0 1 1 0 1</td>
<td>0 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1 1 0 0 1</td>
<td>1 1 0 3 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0 1 0 2 5</td>
<td>3 1 1 1 66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

scale data we fit the Rasch-type model by assuming that $a_1 = a_2 = a_3 = a$.

Under this assumption, we fit model $M_1$ (parameter estimates given in
which assumes a single normal distribution, similar to logistic latent trait theory (i.e., $\mu_1 = \mu_2 = 0$). Also in Table 7, model $M_2$ has two normal distributions with parameters $\mu_1 = -\Delta/2$, $\mu_2 = \Delta/2$, and common $\sigma_1 = \sigma_2 = \sigma = 1$. Model $M_3$ is just $M_2$ with different $a_j$ for different tests. Model $M_4$ is $M_2$ with the added assumption that $T_1 = T_2 = T_3 = T$ for all tests. Statistics relating to goodness of fit for these models are shown in Table 8.

Table 7

Model Parameters for Metastases Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Asymptotic S.E.</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Asymptotic S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta/2$</td>
<td>2.899</td>
<td>0.4331</td>
<td>$b_{22}$</td>
<td>-2.966</td>
<td>0.4670</td>
</tr>
<tr>
<td>$p$</td>
<td>0.601</td>
<td>0.0311</td>
<td>$b_{23}$</td>
<td>-0.783</td>
<td>0.2757</td>
</tr>
<tr>
<td>$a$</td>
<td>0.575</td>
<td>0.0959</td>
<td>$b_{24}$</td>
<td>-0.301</td>
<td>0.2671</td>
</tr>
<tr>
<td>$b_{12}$</td>
<td>-2.353</td>
<td>0.3960</td>
<td>$b_{32}$</td>
<td>-2.799</td>
<td>0.4472</td>
</tr>
<tr>
<td>$b_{13}$</td>
<td>-0.753</td>
<td>0.2716</td>
<td>$b_{33}$</td>
<td>-0.778</td>
<td>0.2734</td>
</tr>
<tr>
<td>$b_{14}$</td>
<td>-0.132</td>
<td>0.2649</td>
<td>$b_{34}$</td>
<td>-0.028</td>
<td>0.2668</td>
</tr>
<tr>
<td>$b_{15}$</td>
<td>1.023</td>
<td>0.3184</td>
<td>$b_{35}$</td>
<td>1.525</td>
<td>0.3529</td>
</tr>
</tbody>
</table>

Clearly, the model with one normal distribution does not fit the data. The two normal distribution model $M_2$ does fit, and the Rasch model fits just as well as do models having different discriminabilities ($\Delta$s) for each test. The assumption of different thresholds ($T_{ij}$) for the different tests is
not rejected. The latent trait finite mixture model provides a parsimonious account of the data whereas the logistic latent trait model (i.e., assuming just one underlying caseness distribution) does not.

Conclusions

We have seen that one can start from a simple conception of diagnostician agreement and garner considerable psychometric information from it. In particular, one can model interdiagnostician agreement by thinking of an underlying model in which diagnoses are either correct or incorrect. Using such latent class models, the epidemiologists' parameters of diagnostic sensitivity, sensitivity and base rate (prevalence) can be estimated provided data from enough raters are available.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>df</th>
<th>$X^2$</th>
<th>$G^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$</td>
<td>One normal distribution</td>
<td>111</td>
<td>138.89</td>
<td>163.68</td>
</tr>
<tr>
<td>$M_2$</td>
<td>Two normal distributions</td>
<td>109</td>
<td>112.70</td>
<td>111.55</td>
</tr>
<tr>
<td>$M_3$</td>
<td>Two normal distributions, different measurement error across tests</td>
<td>107</td>
<td>118.87</td>
<td>110.95</td>
</tr>
<tr>
<td>$M_4$</td>
<td>Two normal distributions, identical thresholds</td>
<td>117</td>
<td>123.40</td>
<td>124.00</td>
</tr>
</tbody>
</table>
More sophisticated models can also be employed. One can think of individuals as lying on a continuum of caseness. Then one can characterize individuals as clearly cases, clearly non-cases, or any degree of “gray” in between. This allows one to account for the fact that observers are in close agreement on presence of disease in some individuals, absence of disease in other patients, and in great disagreement with respect to still others.

However, such increase in model fidelity comes with a price. One must then think of diagnosticians as each having characteristic readiness to diagnose an illness, and each having a characteristic skill level at detecting the difference between cases and non-cases. One can then assume various distributions for the latent, unobserved distribution of “caseness.” The distribution may be unimodal (leading to logistic latent trait theory), multimodal with known disease classifications (leading to signal detection theory), or multimodal with unknown disease classifications (leading to latent trait finite mixture models).
Appendix

Remarks on Maximum Likelihood Estimation

Maximum likelihood estimation involves Fisher’s principle of likelihood inference. According to this principle, we should assign maximal credence to parameter estimates which would, if they were correct, make our data the most likely (when compared to other possible parameter estimates). We have just derived a likelihood function, expressing the probability of observing a set of diagnostic judgments (the data) as a function of some model parameters. Usually one interprets such a probability function as one depending on the data, given a fixed (known or unknown) set of parameters. However, nothing prevents one from formally reinterpreting the likelihood as a function of the parameters, given a fixed and known set of data. In that case, one can differentiate the likelihood with respect to the parameters. The desired maximum of the likelihood is then found when

$$\frac{\partial L}{\partial \theta} = 0 \quad [A1]$$

where

$\theta$ is a $k$-element vector of unknown parameters.

In a few simple cases, there exists a closed-form solution for Equation A1. However, for the types of problems considered here, there is
no analytic solution. The maximum likelihood estimates have to be found by a stepwise process of approximation, or iteration.

There are two main ways of iteratively solving the above equation: hill-climbing methods, and direct search of the likelihood function's surface. The former methods have many subvarieties; the interested reader may consult standard references on nonlinear numerical methods such as Himmelblau (1972). We concentrate here on the latter approach, direct search. This approach explores a \( (k + 1) \)-dimensional “surface” of the likelihood function, where \( k \) of the dimensions correspond to the \( k \) parameters to be estimated and the last dimension is the value of the likelihood function \( L \).

Direct search consists in principle of the following steps, starting from an initial parameter vector estimate and an initial grid across the likelihood surface:

1. Construct a \( k \)-dimensional grid covering that part of the parameter space for which it is reasonable to suppose, on theoretical or empirical grounds, that the true value of the vector \( \theta \) lies in this region. The grid may or may not be regular. Here we assume a non-regular grid, with a closer mesh near parameter boundaries.

For our problem, almost all the parameters are themselves probabilities and hence must lie between 0 and 1. For typical problems,
moreover, $P$ will not lie outside the interval $P \in [.01,.5]$, $\alpha_j$ and $\beta_j$ will all lie in the interval $\alpha_j, \beta_j \in [.5,.995]$, $j = 1, \ldots J$, and $\Delta$ will lie in the interval $[.5, 5]$.

Let us quickly justify these assumptions about parameter ranges. If $\alpha$ or $\beta$ were any lower than $1/2$, interrater reliability would be much worse than it is observed to be for most medical problems, even at the most favorable value of $P$, namely $1/2$. Likewise, if $\Delta$ were less than $1/2$, we would not be able to sort cases from non-cases with any fidelity to speak of. On the other hand, if $\Delta$ were larger than 5, it would be so easy to tell cases from non-cases that, almost no matter what the values of $\alpha$ and $\beta$ parameters, reliability would be nearly perfect at all practical values of $P$. We assume $P$ does not assume extreme values (outside the interval $[.01, .5]$ because most reliability studies are conducted in clinical populations, in which the proportion of cases is artificially enriched over the rate in the general population. $P$ cannot ordinarily exceed $1/2$ since no one disorder predominates in most clinical settings. On the other hand, for a general population reliability study, we might have to consider values for $P$ as low as 0.1%. (See special issue of the Archives of General Psychiatry, October 1984, especially Myers, Weissman, Tischler, Holzer, Leaf, Orvaschel, Anthony, Boyd, Burke, Kramer, & Stolzman, 1984).
2. These boundaries imply that we may construct an initial grid at the points \( P \in \{.01, .05, .1, .2, \ldots, .5\} \), \( \alpha, \beta \in \{.5, .6, .7, .8, .9, .95, .99, .995\} \), \( \Delta \in \{.5, 1, 1.5, 2, 3, 4, 5\} \) in all possible combinations.

3. Set the iteration counter \( i \) to \( i = 0 \).

Repeat the following steps 4-7 until the parameter estimates converge.

4. For each grid point, evaluate the likelihood at the parameter estimate \( \hat{\theta}^i \) where the superscript indexes the iteration number. Note the grid point giving the maximal value of \( L \). Consider this grid point to be the new parameter estimate \( \hat{\theta}^{i+1} \) with corresponding likelihood \( L^{i+1} \).

5. Create a new grid centered on the point \( \hat{\theta}^{i+1} \), extending at least as far as halfway (in each direction) as the most adjacent grid points. One might for example take the previous typical intergridpoint distance and divide it by, say, 1/2, to yield a new typical intergridpoint distance, then generating another grid from this. (This makes the parameter estimation procedure a sort of multidimensional binary search.)

6. Increment \( i \).

7. Evaluate

\[
C_1 = (L^{i+1} - L^i)/L^{i+1} \tag{A2}
\]
Also evaluate
\[
C_2 = \max_j | \tilde{\theta}_j^{i+1} - \tilde{\theta}_j^i |, \quad j = 1, \ldots, k
\]
[A3]

The former indexes the relative change in L, the latter the maximal absolute change in any individual parameter estimate. If these two quantities \(C_1, C_2\) are smaller than preselected small constants \(c_1, c_2\) (e.g., \(c_1 = 10^{-4}, c_2 = 10^{-5}\)) then stop the estimation procedure, as convergence has been achieved. Otherwise, return to step 4.

8. The maximum likelihood estimate of the parameter vector is then taken to be
\[
\text{vec} \theta = \hat{\theta}^i
\]
[A4]

Note that since we are now using maximum likelihood we revert to the usual notation \(\hat{\theta}\) rather than \(\tilde{\theta}\).

Likelihood Maximization Via the EM Algorithm

Other approaches useful in models of the kind discussed here are the iterative proportional fitting method (Agresti, 1990) and the EM method (Dawid & Skene, 1979; Uebersax & Grove, 1990, Appendix). In particular, for some kinds of models discussed below, computer programs have been implemented using a combination of the EM algorithm and hill-climbing methods such as Newton-Raphson (Uebersax & Grove, 1992). Other programs (e.g., Clogg’s [1977] Maximum Likelihood Latent Structure
Program, MLLSA, use the EM algorithm alone. (MLLSA is probably the currently most used program for performing analyses of the kind described in this section.) The reader may consult the cited references for further information.

Since the EM algorithm is widely used in this type of problem, we explain it briefly here in the context of logistic latent trait theory. EM stands for a process called Expectation-Maximization, and proceeds in to two distinct steps. To see these steps as applied to our problem, consider a latent trait model like Equation 28, which we repeat here for convenience:

\[
\Pr \{ X_{ij} = 1 \mid \theta_i \} = \frac{\exp(1.7[b_j + a_j \theta_i])}{1 + \exp(1.7[b_j + a_j \theta_i])}, \quad j = 1, \ldots, J, \quad i = 1, \ldots, N \quad [A5]
\]

where

- \( b_j \) represents a measure of the probability of getting item \( X_j \) correct (i.e. obtaining a positive diagnosis) for an individual of average ability (caseness),
- \( a_j \) represents the capacity of item \( X_j \) to discriminate grades of the latent ability (caseness) \( \theta \),
- \( \theta_i \) is the caseness value of the \( i \)th individual examined, and
- the constant 1.7 is chosen to make the numerical details agree with a similar model in which the logistic cdf is replaced by the Gaussian cdf.
Suppose that a large number of test items (or diagnosticians) are given to each individual. Then a rough idea of the standing of each individual on the caseness continuum is given by the number of items gotten correct, or the number of positive diagnoses earned. Let us scale this for simplicity so that the mean of these scores is zero and the standard deviation unity. Treating these estimated scores

\[
\hat{\theta}_i = \sum_{j=1}^{J} \hat{X}_{ij}
\]  

[A6]

where

\(X_j\) is the score for the \(j\)th diagnosis (0 for not ill, 1 for ill) as before, and the scores have been suitably normed. Let us now regard these \(\theta\) parameters for each individual temporarily as fixed. Treating these as constants, we now estimate the parameters \(a_j\) and \(b_j\) for each item (diagnostician) by maximum likelihood. This is the *Maximization* phase of the EM method.

Now, we have temporarily determined values \(\hat{a}_j\) and \(\hat{b}_j\) parameters. Then, the best estimate of the \(i\)th individual’s standing on the caseness or \(\theta\) dimension is given by the weighted sum

\[
\hat{\theta}_i = \sum_{j=1}^{J} \hat{b}_j + \hat{a}_j \hat{X}_{ij}
\]  

[A7]
where
the score \( \hat{\theta}_i \) is again suitably normed to zero mean and unit variance. This is the *Expectation* phase of the EM method.

These two phases, Expectation and Maximization, alternate until the process converges with stable values for the \( \theta \) parameters for subjects and the \( a_j \) and \( b_j \) parameters for diagnosticians (or test items). Unfortunately, convergence is often slow and may require hundreds or even thousands of iterations. Nonetheless, it converges rather reliably and is hence used frequently for these kinds of problems.

The EM algorithm has the disadvantage that it does not use the likelihood explicitly. Therefore, in order to obtain standard errors (see below) it is necessary to follow a converged EM estimation procedure with an iteration or two of a derivative-based method like Newton-Raphson (Himmelblau, 1972). This gives the required basis for computing standard errors and confidence intervals.

**Numerical Problems in Maximum Likelihood Estimation**

There are two kinds of maxima of the likelihood, local and global. Obviously, the global or overall maximum is the one we seek, yet we may instead converge to a merely local maximum. Starting from a local maximum, the only way to reach the global maximum is to travel in a direction in the \( \theta \)-space that temporarily worsens \( L \). Hill-climbing iteration
methods cannot do this and so they must be run again and again, starting from various points that may (or may not be) appropriately “near” the global maximum. One then chooses that local maximum which is itself maximal, declaring the corresponding parameter estimate to be the maximum likelihood estimate. This method works well for likelihood surfaces that are suitably “smooth.”

For problems like the ones considered here, it often happens that the likelihood is rather bumpy and filled with potholes. In these situations, direct search, while potentially much more computationally intensive, can yield better estimates. However, even with direct search there is no guarantee that the global maximum of the likelihood has been found.

The iteration procedure is subject to certain numerical estimation problems which we will not detail here. Suffice it to say that special care must be paid to the computation of the likelihood and its derivatives, so that parameter estimates do not end up being perverted by accumulated round-off error. It may also be necessary to deal with cells in the $J$-way interjudge agreement table which happen to have zero counts, especially when the ratio of cells to total sample size $2^J/N$ is not large (Agresti, 1990, pp. 244-250).
Standard Errors and Confidence Intervals

Using asymptotic distribution theory, one can evaluate standard errors for parameter estimates and thereby erect confidence intervals. To estimate the standard error for a parameter estimate \( \hat{\theta}_j \), we usually perform numerical double differentiation of the likelihood with respect to the parameter in question. This is accomplished as follows.

1. Evaluate the first partial derivative of \( L \) with respect to \( \theta_j \) at the likelihood maximum. This can be done by the method of “forward differences,” i.e. by the formula

\[
\frac{\partial L}{\partial \theta_j} \approx \frac{L(\hat{\theta}_j + \epsilon) - L(\hat{\theta}_j)}{\epsilon} \quad [A8]
\]

where

\( L(\hat{\theta}_j + \epsilon) \) means to reevaluate the likelihood after changing just the parameter \( \hat{\theta}_j \) to a slightly different value \( \hat{\theta}_j + \epsilon \), leaving all other parameters unchanged. \( \epsilon \) is a small number, e.g. \( \epsilon = 10^{-5} \).

2. Evaluate the second partial derivative of \( L \) with respect to \( \theta_j \) at the likelihood maximum by re-applying the above formula twice, taking differences of the differences:

\[
\frac{\partial^2 L}{\partial \theta_j^2} \approx \frac{L(\hat{\theta}_j + 2\epsilon) - 2L(\hat{\theta}_j + \epsilon) + L(\hat{\theta}_j)}{\epsilon^2} \quad [A9]
\]
In this fashion we thus build up an estimate of the $k \times k$ matrix of second partial derivatives of the (log) likelihood with respect to the $k$ parameters

\[
\begin{pmatrix}
\frac{\partial^2 L}{\partial \theta_i \partial \theta_j} & \frac{\partial^2 L}{\partial \theta_i \partial \theta_2} & \cdots & \frac{\partial^2 L}{\partial \theta_i \partial \theta_j} & \cdots & \frac{\partial^2 L}{\partial \theta_i \partial \theta_k} \\
\frac{\partial^2 L}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_j} & \cdots & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_k} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\frac{\partial^2 L}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 L}{\partial \theta_k \partial \theta_j} & \cdots & \frac{\partial^2 L}{\partial \theta_k \partial \theta_k}
\end{pmatrix}
\]  

[A10]

This matrix is asymptotically equivalent to the Fisher information matrix $\mathbf{J}$. This matrix plays a key role in likelihood-based statistical inference (Lehman, 1983). For example, the standard error for $\theta_j$ is a simple function of this matrix:

\[
\text{s.e.}(\theta_j) \approx \sqrt{\mathbf{J}_{jj}^{-1}}
\]

[A11]

where

$\mathbf{J}_{ij}^{-1}$ is the $(i,j)$th element of the inverse of the Fisher information matrix. A $(100 - c)\%$ confidence interval for $\theta_j$ is then easily obtained as

\[
\hat{\theta}_j \in [\hat{\theta}_j - z_{c/2} \text{s.e.}(\theta_j), \hat{\theta}_j + z_{c/2} \text{s.e.}(\theta_j)]
\]

[A12]

where

$z_{c/2}$ is the $(100 - c/2)$ percentile point of the unit normal distribution.
Hypothesis Testing

The maximum likelihood estimation approach easily allows for tests of hypotheses. Suppose for example one wishes to test the hypothesis \( \alpha = c \) for some constant \( c \). One simply proceeds as follows. First estimate all the model parameters, including \( \alpha \), freely. Denote the log likelihood obtained at this maximum by \( \log L_1 \). Then re-estimate the model parameters, this time applying the constraint \( \alpha = c \) as specified by the hypothesis of interest. Denote the log likelihood obtained at this constrained-parameter maximum by \( \log L_0 \). Then under the usual assumptions the quantity \( 2 \log L_0 - \log L_1 \) is asymptotically distributed as \( \chi^2_1 \). In general, to test a set of \( k \geq 1 \) parameters for equality to constants simultaneously, one does the same thing as before, but when obtaining \( \log L_0 \) one constrains all the desired parameters at once. Twice the resulting difference in log likelihoods is again asymptotically distributed as \( \chi^2 \) with degrees of freedom equal to the number of linearly independent parameters which were constrained in the calculation of \( \log L_1 \) but not in \( \log L_0 \).

A more realistic hypothesis test would occur in more complicated models like those in which there are distinct accuracy parameters for each judge. Then it may be of critical interest to test the hypothesis that the judges are equally sensitive, i.e. \( \alpha_1 = \alpha_2 = \ldots = \alpha_J \), that they are equally
specific, i.e. $\beta_1 = \beta_2 = \ldots = \beta_J$, or both at once. This is easily approached as follows. First estimate $\log L_1$ via a model in which there are distinct $\alpha$ and $\beta$ parameters for each rater. Then estimate $\log L_0$ using a model with a common $\alpha$ and a common $\beta$. Twice the difference in log likelihoods is asymptotically distributed as $\chi^2_{2(J-1)}$, where $J$ is, as usual, the number of judges.

Diagnosing Model Fits

When models do not fit adequately, one typically wishes to know whence this lack of fit arises. Discovering the source of poor model fits is partly art and partly science.

The simplest way to begin looking for the source of a poorly fitting model is to compute expected cell counts in the interrater agreement table based on the parameter estimates. For example, in a three-judge study (all judges equally accurate) the expected number of individuals with all three diagnoses positive is given by

$$\hat{n}_{+++} = N(\hat{P} \hat{\alpha}^3 + \hat{Q}(1 - \hat{\beta})^3)$$

The deviation of $\hat{n}_{+++}$ from its predicted value can be assessed for statistical significance by the formula

$$\frac{(\hat{n}_{+++} - n_{+++})^2}{\hat{n}^2_{+++}} \sim \chi^2_1$$  \[A13\]
which says that the left hand side of Equation A10 is asymptotically
distributed as $\chi^2_1$. If $\hat{n}_{+++}$ is far from expectation, one should then closely
examine the marginal frequencies $n_{++}$, $n_{++}$, and $n_{++}$, where the subscript
“+” denotes a positive diagnosis and the dot subscript denotes summing
over the missing subscript. This will often give clues to the cause of the
discrepant count. Usually at least one rater is not well modelled, leading to
a poor fit for the corresponding marginal as well.

It can also happen that there exists a dependence between two raters
not well described by the above model. In that case, the residuals for
a collapsed contingency table (summing over the third diagnostician’s
judgments) will show a pattern in which one diagonal of the (collapsed)
$2 \times 2$ table is greater than the other.
References


