Psychological Bulletin
1987, Vol. 101, No. 1, 140-146

0033-2909/87/$00.75

Diversity of Decision-Making Models and the Measurement of Interrater Agreement

John S. Uebersax
Center for Health Policy Research and Education
Duke University

Several papers have appeared criticizing the kappa coefficient because of its tendency to fluctuate with sample base rates. The importance of these criticisms is difficult to evaluate because they are presented with regards to a highly specific model of diagnostic decision making. In this article, diagnostic decision making is viewed as a special case of signal detection theory. Each diagnostic process is characterized by a function that relates the probability of a case receiving a positive diagnosis to the severity or salience of symptoms. The shape of this diagnosability curve greatly affects the value of kappa obtained in a study of interrater reliability, how it changes in response to variation in the base rates, and how closely it corresponds to the validity of diagnostic decisions. The common practice of evaluating a diagnostic procedure, when criterion diagnoses for comparison are unavailable, on the basis of the magnitude of the kappa coefficient observed in a reliability study is questionable. New methods for measuring interrater agreement are necessary, and possible directions for research in this area are discussed.

The kappa coefficient (Cohen, 1960) is generally regarded as the statistic of choice for measuring agreement on ratings made on a nominal scale. It is relatively easy to calculate, can be applied across a wide range of study designs, and has an extensive history of use, particularly in the area of the reliability of psychiatric diagnosis, to recommend it. Recently, however, there has been growing concern that it may not be entirely satisfactory as an index of interrater agreement. This article examines the arguments that have been raised, noting that in some ways they are correct and in other ways they are not. However, in the course of doing this, additional problems will become apparent that ultimately cast further doubt on the usefulness of kappa as a general tool for evaluating interrater agreement. It is shown that kappa's ostensible purpose of providing a measure of interrater agreement that is corrected for chance is questionable, and that an observed value of kappa can be interpreted only in the context of a specific decision-making model known to govern classificatory judgments, but that this model is typically not known to the researcher.

The approach taken is to view the measurement of interrater reliability from the standpoint of signal detection theory. Although signal detection theory has been applied extensively to the assessment of the validity of diagnostic procedures (Lusted, 1968; Metz, 1978; Swets, 1986; Swets & Pickett, 1982), comparatively little attention has been paid to applying it to problems arising in conjunction with the measurement of interrater agreement. In addition to clarifying kappa's limitations as an agreement index, the signal detection model has many important implications for further research in this area.

The author thanks two anonymous referees and the editor for many helpful suggestions concerning the form and content of this article.

Correspondence concerning this article should be addressed to John S. Uebersax, Center for Health Policy Research and Education, Box GM, Duke Station, Duke University, Durham, North Carolina 27706.

Chance Correction

Originally, the kappa coefficient was proposed as a measure of agreement that circumvents certain problems with the simpler agreement statistic, the observed percentage of times pairs of raters agree on assigning a case to the same category, or p_o. Cohen (1960) argued that in addition to p_o, a second term, p_c, should be calculated, representing the level of agreement expected were raters to make classifications (a) on a random basis, and (b) according to probabilities that correspond to the base rates with which they make each diagnosis. Kappa combines both terms in the formula $\kappa = (p_o - p_c)/(1 - p_c)$, providing a "chance-corrected" measure of agreement. The term p_c, however, represents the amount of agreement that would occur under a null hypothesis of random decision making by all raters. If the null hypothesis were true, kappa would be exactly equal to zero; to the extent that it is not, the null hypothesis is unlikely, and there is evidence for concluding that raters are not making their decisions on a random basis. However, because p_c is derived under the conditions of a null hypothesis of completely random agreement, it is not clear how the magnitude of kappa is to be interpreted once the null hypothesis is known not to be true.

Base Rate Problem

Much of the recent concern about the kappa statistic has centered around what has been called the "base rate problem" (Spitznagel & Helzer, 1985). That is, the same diagnostic process may yield different values of kappa depending on the proportions of positive and negative cases in the sample. The immediate concern is that kappa values obtained from different studies may not be comparable, although there are other equally important consequences associated with base rate variation as well.

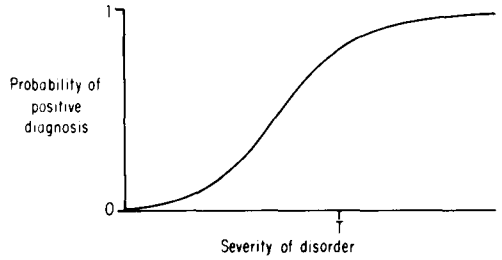

Figure 1. Curve relating the probability of positive diagnosis to symptom intensity or salience (*T*: diagnostic threshold, or the point along the continuum separating positive and negative cases).

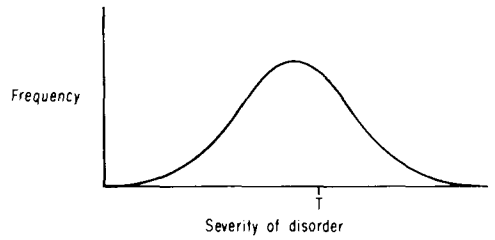

Figure 2. Frequency distribution showing the number of cases at each symptom level (*T*: the diagnostic threshold).

Some degree of responsivity to base rates is to be expected for any index that globally assesses the reliability or validity of both positive and negative diagnoses. In general, we would expect some difference in the accuracy with which diagnosticians are able to detect positive and negative cases. Moving the sample base rates in the direction of the more recognizable category will tend to raise both reliability and validity. For the kappa coefficient, however, this is complicated by the fact that the base rates themselves figure prominently in its calculation.

Spitznagel and Helzer (1985) have made the best technical investigation of the subject to date. They presented evidence suggesting that kappa is too responsive to base rates to be useful as an index of interrater agreement; however, that conclusion's generality is limited by the assumption that the decisions of one diagnostician are independent of those of another, conditional on the true diagnosis. This assumption of *conditional independence* is plausible when the disorder being diagnosed is dichotomous, but not when, as is more generally the case, there are gradations in terms of symptomatology or the salience of diagnostically relevant information. For example, suppose that all positive cases have a .7 probability of being diagnosed positive. The probability of a positive case receiving a positive diagnosis by two diagnosticians is $.7 \times .7 = .49$. However, now suppose that there are two types of positive cases, present in equal proportions, one with a .5 probability of receiving a positive diagnosis and the other with a .9 probability. In this case, the probability of a positive case receiving two positive diagnoses is $.5(.5 \times .5) + .5(.9 \times .9) = .53$. What is needed is a way to evaluate kappa across a wider and more representative range of diagnostic situations.

Models of Diagnostic Decision Making

The ensuing discussion is restricted to the simple case in which diagnosticians are asked to decide whether cases either do or do not have a particular disorder. Figure 1 illustrates a hypothetical curve characterizing diagnosticians' responses to varying levels of a patient's symptomatology. The *x*-axis corresponds to symptom level, and may, depending on the context, indicate either symptom intensity or the salience of diagnostically relevant information. In general, the more severe the disorder, the greater the case's value on the *x*-axis. This dimension may correspond to one clinically relevant symptom, such as the degree of depression, or, given certain assumptions of additivity, a composite representing several different symptoms, such as Minnesota Multiphasic Personality Inventory profile elevation.

The *y*-axis corresponds to the probability that a patient with a given symptom level will be diagnosed as having the disease. This is closely related to the concept of an item characteristic curve in test theory (Hulin, Drasgow, & Parsons, 1983), but for convenience it is referred to here as the *diagnosability curve* for a disorder. The point T indicates the diagnostic threshold, the dividing point in terms of symptom level above which cases qualify as having the disorder and below which they do not.

Figure 1 is drawn arbitrarily, but has two factors that are likely common to most curves of this type. First, it levels out at both extremes; we expect that there is a point on the *x*-axis above which increasing the level of disease does not raise the probability of a positive diagnosis, and one below which decreasing levels do not decrease the probability. Second, the curve is monotonic nondecreasing; cases with lower levels of symptomatology are never associated with higher probabilities of positive diagnosis. Even without specifying the exact shape of a diagnosability curve, however, we can derive a number of implications concerning the nature and measurement of diagnostic processes. We begin by defining the following:

x = level of symptomatology;
$p(x)$ = function specifying the probability of a positive diagnosis for a case with symptom level x;
$q(x) = 1 - p(x)$;
$f(x)$ = frequency of cases at symptom level x (Figure 2);
T = diagnostic threshold; and
N = total number of cases.

The frequency distribution for case severity, f, is shown here as a normal curve. This would correspond to a disorder defined as an extreme level of some trait that is normally distributed across the population. This may be an appropriate model for diagnoses such as personality disorders. Alternatively, if positive cases can be thought of as constituting a distinct population, such as a disorder with an organic basis, the situation may be more accurately represented as the sum of two overlapping distributions of positive and negative cases.

Knowing T and the shapes of p and f is sufficient to calculate the entire range of statistics pertaining to the evaluation of diagnostic procedures. Suppose that x has been divided into a large number of equal-sized intervals. Defining $a(x) = p^2(x) + q^2(x)$, the expected value of p_o is equal to the sum of the product $a(x)f(x)$ across all intervals, divided by the total number of cases, N. Similarly, the proportion of positive diagnoses made equals the sum of $p(x)f(x)$ over all intervals of x, divided by N, and the proportion of negative diagnoses equals the sum of $q(x)f(x)$ across all intervals, again divided by N; squaring both

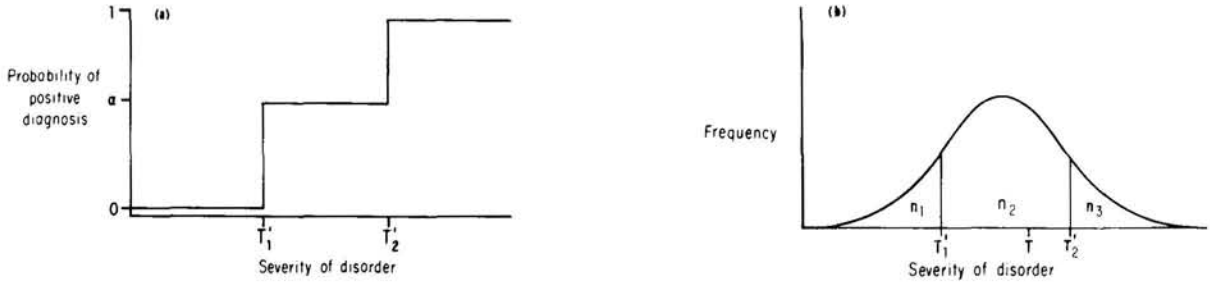

Figure 3. Decision-making model with provision for random guessing on uncertain cases, based on Maxwell (1977); (a) diagnosability curve for decision-making model (T'_1: the level of severity below which a negative diagnosis is certain; T'_2: the level at or above which a positive diagnosis is certain); (b) frequency distribution and numbers of cases in each region defined by the model (n_1, n_2, and n_3: numbers of cases in each region defined by T'_1 and T'_2; T: the point dividing positive and negative cases).

of these results and adding them gives the expected value for p_c. As shown in the Appendix, similar formulas can be applied to calculate commonly used indices of diagnostic validity.

Let us now consider some hypothetical diagnosability curves. Figure 3a shows one of a class of curves described in connection with psychiatric diagnosis by Maxwell (1977) and Janes (1979). Curves of this type are characterized by a three-tiered shape, corresponding to three types of cases in the sample: one for which the probability of a positive diagnosis is 1, one for which the probability is 0—that is, positive or negative diagnoses can be made with certainty for both groups—and a third group of questionable cases, where the probability of positive diagnosis is some value, α, between 0 and 1.

The three ranges are divided on the x-axis by T'_1 and T'_2. Like T in Figure 1, these are diagnostic thresholds. However, unlike T, they are not points that separate cases according to which category they actually belong to, but ones separating regions of the x-axis with differing probabilities of positive diagnosis. They are, in effect, thresholds for classification as opposed to a definitional division point represented by T. For the present, it is assumed that the curve, including T'_1 and T'_2, is constant across all diagnosticians, though ultimately it would be desirable to make some provision for individual differences. Maxwell considered the situation where $\alpha = .5$, that is, where positive and negative diagnoses are equiprobable for questionable cases; however, this might just as easily correspond to some other value. Here we shall consider the case where α is equal to the proportion of positive cases in the sample, which would mean that diagnosticians respond to uncertain cases by displaying the phenomenon of "probability matching" (Atkinson, Bower, & Crothers, 1965). Referring to Figure 3b, n_1 is the number of cases with symptom levels from the origin to T'_1, n_2 is the number of cases between T'_1 and T'_2, and n_3 is the number of cases at or above T'_2. If we assume that the true diagnostic threshold T, the point separating positive and negative cases, is between T'_1 and T'_2, we can calculate upper and lower bound estimates for the validity of diagnostic decisions. Sensitivity ranges from $(\alpha n_2 + n_3)/(n_2 + n_3)$ to 1.0, and specificity from 1.0 to $[n_1 + (1 - \alpha)n_2]/(n_1 + n_2)$ as T moves from T'_1 to T'_2.

Figure 4a illustrates a different diagnosability curve corresponding to a diagnostic model discussed by Kraemer (1980) and Kaye (1980), among others. This is also the diagnostic situation considered in analyses by Spitznagel and Helzer (1985) and Grove, Andreason, McDonald-Scott, Keller, and Shapiro (1981). This model differs from the preceding one in that cases are divided into only two groups, those for which the probability of a positive diagnosis is α and those for which it is β. T' is the point on the continuum of the disorder level that separates the two groups. Again, T' divides groups for which the probability of eliciting a positive diagnosis is different, hence a function of the cognitive framework of the diagnosticians, and is to be distinguished from T, the point dividing cases that actually have and do not have the disorder. It is clear that T' and T may be located at different points along the x-axis, but for the sake of discussion we assume that they coincide. Sensitivity is equal to α, and specificity is equal to $(1 - \beta)$. Defining n_1 and n_2 now in the manner illustrated in Figure 4b, the total proportion of correctly diagnosed cases is $[(1 - \beta)n_1 + \alpha n_2]/(n_1 + n_2)$.

At this point we have seen two contrasting models of diagnostic decision making, each characterized by a different diagnos-

Figure 4. Decision-making model for a disorder manifest as a simple dichotomy: (a) two-tiered diagnosability curve (T': threshold below which probability of positive diagnosis is β and at or above which is α); (b) frequency distribution and numbers of cases (n_1 and n_2) in two regions defined by T'.

Table 1
Joint Rating Distribution

	Diagnostician 1	
Diagnostician 2	Positive diagnosis	Negative diagnosis
Positive diagnosis	*a*	*b*
Negative diagnosis	*c*	*d*

Note. Letters correspond to the proportion of cases falling in each cell.

ability curve. Though simplistic, neither is entirely implausible and both have been presented previously in the literature as at least approximations of what actually occurs in some diagnostic situations. They are also identical to threshold models familiar in signal detection theory. We shall now observe the behavior of kappa in response to varying the parameters associated with each model. Let us first consider the impact of changing base rates under the conditions of the diagnostic model shown in Figure 3. As an intermediate step in the analyses it is useful to consider the method of representing interrater agreement data shown in Table 1. The letters in each cell correspond to the proportion of cases categorized in that combination of ways by two raters; for example, a is the proportion given a positive diagnosis by both diagnosticians, d is the proportion given a negative diagnosis by both diagnosticians, and so on. Recalling the earlier assumption that diagnosticians share a common diagnosability curve, the expected values for the proportions b and c are equal, and kappa can be calculated as $(ad - b^2)/[(a + b)(d + b)]$.

It remains now only to calculate a, b, and d from the model parameters. Let P_1 and P_2 be the proportions of positive and negative cases, respectively, in the sample, and let D_1 be the proportion of positive cases clearly recognizable as having the disorder and D_2 the proportion of negative cases clearly recognizable as not having the disorder. The proportion of cases receiving positive diagnoses by both diagnosticians consists of the clearly recognizable positive cases plus the questionable cases for which they both happen to make a positive diagnosis. Therefore

$$a = P_1D_1 + \alpha^2[1 - (P_1D_1 + P_2D_2)], \quad (1)$$

where we have stipulated that for this example $\alpha = P_1$. Reasoning similarly, the proportion of cases given a negative diagnosis by both diagnosticians is

$$d = P_2D_2 + P_2^2[1 - (P_1D_1 + P_2D_2)]. \quad (2)$$

Cell b contains only cases given a positive diagnosis by one diagnostician and a negative diagnosis by the other, so

$$b = P_1P_2[1 - (P_1D_1 + P_2D_2)]. \quad (3)$$

Substituting a, b, and d into the equation for kappa yields expected values for a reliability study with parameters P_1, P_2, D_1, and D_2.

Examining the results in Table 2, the tendency of kappa to vary across base rates is apparent. For example, when the detection rates for positive and negative cases (i.e., sensitivity and specificity) are .50 and .80, kappa values are shown ranging from .62 to .69.

We now consider the behavior of kappa under the conditions of the decision-making model depicted in Figure 4. To promote consistency, we define the proportion of positive cases detected as $D_1 = \alpha$ and the proportion of negative cases detected as $D_2 = 1 - \beta$. As before, P_1 and P_2 are the proportions of positive and negative cases in the sample. A pair of diagnosticians will agree on a positive diagnosis either if they both recognize a positive case or if they both fail to recognize a negative case. Thus

$$a = P_1D_1^2 + P_2(1 - D_2)^2. \quad (4)$$

By similar logic,

$$d = P_2D_2^2 + P_1(1 - D_1)^2 \quad (5)$$

and

$$b = P_1D_1(1 - D_1) + P_2(1 - D_2)D_2. \quad (6)$$

Values of kappa obtained for varying prevalences and detection rates under the conditions of this model are shown in Table 3. Again, kappa tends to vary across rows and down columns of the table. However, this finding is overshadowed by the strikingly low values obtained for all combinations of parameters with this model. Even when the positive and negative detection rates are both .80, indicating a reasonably valid diagnostic pro-

Table 2
Expected Values of Kappa for Diagnosability Curve in Figure 3

	Detection rates: Positive cases								
	.20			.50			.80		
	Negative cases								
Prevalences[a]	.20	.50	.80	.20	.50	.80	.20	.50	.80
.10–.90	.20	.30	.47	.37	.50	.69	.49	.62	.80
.30–.70	.20	.31	.44	.36	.50	.66	.47	.63	.80
.50–.50	.20	.34	.45	.34	.50	.64	.45	.64	.80
.70–.30	.20	.36	.47	.31	.50	.63	.44	.66	.80
.90–.10	.20	.37	.49	.30	.50	.62	.47	.69	.80

[a] Sample proportions of positive and negative cases, respectively.

Table 3
Expected Values of Kappa for Diagnosability Curve in Figure 4

	Detection rates: Positive cases								
	.20			.50			.80		
	Negative cases								
Prevalences[a]	.20	.50	.80	.20	.50	.80	.20	.50	.80
.10–.90	.17	.03	.00	.05	.00	.05	.00	.03	.17
.30–.70	.32	.08	.00	.09	.00	.09	.00	.08	.32
.50–.50	.36	.10	.00	.10	.00	.10	.00	.10	.36
.70–.30	.32	.09	.00	.08	.00	.08	.00	.09	.32
.90–.10	.17	.05	.00	.03	.00	.03	.00	.05	.17

[a] Sample proportions of positive and negative cases, respectively.

cess, the values of kappa are such that a researcher relying on conventional interpretative guidelines for kappa (e.g., Landis & Koch, 1977) could end up disregarding what is in actuality a useful procedure.

In summary, kappa values obtained from samples with different base rates may not be comparable, and by extension, when sample base rates are not representative of population base rates, generalizations of a sample kappa value to populations may be similarly subject to error. Secondly, how kappa varies across base rates differs according to the mathematical characteristics of the particular decision-making process. It is dependent both on the shape of the diagnosability curve and the frequency distribution for case severity. The "base rate problem," therefore, is really only one part of a much broader problem concerning kappa's dependency on a variety of factors unique to each diagnostic situation. Given the complexity and diversity of diagnosability curves that must certainly exist, developing correction formulas to equate kappas obtained from samples with different base rates seems pointless. Further, not only is the comparison of kappa values across different base rates unwarranted, so are comparisons of values obtained from different diagnostic categories or procedures, even though the base rates might be similar. Extending this argument, diagnosticians might differ so much in terms of their diagnosability curves that kappa values obtained for the same disorder, with similar base rates, may still not be comparable across studies. Ultimately, the problem is that there is not a symmetrical mapping between diagnostic processes and kappa values. Instead, there is a many-to-one correspondence, such that any one of a large number of different combinations of diagnosability curves and frequency distributions can result in the same value of kappa. To go backward and try to make specific inferences concerning the nature or quality of a diagnostic procedure on the basis of an obtained kappa value is, without this information, impossible, and, with it, complicated.

If the intent is simply to verify that agreement is at a level beyond chance, then calculating the kappa coefficient and testing whether it differs significantly from 0 can be useful, although other, potentially more flexible methods (cf. Tanner & Young, 1985) that accomplish essentially the same thing should also be considered. In connection with diagnostic decision making, however, verifying that raters are agreeing more often than those who are merely guessing is a minor concern. Instead, what is really required is a method for measuring in a meaningful way the amount of agreement actually present, and in that sense kappa does not meet the needs of the typical researcher or clinician.

Alternatives

The best procedure would be to directly measure the validity of diagnoses by comparing those obtained by the source or method in question with those coming from some more accurate or definitive source. The problem is ordinarily that such criterion diagnoses are comparatively difficult or expensive to obtain. The operative term here, though, is "comparative." What we are really confronted with is a decision where the important issue is whether or not the incremental cost of obtaining criterion diagnoses is matched by a corresponding increase in the information provided beyond agreement data. Because the information value of the kappa statistic appears to be lower than has been supposed, the additional expense of obtaining criterion diagnoses may in fact be justified in situations where this was previously thought not to be the case. Of course, this is complicated by the fact that for psychiatric diagnoses, unambiguous methods for obtaining criterion diagnoses are not just expensive or difficult, they are usually nonexistent. We therefore must be more willing to explore new avenues for obtaining such diagnoses, and where new procedures are not feasible, to make better use of the information we do have. For example, the quality of diagnostic decisions made on the basis of routine interviews could be compared to diagnoses made by a panel of experts provided with extensive case information, who formulate, if not exactly a criterion diagnosis, something very much like one. This reintroduces the possibility of obtaining separate estimates for sensitivity, specificity, true and false positive rates, true and false negative rates, and so on, instead of lumping all information pertaining to the quality of a diagnostic procedure under the rubric of one statistic.

Another possibility is to convey the information concerning a diagnostic procedure's reliability directly in terms of a diagnosability curve. Explicit mathematical methods can be developed to measure the steepness of a diagnosability curve, or the extent to which it clearly distinguishes between cases with high

and low probabilities of positive diagnosis. In some cases the issues causing conditions with different severity levels to be associated with particular probabilities of positive diagnosis may be clear-cut enough to approximate a curve on the basis of a priori knowledge. This may be less feasible for psychiatric diagnoses, where the categories and diagnostic procedures are by their nature less precise, but a thorough consideration of the issues pertaining to a particular type of diagnosis should still reveal at least some information about the curve. It is also possible to construct laboratory studies of a diagnostic process, where diagnosticians are provided with hypothetical clinical vignettes or other similar data in order to determine how their decisions are affected by varying symptomatic information.

By using designs in which patients are diagnosed by several diagnosticians, it is also possible to make inferences concerning the diagnosability curve directly from data obtained in a clinical study. Ordinarily we are accustomed to thinking in terms of studies in which patients are independently diagnosed by two diagnosticians. However, the same general design could be applied using n raters and results summarized in terms of the number of patients receiving a positive diagnosis 1, 2, 3, . . . n times. The resulting frequency distribution can be compared with an expected distribution generated under the conditions of a specific diagnosability curve. Similarly, by considering several diagnosability curves, more or less likely ones can be differentiated on the basis of the observed data.

Table 4 contains hypothetical results of a study in which 100 patients are each evaluated by 10 diagnosticians, and shows the frequency with which various numbers of diagnosticians agree on making a positive diagnosis. Figure 5 contains five diagnosability curves, again hypothetical. These curves are compared in terms of the likelihood that they could account for the observed frequencies. This can be done simply by (a) dividing the x-axis into intervals and calculating the value of $p(x)$ at the midpoint of each; (b) applying the binomial formula to calculate the expected probability distribution of the number of successes on a 10-trial experiment, using the values of $p(x)$ previously defined for each interval; (c) forming a weighted average of the binomial distributions across all intervals to determine the overall distribution—which requires the specification of the shape of $f(x)$ (for illustrative purposes the portion of the normal distribution between $\pm 2\sigma$ is assumed); and finally, (d) using the resulting distribution to determine expected frequencies for various numbers of positive diagnoses. Chi-square is used here to compare expected and observed frequencies, although another procedure, such as least squares or maximum-likelihood estimation could be used in an equivalent way. As shown, the second curve results in the lowest chi-square value and hence provides the best fit. Five curves were used in this example, but the actual computations being trivial by computer, there is no reason why a larger number could not be considered, resulting in the specification of a range of plausible diagnosability curves. In the ideal case, a suitable heuristic would be applied to generate a plausible set of curves for comparison.

Table 4
Hypothetical Distribution of Raters' Judgments

No. making positive diagnosis	Frequency
0	6
1	8
2	7
3	7
4	6
5	5
6	11
7	7
8	9
9	15
10	19

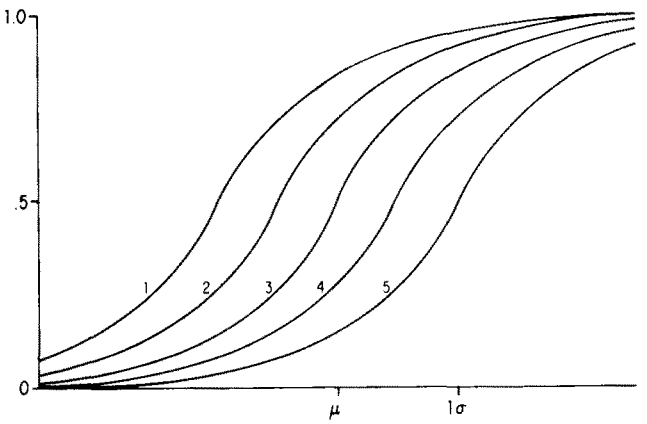

Figure 5. Alternative diagnosability curves compared in terms of the likelihood of leading to observed frequencies in Table 4; μ and σ show the mean and standard deviation of the distribution of cases on the symptom severity dimension (x-axis). (Chi-square values measuring the correspondence of expected frequencies derived from each curve to the hypothetical observed frequencies in Table 4 are Curve 1, 20.3; Curve 2, 3.2; Curve 3, 13.3; Curve 4, 60.7; and Curve 5, 160.7.)

Although it would be highly desirable to develop an entirely new paradigm for thinking about interrater agreement and its measurement, we should also consider the behavior of other statistical indices in response to varying parameters of the diagnostic process to see whether any are more satisfactory than the kappa coefficient. In fact, a wide range of alternatives have been suggested, including p_s, the proportion of specific agreement (Fleiss, 1971) and agreement with majority opinion (Schouten, in press), both of which offer the advantage of providing separate measures of agreement on positive and negative diagnoses, the random error coefficient (Janes, 1979; Maxwell, 1977), the log of the odds ratio, and contingency table analysis (Tanner & Young, 1985), in addition to Yule's Y, which has been considered most recently (Spitznagel & Helzer, 1985). The potential usefulness of each of these indices can be better understood by observing their dependence on changes in base rates across a wide range of plausible diagnosability curves. A computer simulation study of this is currently under way.

References

Atkinson, R. C., Bower, G. G., & Crothers, E. J. (1965). *An introduction to mathematical learning theory.* New York: Wiley.

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

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–383.

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

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

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

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

Kraemer, H. C. (1980). Extension of the kappa coefficient. *Biometrics, 36,* 207–216.

Landis, J. R., & Koch, G. C. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lusted, L. B. (1968). *Introduction to medical decision making.* Springfield, IL: Charles C Thomas.

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

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283–298.

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

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

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

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York: Academic Press.

Tanner, R., & Young, M. A. (1985). Modelling agreement among raters. *Journal of the American Statistical Association, 80,* 175–180.

Uebersax, J. S. (1986). *Validity inference from interobserver agreement data.* Manuscript submitted for publication.

Appendix

Formulas for Calculating Validity Indices

Referring to Figures 1 and 2, the total percentage of correct diagnoses is given by the equation

$$\% \text{ correct} = 1/N \int_0^T q(x)f(x) + 1/N \int_T^\infty p(x)f(x). \quad (7)$$

The sensitivity (*Se*) of the procedure is given by

$$Se = \int_T^\infty p(x)f(x) \div \int_T^\infty f(x), \quad (8)$$

and the specificity (*Sp*) by

$$Sp = \int_0^T q(x)f(x) \div \int_0^T f(x). \quad (9)$$

Received October 17, 1985
Revision received May 15, 1986 ■