Psychological Bulletin
1980, Vol. 88, No. 2, 458–468

Estimating False Alarms and Missed Events From Interobserver Agreement: A Rationale

Kenneth Kaye
Department of Education and Department of Behavioral Sciences,
University of Chicago

In the analysis of temporal, sequential, or contingent relations among events within sessions, the observers' or coders' false-alarm rate and missed-event rate are more useful than coefficients of interobserver agreement as indices of reliability. False-alarm and missed-event probabilities can be estimated from comparisons of two codings, either in the fixed-unit situation or when behavior has been coded in continuous time. Complex problems, such as those that arise when one computes rates in the context of two or more mutually exclusive continuous states, can often be reduced to a simple model, so long as investigators keep in mind the precise meaning of the reliability indices they use.

Complex analyses of sequential data are becoming increasingly important to investigators of be!avior in its natural contexts, including the interpersonal behavior of dyads or groups. The data in such studies often consist of series of observed events (Kaye, 1978). These may be coded continuously as they occur, or in slow motion, or by a stop-frame method, or in unit time intervals checked as containing or not containing an event of a given kind, or even from a transcript in which an utterance, answer, or trial takes the place of unit time.

So long as one is only concerned with a summary measure from each subject or session, such as the frequency or rate with which a particular category of behavior occurs, reliability is a matter of how well the measure discriminates among the subjects (correlation over observers, over sessions, etc.), and validity is a matter of how well the measure generalizes to the subjects' behavior in other situations. This article is concerned with cases in which that is not sufficient. When data are to be microanalyzed—that is, whenever one is concerned with temporal, sequential, or contingent relations among events within a session—three questions arise: (a) How confident am I, when a particular event was coded, that it actually occurred? (b) How confident am I that when a particular event occurred, it was coded? and (c) How precisely did we code the times of occurrence?

Suppose an ethologist observes interaction between two macaques, a mother and her infant. One of the behaviors of interest is "infant activity" (I); each discrete onset of activity is coded. Another category, coded independently of the infant's behavior, is "mother pats, strokes, or jiggles" (M). It is found that this category of maternal behavior occurs in 60% of the 10-sec intervals immediately following "infant active" but has only a 15% likelihood of occurring at other times.[1]

Each of these results is highly sensitive to the observers' two kinds of errors: missed events and false alarms. Thus the simple probability that they will record the event (p_r) is a function of the proportion of intervals in which it actually occurs (p_a), the proportion in which they fail to code it when it does occur (β), and the proportion in which they code it

I am grateful to Donald Fiske, William Kruskal, Lalitha Sanathanan, and Jeanne Altmann for helpful comments on earlier drafts of this article.

Requests for reprints should be sent to Kenneth Kaye, 5835 South Kimbark Avenue, Chicago, Illinois 60637.

[1] These data come from Sackett (1979), but I have taken liberties with his study to simplify the illustration.

when it does not occur (α):

$$p_r = (1-\beta)p_a + \alpha(1-p_a) \\ = \alpha + (1-\alpha-\beta)p_a. \quad (1)$$

Note that p_r is a linear function of p_a, with negative slope. It will underestimate p_a only when $\alpha < (\alpha+\beta)p_a$. Otherwise, p_r will overestimate p_a. The slope is steeper, the greater the proportion of coding errors of either kind. The proportion of missed events alone, β, does not allow one to judge whether p_r is a good estimate of p_a.

Now suppose that the true behavioral contingency between the monkeys is shown in Table 1 under "Actual events"; Each of 1,000 10-sec intervals falls into one of the four cells. Suppose also that for both categories the observer misses 15% of the events that occur ($\beta = .15$), and in only 3% of the intervals when no event occurs, he or she codes a false alarm ($\alpha = .03$). Equation 1 provides the marginal frequencies the observer would be likely to record; the unconditional probabilities of observing I would be [.03 + .085(.82)] = .100, and those of M would be [.03 + .201(.82)] = .195, shown in Table 1 under "Data as coded." Then the total number of cases of I → M in the observer's data (the first cell of Table 1 under "Data as coded") should include only $.85^2$ of the 78 actual cases of I → M, plus .03 × .85 of the 7 cases of I → $\bar{\text{M}}$, plus .03 × .85 of the 123 cases of $\bar{\text{I}}$ → M, plus $.03^2$ of the 792 cases of $\bar{\text{I}}$ → $\bar{\text{M}}$, for a total of about 60. Filling in the other cells by subtraction, one finds that the investigator would be likely to compute the conditional likelihood of "mother pats, strokes, or jiggles," given "infant active," as about .60 when it was actually about .92. The observer would also slightly overestimate the unconditional likelihood or baseline; so the net effect would be to make the contingency look weaker than it really was. In fact, in Table 1 under "Actual events," the likelihood of a maternal response is seven times as great following the onset of "infant activity" as it is at other times, whereas the investigator (Sackett, 1979) thought it was only four times as great.

Now imagine that one had found a more modest contingency, so that I → M appeared to be only twice as great as $\bar{\text{I}}$ → M or less than that. Then if one knew α and β for categories I and M, one could predict whether more substantial differences would be likely to be revealed by a replication in which the coding was done more accurately.

Table 1
Hypothetical Contingency As Occurring and As Coded

	I→	$\bar{\text{I}}$→	Total
	Actual events		
M	78	123	201
$\bar{\text{M}}$	7	792	799
Total	85	915	1,000
	Data as coded		
M	60	135	195
$\bar{\text{M}}$	40	765	805
Total	100	900	1,000

Note. I→ = intervals following "infant active"; $\bar{\text{I}}$→ = all other intervals; M = "Mother pats, strokes, or jiggles"; $\bar{\text{M}}$ = mother does not.

Conditional likelihoods are only one of many kinds of measures used in microanalytic studies, on which false alarms and missed events can have a large impact. For example, the chi-square is a straightforward way of testing contingency between events of particular kinds, but it can be affected by both kinds of errors. The statistical correction procedures (Assakul & Proctor, 1967; Bross, 1954; Hayashi, 1968; Mote & Anderson, 1965) assume that one has good probability estimates of false alarms and omissions.

In this article I define r_β as the probability of coding an event, given that one occurs, and r_α as the probability that a coded event really has occurred. These are more meaningful and more useful measures of reliability than any measure of interobserver agreement can be. Whenever sequential or temporal behavioral observations, codings, or ratings are to be analyzed in relation to one another within a session (microanalysis), the problem of reliability estimation is more than just a technical requirement of journals: It is an indispensable part of the analysis.

In some situations r_α and r_β can be measured directly, though that requires a perfectly accurate record of the actual behavior to compare with the observers' codings. Imagine a situation in which 5-minute videotaped sessions

Table 2
Agreements and Disagreements Between Two Observers

First observer (O_1)	Second observer (O_2)	
	0	1
0	I	D_2
1	D_1	J

can be coded with nearly perfect accuracy by stop-frame methods in 10 hours each, but reasonable agreement with those exhaustive codings can be obtained by watching the tape at full speed and using a microcomputer keyboard. An investigator who has 200 sessions to code might estimate α and β for the efficient method by comparing half a dozen codings with exhaustive transcripts of the same sessions. Projections such as those in Table 1 could then be used to determine in advance whether the αs and βs for the various event categories will be small enough to allow contingencies of a particular magnitude to be detected.

Often, however, it is impossible to obtain a "perfect" record of the events, no matter how slowly and carefully one codes them. The error is not only in the coder's performance but in the nature of the category itself. The "true" onset of "infant activity" is an idealization of the category definition. The concepts of α errors and β errors refer to that idealization: To what extent is "infant activity" an observable event among captive crab-eating macaques?[2] The only available data from which α and β can be estimated are the agreements and disagreements between two codings of the same session.

Estimating α and β in these circumstances depends on a number of assumptions and specifications. There cannot be one correct formula or magic criterion for all investigations. Although I propose several sets of formulas that may have wide utility, my main purpose is to illustrate a way of conceptualizing reliability in signal-detection terms. The kind of reasoning I outline may lead other investigators to different formulas for their own special cases.

Occurrence–Nonoccurrence in Fixed-Time Intervals or Atemporal Units

Let us begin with the situation in which observers attend to a series of units (which may be fixed-time intervals, trials, patients, utterances, etc.) and code each unit 0 or 1 according to some categorical definition. There may be multiple categories, and they may or may not be mutually exclusive; but I am concerned with the reliability of each category singly.

Assume that it is not known when the events actually occurred (i.e., which units truly belong to the category). There is only a pair of codings by two observers, each of whom tried to produce a perfect rendering of the events but who do not agree perfectly with one an-

Table 3
Occurrences and Nonoccurrences Partitioned Into Agreements and Disagreements

	$\bar{O}_2$	O_2	Total
	Units (N) actually containing event		
	(Miss)	(Correct)	
$\bar{O}_1$ (Miss)	$\beta^2 N$	$\beta(1-\beta)N$	βN
O_1 (Correct)	$(1-\beta)\beta N$	$(1-\beta)^2 N$	$(1-\beta)N$
	Units (T–N) actually not containing event		
	(Correct)	(False alarms)	
$\bar{O}_1$ (Correct)	$(1-\alpha)^2(T-N)$	$(1-\alpha)\alpha(T-N)$	$(1-\alpha)(T-N)$
O_1 (False alarms)	$\alpha(1-\alpha)(T-N)$	$\alpha^2(T-N)$	$\alpha(T-N)$

Note. $\bar{O}_2$ = no event is coded by second observer; O_2 = event is coded by second observer; $\bar{O}_1$ = no event is coded by first observer; O_1 = event is coded by first observer.

[2] Observable by the psychologist. The research results will indicate how well those categories correspond to events that are salient and meaningful to the organisms being studied.

other. All one can do is partial the total number of time slots or units coded T into agreements and disagreements, as shown in Table 2. There are, of course, two kinds of agreement:

J = the number of slots in which both observers agreed the event occurred, and

I = the number of slots in which both agreed it did not occur.

Let $D = D_1 + D_2$ = the number of slots in which one observer did and the other did not see an event.

Let $F = D + 2J$ = the total number of ones coded by the two observers.

Traditionally, the question of reliability has been restricted to the question of strength of association in this table. Different coefficients of association have been suggested—Q (Yule & Kendall, 1950), λ_r (Goodman & Kruskal, 1954), π (Scott, 1955), κ (Cohen, 1960), and ϕ (Hays, 1973)—depending on how one calculates the expected chance agreement between the two observers. But all of these coefficients estimate the observers' strength of agreement with one another, not their agreement with an ideal. In the situations dealt with here, on the other hand, the question of reliability has to do with the relation between the average coder and the actual events he or she was supposed to code (as ideally defined). I am concerned with the probability of false alarms and the probability of missed events on the part of any one observer as he or she goes on to code subsequent sessions. For this purpose one should assume that the numbers J, D, I, and F actually resulted from the eight cells of Table 3, in which T has been divided into the number of units in which an event actually occurred (an unknown, N) and the number in which no event occurred $(T - N)$. Note that β, the rate of missed events, is defined as a proportion of N: It is the probability that an event that occurs will be missed. Similarly, α is defined as a proportion of $(T - N)$: It is the probability of coding an event when no event occurred. Thus

$$r_\beta = 1 - \beta,$$
$$r_\alpha = (1 - \beta)N/[(1 - \beta)N + \alpha(T - N)].$$

The problem, then, is to obtain estimates of α, β, and N from our counts of J, D, and I. The expected values of J, D, and I can be seen from Table 3:

$$E(J) = (1-\beta)^2N + \alpha^2(T-N), \qquad (2)$$

$$E(D) = 2\beta(1-\beta)N + 2\alpha(1-\alpha)(T-N), \qquad (3)$$

$$E(I) = \beta^2N + (1-\alpha)^2(T-N). \qquad (4)$$ [3]

Since the three equations sum to T, they are two independent equations in three unknowns, which means they can only be solved if the investigator makes some further assumption about α, β, or N. For example, if $\alpha = 0$ (i.e., there are no false alarms),

$$J = (1 - \beta)^2N,$$
$$D = 2\beta(1 - \beta)N,$$

thus

$$F = 2J + D = 2(1 - \beta)^2N + 2\beta(1 - \beta)N$$
$$= 2(1 - \beta)N,$$

thus

$$D/F = \beta,$$

$$r_\beta = 1 - D/F = \frac{F - D}{F} = \frac{J}{J + \frac{1}{2}D}. \qquad (5)$$

This formula has actually been used for some time (Wright, 1967) and is superior to the more popular $r_\beta = J/(J + D)$, "agreements over agreements plus disagreements," because the latter assumes that when two observers disagreed, both of them must have been wrong (i.e., It counts all disagreements D as errors.) In Equation 5, only half of the disagreements appear in the denominator, reasonably enough; the event either occurred or did not occur, so only one of the observers can have been wrong. What is harder to defend is the assumption of no false alarms, the assumption that when the observers agreed that an event occurred, it did. The only utility of this assumption is to place an upper limit on β: If all disagreements were due to missed events, then $\beta = D/F$. However, it also means that the observers must always underestimate N, since they miss events but code no false alarms.

In the more general case, Table 3 gives a basis for expressing the most likely value of N. Each observer codes this many ones on the average:

$$\tfrac{1}{2}F = (1 - \beta)N + \alpha(T - N).$$

[3] Henceforth the "$E(\,)$" notation will be understood.

Thus

$$\tfrac{1}{2}F - \alpha T = N - \beta N - \alpha N,$$

$$N = \frac{\tfrac{1}{2}F - \alpha T}{1 - \alpha - \beta}.$$

If $\alpha = 0$, then

$$N = \frac{F}{2(1-\beta)} = \frac{F^2}{2(F-D)} = \frac{F^2}{4J}. \qquad (6)$$

This can be shown to be identical to Feller's (1957) maximum likelihood estimate of N, in which one samples a population with replacement (e.g., capturing and tagging wildlife) and then samples again (Goodman, 1953). The proportion of tagged animals in the second sample is a direct analogy to the proportion of agreements with the first observer, among a second observer's codes. In fact the model was developed for exactly that purpose by Geiger and Werner (1924); the observers were scanning photographs of particles in a cloud chamber. They were not concerned about false alarms; neither is a game warden.[4]

Now imagine the opposite case, in which $\beta = 0$:

$$N = \frac{\tfrac{1}{2}F - \alpha T}{1 - \alpha},$$

$$D = 2\alpha(1-\alpha)\left(T - \frac{\tfrac{1}{2}F - \alpha T}{1-\alpha}\right) = 2\alpha T - \alpha F,$$

$$\alpha = D/(2T - F). \qquad (7)$$

This is an upper limit on α: the rate of false alarms if those were responsible for all the disagreements. From the definition of r_α, in this case,

$$r_\alpha = \frac{\tfrac{1}{2}F - \alpha T}{1-\alpha}\Big/\tfrac{1}{2}F. \qquad (8)$$

Substituting Equation 7 for α in Equation 8 yields

$$r_\alpha = \frac{2FT - F^2 - 2DT}{2FT - F^2 - FD}. \qquad (9)$$

Somewhere between Equations 5 and 9 lie the more likely cases in which observers make both kinds of errors. If N can be guessed, of course, any two of Equations 2–4 can be solved for α and β. In the general case we could reduce the number of unknowns by assuming some ratio of false alarms to misses,

$$\alpha(T - N) = k\beta N, \qquad (10)$$

but it is difficult to conceive how a researcher's experience might provide a value for k without also directly providing α and β. However, many investigators implicitly assume that $k = 1$. For example, if the average number of events seen by one coder is assumed to be a good estimate of the actual number of occurrences, $E(N) = \tfrac{1}{2}F$, the implication can only be that the number of false alarms on the average equals the number of missed events. Otherwise one is consistently either underestimating or overestimating the likelihood of occurrence of the event category in question, p_a or N/T. Although there is no reason to believe that $\alpha(T - N) = \beta N$ will actually be true, at least it has the advantage of yielding medium rather than extreme values for α and β. If a convention is desirable for practical purposes, this is the convention I recommend. Let us see what formulas it produces for α, β, r_α, and r_β. From Equations 3 and 10,

$$D = 2\beta(1-\beta)N + 2\beta\left(\frac{N}{T-N}\right) \times \left(1 - \beta\frac{N}{T-N}\right)(T-N)$$

$$= 2N\left(\beta - \beta^2 + \beta - \beta^2\frac{N}{T-N}\right),$$

$$D/2N = D/F = 2\beta - \beta^2\left(1 + \frac{N}{T-N}\right).$$

It is easiest to begin by computing the ratio $Q = N/(T - N)$, so that the quadratic equation[5] can be solved for β:

$$\beta = \frac{1 - \sqrt{1 - D(1+Q)/F}}{1+Q}, \qquad (11)$$

$$\alpha = \beta Q,$$

$$r_\alpha = (1-\beta)N/[(1-\beta)N + \alpha(T-N)],$$

$$= (1-\beta)N/(N - \beta N + \beta N)$$

$$= 1 - \beta = r_\beta. \qquad (12)$$

If an investigator feels the number of missed events probably exceeds the number of false alarms ($N > F/2$), then r_β would lie between

[4] This is what Thoreau (1927) meant when he wrote, "Some circumstantial evidence is very strong, as when you find a trout in the milk" (p. 58).

[5] Take the lesser root of $-(1+Q)\beta^2 + 2\beta - D/F = 0$ on the assumption that β is never more than .5.

Equation 5 and (1 − Equation 11), and r_α would lie between (1 − Equation 11) and 1. If the number of false alarms is greater than the number of missed events ($N < F/2$), then r_β would lie between (1 − Equation 11) and 1, and r_α would lie between Equation 9 and (1 − Equation 11).

Continuous-Time Coding

When the data consist of events coded continuously in real time, *agreement* has a different meaning. In counting the number of first-observer (O_1) events that fall close to second-observer (O_2) events, one must use some criterion tolerance lag t. This changes the definition of β, which must be something like the average rate of failure to code an event within $\pm\, t$ of the average reaction time. Furthermore:

1. Two of O_1 events could fall within the criterion interval $\pm\, t$ surrounding a single O_2 event; only one of these can be considered to be an agreement with that event.
2. In the fixed-interval case, I considered that either zero or one false alarm might occur in any interval when a true event did not occur; in the continuous case, false alarms might occur in any number and at any proximity to correctly coded events.
3. There is no fixed number of units to partition as I did in Table 3.

For these reasons I must make some simplifying assumptions about α, β, and N before constructing any formulas.

One might begin by imagining that $\alpha = 0$. But what happens to the βN events that each partner fails to code within the criterion t? At least some of these are likely to be coded late rather than missed entirely. If coded late, they will often contribute to D and occasionally to J. So one must assume that $\alpha > 0$, if only to include the late-coded events. In fact I make no distinction between late-coded events and false alarms, treating them as a random pool of codes expected to fit a Poisson process with parameter λ:

$$\lambda = \frac{\text{the total number of such ``random'' codes}}{\text{the total time period over which events occurred}}.$$

Manageable formulas will result only if it is also assumed that the total number of false alarms plus late codes will add up to approximately βN, the number of missed events, so that $E[(O_1 + O_2)/2] = (F/2) = N$.

Now J and D can be partitioned. J consists of the agreements on true events, $(1-\beta)^2 N$, plus some accidental number of pairs of codes that happen to fall within $\pm\, t$ of one another. D consists of all the remaining codes. Each observer's expected share of the latter, or $\frac{1}{2}D$, ought to be equal to the number of times one of his or her own "random" codes failed to match by chance one of the other observer's "random" codes. In the fixed-interval case, α is defined as the likelihood of a false alarm in any given interval; here I define it as the likelihood of any "random" code falling within the tolerance criterion $\pm\, t$. If each observer produces on the average N codes, of which $(1-\beta)^2 N$ accurately match those of the other observer, then there remain $N - (1-\beta)^2 N = (2\beta - \beta^2)N$ additional opportunities to agree, each with probability α. So each partner's failures to match the other by chance are

$$\tfrac{1}{2}D = (1-\alpha)(2\beta - \beta^2)N,$$

$$D/F = (1-\alpha)(2\beta - \beta^2).$$

Now $(1-\alpha)$ is the chance of finding, for any point in time selected at random, that the next occurrence in the Poisson process is more than $2t$ away:

$$1 - \alpha = e^{-\lambda \cdot 2t},$$

where $$\lambda = (2\beta - \beta^2)N/T,$$

thus $$D/F = (2\beta - \beta^2)e^{[-(2\beta - \beta^2)Ft/T]}. \qquad (13)$$

One can see how greatly the estimate of β from D/F will depend on T, which (as in the fixed-interval case) must be the actual effective period in which events were free to occur (not, e.g., the total length of a videotape including time-outs).

Equation 13 can be solved by a computer (by iteration), but it is more convenient to obtain β by referring values of D/F and Ft/T to Figure 1. Follow the curve corresponding to

[6] Actually, this formula slightly underestimates β and overestimates α because it does not take account of a small reduction in λ every time a pair of false alarms happens to match. In practice, however, this error will always be too minuscule to affect the rounded estimate.

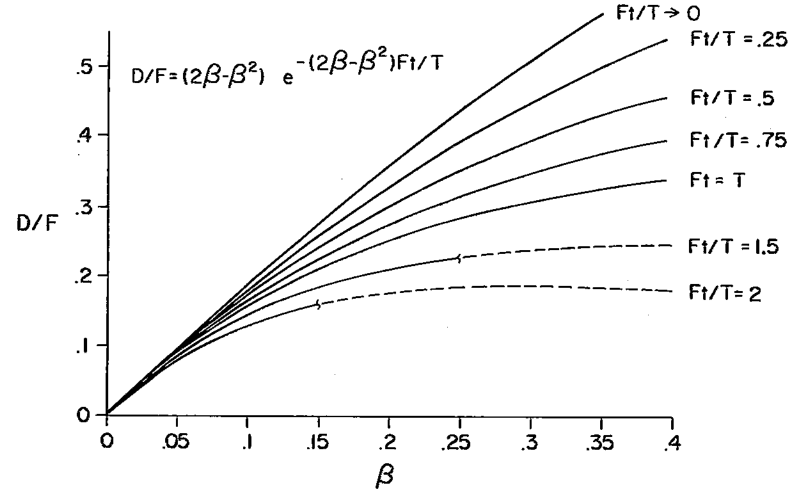

Figure 1. Ratio of disagreements to total events coded by two observers, as affected by the criterion lag accepted for matches and as affected by β when the number of missed events is balanced by the number of false alarms plus late-coded events.

the ratio Ft/T (or interpolate between two curves) to the point corresponding to D/F, from which the estimate of β can be read along the abscissa. Then

$$\alpha = 1 - e^{[-(2\beta-\beta^2)Ft/T]},$$

but α may be of little interest, since $r_\alpha = r_\beta = 1 - \beta$. Note that the dotted portions of the curves in which $Ft > T$ are useless; t is unlikely to be an appropriate tolerance interval if it is greater than T/F. Note also that

$$\lim_{Ft/T \to 0} D/F = 2\beta - \beta^2,$$

so that

$$\lim_{Ft/T \to 0} r_\beta = \sqrt{1 - D/F}.$$

To illustrate, suppose a 5-min film is coded twice with the result that $F = 125$. Using a 2-sec tolerance, we find that 54 pairs of codes match, so $D = 17$. If there were no false alarms and no late codes (only missed events), $r_\beta = 1 - D/F = 86\%$ (Equation 5; T and t do not enter into the calculations if there are no false alarms). With as many as βN false alarms, Figure 1 shows that when $Ft/T = .83$, the obtained percentage disagreement (i.e., D/F) of about 14% would be expected when $\beta = .08$. So reliability $r_\alpha = r_\beta = 92\%$ can be estimated rather than $r_\alpha = 100\%$, $r_\beta = 86\%$.

Enduring Events

Frequently coders are asked to record the onset and offset of some continual state, and the investigator is interested in what else goes on between an onset and the following offset, for example, the rate of wives' smiling when their husbands are looking at them versus when the husbands are looking elsewhere. The wives' smiling, a discrete event, is coded with a certain reliability whose r_α and r_β can be estimated by one or the other of the sets of formulas given previously. But the enduring behavior, husbands' gaze direction, appears at first to pose a more complex problem. One could separately estimate the reliability of coding onsets or of coding offsets. When an onset is coded a few seconds late, however, the effect on the data depends on how long the behavior continues. To integrate information about the onset reliability, offset reliability, and distribution of durations would be difficult enough, but it is further complicated by the fact that some kinds of behavior continue to clamor for the coder's attention so that a code may be late but is unlikely to be missed entirely, whereas other kinds of behavior are codable only at their onset, which when missed leads to the entire duration being added to the offs instead of to the ons.

Fortunately, these considerations turn out to be superfluous. In the example just given, the concern is the likelihood of misclassifying a smile as occurring during an off versus an on: What is the probability that at any given time when the husband was looking, the coding procedure resulted (for whatever reason) in the conclusion that he was not looking? That probability β is equal to the proportion of true looking time that the coder missed, α is the proportion of true nonlooking time that was coded as looking, and N is the actual time looking at the wife, and so long as the coders' errors are independent, these unknowns can be estimated from Equations 5–12. In this case J is the number of seconds that both coders are in agreement that the husband is looking, I is the number of seconds both are in agreement he is not looking, and $D = T - J - I$.

If S_1 and S_2 are the actual rates of smiling during looking and nonlooking, respectively, then modest errors in coding the looking category alone can produce significant distortions in the relative magnitudes of the two rates. Figure 2 shows how rapidly the distortion increases as a function of α and β. (These curves result when the enduring event—e.g., looking—adds up to 50% of the total time. When the state producing the higher rate occurs during a greater proportion of time, the distortion is much greater; when it occurs during a smaller proportion of time, the same α and β have much smaller effects.)

Confidence Limits

A frequent puzzle for investigators is how large a sample of sessions should be coded twice to estimate reliability. The answer is not any particular proportion of the total (though one does want to sample throughout the series to check for drift) but enough to establish confidence limits for α, β, r_α, and r_β. Since these best guesses are all derived from direct measurements—basically from D/F—they will vary as D/F varies due to sampling error. Regardless of the soundness of one's assumptions and even if a bracketing method is used, as illustrated previously, to provide probable bounds for the true value, the accuracy of those bounds depends on how well D/F is sampled. One can estimate its mean and variance from a sample of sessions (calculating D/F separately from each session), or if that is impossible one can generate the theoretical sampling distribution. Since D/F is a proportion based on a sample of size N, its theoretical standard error is

$$\sigma_{D/F} = \sqrt{\frac{D/F \times 2J/F}{N}} = \sqrt{4JD/F^3}.$$

From this theoretical value or preferably from the empirical standard error, one can find 95% confidence limits for D/F and use these to get an idea of the upper and lower bounds for r_β. The population to which this estimate then applies is the set of coding reliabilities of the same scheme by the same coders for any random sample of the same class of behavior produced with about the same frequency by similar subjects under similar conditions.

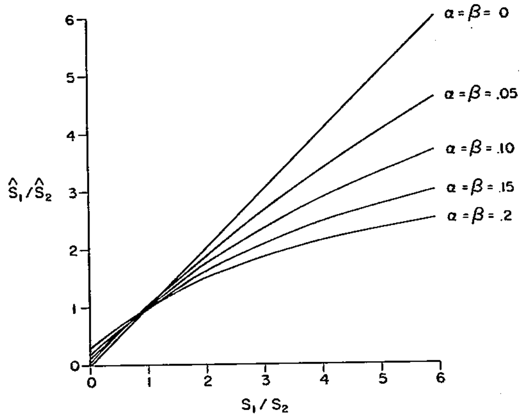

Figure 2. Ratio of two rates of behavior, as affected by reliability of coding the continual state on which they are conditioned. (Curves shown are for $N = T - N$.)

Precision

Often independent codings are synchronized so that hypotheses can be tested regarding the latencies between events. However, for each category there is a variance in observer reaction time. This has an important implication: When Event X actually precedes Event Y by a small latency, Y might occur before X in the record. Define the confidence lag L as the minimum latency at which one can be 95% sure the events actually occurred in the order they appear to have occurred in. Since the variance in their latencies is likely to be the sum of the variances in the two independent categories, the confidence lag is the sum of the times corresponding to the one-tailed 95% confidence intervals for Events X and Y. However, those intervals in turn may have to be deduced from the distribution of interobserver lags for the J agreed-on events. In other words, the 95% confidence lag is the mean of the 95% agreement lags for X and Y. These are easily computed. When counting agreements, one should accumulate the squared logs of the lags. Let g be the absolute lag between observers of any one event. Then

$$\log \sigma_L = \sqrt{\frac{\Sigma(\log^2 g)}{J-1}},$$

$$L = \text{antilog } 2\sqrt{\frac{\Sigma(\log^2 g)}{J-1}}.$$

One converts back from the log representing two standard errors (i.e., where $z = 2$) to get an approximate 95% confidence lag in seconds. If α was reasonably small, then few of the J agreements were random matches (double false alarms), so their effect on the lognormal distribution of true agreements should be negligible. There can be a problem, however, if L approaches t, the tolerance used for agreements. Inspection of any table showing areas under the normal curve as a function of z scores will make it clear that so long as t is at least 25% greater than L, no more than 1% of agreements will have been discounted. As L approaches t, a considerable number of interobserver agreements must have been discounted because their lags exceeded t. This in turn means that L has been underestimated (i.e., based on a distribution without tails) because the sample J is a restricted sample of the actual interobserver lags. Agreement should be recomputed using a larger tolerance ± t. This will lead to a greater J, a smaller D, a smaller β (but only somewhat smaller, because Equation 13 corrects for the higher number of random matches), and a larger L so that one does not interpret interevent times with a greater precision than the coding has warranted.

Conclusions and Discussion

I propose that reliability of coding in sequential units, fixed-time intervals, or continuous time ought to be conceptualized in terms of a false-alarm rate as well as a missed-event rate. Agreement between coders is only a preliminary calculation to be used in the more important step of estimating the probable error rates of an individual coder. Similarly, the lags in coding times between two observers in the continuous-time case are data to be used in estimating the precision of an individual observer. When certain broad principles are understood, different methods of coding can be treated as equivalent for purposes of estimating accuracy and precision. In the remainder of this article, I try to establish some restrictions and principles that, if observed, allow generalization of the preceding formulas to different research paradigms.[7]

There are as many kinds of reliability as there are steps in the gathering and analysis of data. Psychologists are concerned with the reliability of their recording equipment (*sensitivity*), of transcribing (*accuracy*), of coding (*agreement*), of assignment of times or other quantitative measurements to codes (*precision*), of selected measures (*robustness*), of assessments of individuals' performance at different times (*stability*), and of whole experiments (*replicability*). Any of these judgments can be reduced, under certain circumstances, to a question of correlation. Reliability of individual events may be of less concern than whether the scores produced from whole sessions are generalizable measures (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

As soon as one decides to analyze relations among events occurring within the session, especially when one's whole corpus of data may come from a small number of sessions, questions of sensitivity, accuracy, agreement, and precision have to be framed in terms of confidence that events actually occurred when they were coded and vice versa. Generalizability, on the other hand, becomes a matter of replication.

[7] Sanathanan (1972) has developed estimation methods for the particle-scanning experiments mentioned earlier, at the level of complexity that is necessary when some of my assumptions are dropped (e.g., the assumption of equally reliable coders and the assumption of independent errors). Similar formulas might be developed for cases in which false alarms can occur, but the resulting formulas are likely to prove unwieldy and to provide more precision than a psychologist can really use.

The emphasis in the literature on strength of agreement between coders was a misguided attempt to apply the logic of generalizability or correlation at the microanalytic level. This article argues instead in favor of using agreement data to estimate reliability indices that will help to correct for systematic biases in conditional rates, transitional likelihoods, and the like.

Fortunately, one does not need half a dozen different sets of formulas to deal with sensitivity in filming or recording, accuracy in transcribing, agreement in coding, and so forth, provided a fundamental principle is observed: EVERY RELIABILITY ESTIMATE IS A REPORT ON A BOUNDED PHASE OF THE PROCEDURE RATHER THAN ON THE ADEQUACY OF THE DATA GATHERING AS A WHOLE. Measurements can be made on the agreement between repeated applications of any particular phase of the procedure, for example, from the transcription to the derivation of certain complex combinations of coded events, without worrying about the individual steps in that process.

However, when a phase of the procedure is chosen whose reliability is to be estimated, there should be some justification for one's confidence in all of the prior phases. For although their reliability or unreliability is not measured, those prior phases (filming, transcribing, even selection of a baboon troop or an interviewer) are bound to affect the measures obtained. A particular camera angle may make a subset of the relevant events highly salient and others imperceptible. There is no guarantee that this reliably coded subset will be systematically related to the phenomena under study; and even if results are obtained, the study will not be replicable. So there is good reliability, and there is bad reliability. RELIABILITY THAT IS PURCHASED AT THE COST OF GENERALIZABILITY IS NO VIRTUE.

Consider an example of bad reliability that occurs with disturbing frequency in published articles. A set of transcripts includes hesitations in the subjects' speech: "uh," "er," and so forth. A coding category applied to these transcripts—the category "hesitation"—turns out to be highly reliable. But this is of no redeeming importance. What matters is that the original transcribing should have been reliable; and this can only be discovered by hiring a second transcriber. The tape, too, could be a source of bad reliability, a question that can only be answered by making two tapes.

It is well established that observers' reliability is much higher when they know it is being tested than when they are engaged in routine coding. Reliability is also inflated by any biases in the observers' instructions that induce expectations about the fequency or context of particular behavior and by any predictable relations among categories. Kazdin (1977) has reviewed evidence for these and other sources of bad reliability. Good reliability, established independently for each of the categories one is going to analyze in relation to other categories, contributes to validity; without it, one can never be sure that the categories have meaning beyond one's own coding procedure.

Many factors affect reliability. The discriminability of the events and the clarity of the category definitions are obvious factors, but there are more complex ones. For example, if the coders are simultaneously monitoring two or more categories, their attention to one event is likely to impair their coding of others. Another problem arises when the events of concern are not the ones directly coded but are some derived composite of those events. How can the reliability of the higher order category be estimated from the separate coefficients for each of the lower order ones?

A third basic principle makes it possible to ignore complications of those kinds: THE RELIABILITY THAT MATTERS IS THE RELIABILITY OF THE DATA THAT WILL ACTUALLY BE USED IN THE ANALYSIS, AFTER IT HAS BEEN RECODED, TRANSFORMED, COMBINED, CONCATENATED, OR SMOOTHED IN PRELIMINARY WAYS. One should not compute reliability for lower order events at all. Let the computer or the research assistant construct the events that are actually going to be used (e.g., "mother pats, strokes, or jiggles") separately from the two independent original codings. The agreements and disagreements between the results of this process are what one would use as the values of J, D, and so forth. One can estimate r_α and r_β from any two records of a sequence of events, regardless how those records were produced and regardless of the nature of the categories they represent.

The whole investigation should be evaluated from the point of view of these principles before and after estimating reliability. The formulas described previously provide no guarantee against misconceiving the limitations of one's data. Happily, if the three principles are kept in mind, a few formulas can be used across a wide variety of research domains, types of observational category, and questions. For example, the formulas for reliability of events that have a duration (i.e., that are either on or off at any point in time), coded continuously, turned out to be the same as those for codes checked once per fixed interval. The choice of formulas depends on the logic of the questions one is asking about the data rather than on the method used to code the data.

References

Assakul, K., & Proctor, C. H. Testing independence in two-way contingency tables with data subject to misclassification. *Psychometrika*, 1967, *32*, 67–76.

Bross, I. Misclassification in 2 × 2 tables. *Biometrics*, 1954, *10*, 478–486.

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

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements*. New York: Wiley, 1972.

Feller, W. *An introduction to the theory of probability and its applications* (Vol. 1). New York: Wiley, 1957.

Geiger, H., & Werner, A. Die Zahl der von Radium ausgesandten α-Teilchen. I. Teil: Szintillationszählungen. *Zeitschrift für Physik*, 1924, *21*, 187–215.

Goodman, L. A. Sequential sampling tagging for population size problems. *Annals of Mathematical Statistics*, 1953, *24*, 56–59.

Goodman, L. A., & Kruskal, W. Measures of association for cross classification. *Journal of the American Statistical Association*, 1954, *49*, 732–764.

Hayashi, C. Response errors and biased information. *Annals of the Institute of Statistical Mathematics, Tokyo*, 1968, *20*, 211–228.

Hays, W. L. *Statistics for the social sciences*. New York: Holt, Rinehart & Winston, 1973.

Kaye, K. *CRESCAT: Software system for analysis of sequential or real-time data*. Chicago: University of Chicago Computation Center, 1978.

Kazdin, A. Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, 1977, *10*, 97–105.

Mote, V. L., & Anderson, R. L. An investigation of the effect of misclassification on the properties of χ^2 tests in the analysis of categorical data. *Biometrika*, 1965, *52*, 95–109.

Sackett, G. The lag-sequential analysis of contingency and cyclicity in behavioral interaction research. In J. Osofsky (Ed.), *Handbook of infant development*. New York: Wiley, 1979.

Sanathanan, L. Models and estimation methods in visual scanning experiments. *Technometrics*, 1972, *14*, 813–829.

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 1955, *19*, 321–325.

Thoreau, H. D. Journal entry of November 11, 1850. In O. Shepard (Ed.), *The heart of Thoreau's journals*. Boston: Houghton Mifflin, 1927.

Wright, H. *Recording and analyzing child behavior*. New York: Harper & Row, 1967.

Yule, G., & Kendall, M. G. *An introduction to the theory of statistics*. London: Griffin, 1950.

Received November 26, 1979 ■