4

Assessing observer agreement

4.1 Why bother?

Imagine that we want to study how objects are used in communication between mothers and their 15-month-old infants, and have made a number of videotapes of mothers and infants playing together. We might focus our attention on those times when either the mother or the infant tries to engage the other's interest in some object, and then describe what those attempts look like and what their consequences are. After viewing the videotapes several times, sometimes in slow motion, we might be convinced that we had detected all such episodes, had accurately described them, and were now ready to make statements about how mothers and infants go about attempting to interest each other in objects.

We should not be surprised, however, if other investigators do not take our conclusions as seriously as we do. After all, we have done nothing to convince them that others viewing the videotapes would see the same things, much less come to the same conclusions. We are probably all aware how easy it is to see what we want to see, even given the best of intentions. For that reason, we take elaborate precautions in scientific work to insulate measuring procedures from the investigator's influence.

When measurements are recorded automatically and/or when there is little ambiguity about the measurement (for example, the amount of sediment in a standard sample of seawater), as is often the case in the "hard" sciences, the problem of investigator bias is not so severe. But in observational studies, especially when what we call "socially based" coding schemes are being used, it becomes especially important to convince others that what was observed does not unduly reflect either the investigator's desires or some idiosyncratic worldview of the observer. The solution to the first problem is to keep observers naive as to the hypotheses under investigation. This in fact is done by most investigators, judging from the reports they write, and should be regarded as standard practice in observational work. The solution to the second problem is to use more than one observer and to assess how well they agree.

Accuracy

The major conceptual reason for assessing interobserver agreement, then, is to convince others as to the "accuracy" of the recorded data. The assumption is, if two naive observers independently make essentially similar codings of the same events, then the data they collect should reflect something more than a desire to please the "boss" by seeing what the boss wants, and something more than one individual's unique and perhaps strange way of seeing the world.

Some small-scale studies may require only one observer, but this does not obviate the need for demonstrating agreement. For example, in one study Brownlee and Bakeman (1981) were concerned with communicative aspects of hitting in 1-, 2-, and 3-year-old children. After repeated viewings of 9 hours of videotape collected in one day-care center, they developed some hypotheses about hitting and a coding scheme they thought useful for children of those ages. The next step was to have a single observer collect data "live" in another day-care center. There were two reasons for this. First, given a well-worked-out coding scheme, they thought observing live would be more efficient (no videotapes to code later), and second, nursery school personnel were concerned about the disruption to their program that multiple observers and/or video equipment might entail. Two or more observers, each observing at different times, could have been used, but Brownlee and Bakeman thought that using one observer for the entire study would result in more consistent data. Further, the amount of observation required could easily be handled by one person. Nonetheless, two observers were trained, and agreement between them was checked before the "main" observer began collecting data. This was done so that the investigators, and others, would be convinced that this observer did not have a unique personal vision and that, on a few occasions at least, he and another person independently reported seeing essentially the same events.

Calibration

Just as assuring accuracy is the major conceptual reason, so calibrating observers is probably the major practical reason for establishing interobserver agreement. A study may involve a large number of separate observations and/or extend over several months or years. Whatever the reason, when different observers are used to collect the same kind of data, we need to assure ourselves that the data collected do not vary as a function of the observer. This means that we need to calibrate observers with each other or, better yet, calibrate all observers against some standard protocol.

Reliability decay

Not only do we need to assure ourselves that different observers are coding similar events in similar ways, we also need to be sure that an individual observer's coding is consistent over time. Taplin and Reid (1973) conducted a study of interobserver reliability as a function of observer's awareness that their coding was being checked by an independent observer. There were three groups – a group that was told that their work would not be checked, a group that was told that their work would be spot-checked at regular intervals, and a group that was told that their work would be randomly checked. Actually the work of all three groups was checked for all seven sessions. All groups showed a gradual decay in reliability from the 80% training level. The no-check group showed the largest decay. The spot-check group's reliability increased during sessions 3 and 6, when they thought they were being checked. The random-check group performed the best over all sessions, though lower than the spot-check group on session 3 and 6.

Reliability decay can be a serious problem when the coding process takes a long time, which is often the case in a large study that employs a complex coding scheme. One solution to the problem was reported by Gottman (1979a). Gottman obtained a significant increment in reliability over time by employing the following procedure in coding videotapes of marital interaction. One employee was designated the "reliability checker"; the reliability checker coded a random sample of *every* coder's work. A folder was kept for each coder to assess consistent confusion in coding, so that retraining could be conducted during the coder's periodic meetings with the reliability checker. To test for the possibility that the checker changed coding style for each coder, two procedures were employed. First, in one study the checker did not know who had been assigned to any particular tape until after it was coded. This procedure did not alter reliabilities. Second, coders occasionally served as reliability checkers for one another in another study. This procedure also did not alter reliabilities. Gottman also conducted a few studies that varied the amount of interaction that the checker coded. The reliabilities were essentially unaffected by sampling larger segments, with one exception: The reliabilities of infrequent codes are greatly affected by sampling smaller segments. It is thus necessary for each coding system to determine the amount that the checker codes as a function of the frequency of the least frequent codes.

What should be clear from the above is that investigators need to be concerned not just with inter-, but also with intraobserver reliability. An investigator who has dealt with the problems of inter- and intraobserver agreement especially well is Gerald Patterson, currently of the Oregon Social Learning Center. Over the past several years, Patterson and his coworkers have trained a number of observers to use their coding schemes.

Although observers record data live, training and agreement assessments depend on the use of videotapes. First, presumably "correct" codings or "standard" versions were prepared for a number of videotaped sequences. Then these standards were used to train new observers, who were not regarded as trained until their coding reached a preset criterion of accuracy, relative to the standard. Second, observers periodically recoded standard versions, and their agreement both with the standard and with their own previous coding was assessed. Such procedures require time and planning, but there is probably no other way to ensure the continued accuracy of human observers when a project requires more than one or two observers and lasts for more than a month or two.

4.2 Reliability versus agreement

So far in this chapter we have used the term "observer agreement," yet the term "observer reliability" often occurs in the literature. What, if any, is the difference? Johnson and Bolstad (1973) make a nice distinction. They argue that agreement is the more general term. It describes, as the word implies, the extent to which two observers agree with each other. Reliability is the more restrictive term. As used in psychometrics, it gauges how accurate a measure is, how close it comes to "truth." Hence when two observers are just compared with each other, only agreement can be reported. However, when an observer is compared against a standard protocol assumed to be "true," then observer reliability can be discussed.

Others would argue that reliability is the more general term. Interobserver agreement only addresses potential errors among observers and ignores many other sources of potential errors, which in the context of observational research may be many (Pedhazur & Schmelkin, 1991, pp. 114–115, 145–146). Yet, when two observers independently agree, the usual presumption is that they are therefore accurate, even though it is possible, of course, that they simply share a similar but nonetheless deviant worldview.

But this presumption is questionable because other sources of error may be present. In observational research (and in this book as well), interobserver agreement is emphasized, but it is important to remember that, although important, indices of interobserver agreement are not indices of reliability, and that reliability could be low even when interobserver agreement is high.

Not wishing to shortchange a complex topic (i.e., assessment of reliability), we would nonetheless argue that there is some merit in preparing standard protocols, presumed true, which can then be used as one, simple index of *observer reliability*. If the first reason for assessing observer agreement is to assure others that our observers are accurate and our procedures replicable, and the second reason is to calibrate multiple observers, then a third reason is to assure ourselves as investigators that observers are coding what we want (i.e., are seeing the world as we do). And one way to test this is to let observers code independently a sequence of events for which we have already prepared a standard protocol. Clearly, this is easiest to do when videotaped behavior or transcripts of conversations are coded, and less easy to do when live behavior is coded. However, such procedures let us speak, albeit relatively informally, of "observer reliability" (assuming of course that investigators are relatively infallible), and they also give investigators additional confidence in their observers, a confidence that probably becomes more important, the more socially based the coding scheme is.

In sum, "reliability" invokes a rich and complex psychometric tradition and poses problems that lie well beyond the scope of this book. From this point of view, "interobserver agreement" is the more limited and straightforward term. Moreover, it is the one that has been emphasized in observational research and, consequently, is addressed in the remainder of this chapter. The question now is, how should observer agreement be assessed and computed?

4.3 The problem with agreement percentages

Perhaps the most frequently encountered, and at the same time the most misleading, index of observer agreement is a percentage of some sort. This is usually referred to as a "percentage of agreement" and in its most general form is defined as follows:

$$P_A = \frac{N_A}{N_A + N_D} \times 100$$

P_A refers to the percentage of agreement, N_A the number of agreements, and N_D the number of disagreements. In any given application, the investigator would need to specify further the recording unit used (events or intervals), which is after all the basis for determining agreement and disagreement, and exactly how agreement and disagreement are defined.

For example, Tuculescu and Griswold (1983), in their study of prehatched chickens (section 2.10), defined four kinds of embryonic distress calls (Phioo, Soft Peep, Peep, and Screech). Observers coded events. Whenever an event of interest occurred, they recorded which it was, when it began, and when it ended. Given this coding scheme and this recording strategy, observer agreement could have been computed as follows.

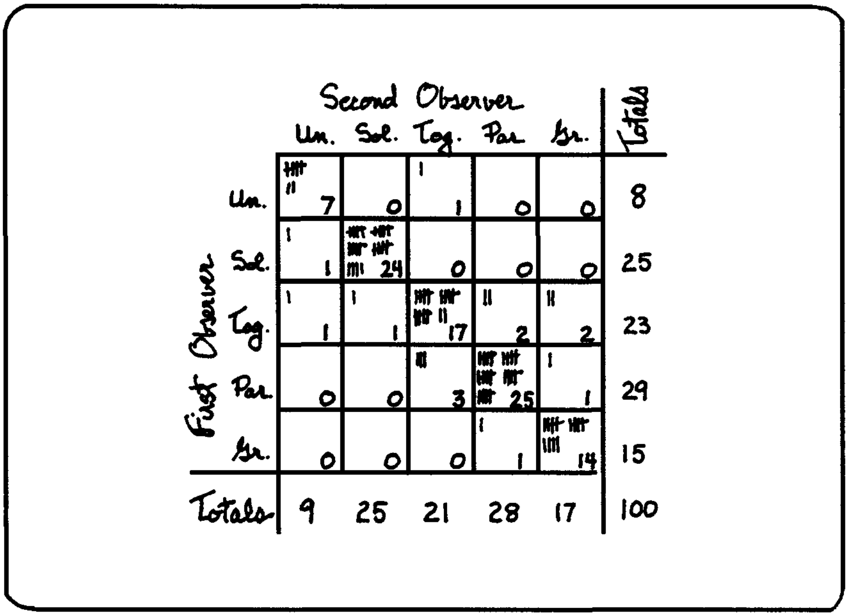
First, what constitutes an agreement needs to be defined. For example, we might say that two observers agree if they record the same kind of distress call at times that either overlap or are separated by no more than two seconds. They disagree when one records a distress call and the other does not, or when they agree that a distress call occurred but disagree as to what kind it is. (Following an old classification system for sins, some writers call these disagreements "omission errors" and "commission errors," respectively.)

Once agreements and disagreements have been identified and tallied, the percentage of agreement can be computed. For example, if two observers both recorded eight Phioos at essentially the same time, but disagreed three times (each observer recorded one Phioo that the other did not, and once one observer recorded a Phioo that the other called a Soft Peep), the percentage of agreement would be 73 (8 divided by 8 + 3 times 100). Percentage agreement could also be reported, not just for Phioos in particular, but for embryonic distress calls in general. For example, if the two observers agreed as to type of distress call 35 times but disagreed 8 times, then the percentage of agreement would be 81 (35 divided by 35 + 8 times 100).

Given a reasonable definition for agreement and for disagreement, the percentage of agreement is easy enough to compute. However, it is not at all clear what the number means. It is commonly thought that agreement percentages are "good" if they are in the 90s, but there is no rational basis for this belief. The problem is that too many factors can affect the percentage of agreement – including the number of codes in the code catalog – so that comparability across studies is lost. One person's 91% can be someone else's 78%.

Perhaps the most telling argument against agreement percentage scores is this: Given a particular coding scheme and a particular recording strategy, some agreement would occur just by chance alone, even with blindfolded observers, and agreement percentage scores do not correct for this. This becomes most clear when an interval coding strategy is coupled with a simple mutually exclusive and exhaustive scheme, as in the study of parallel play described in section 1.7. Recall that for this study, Bakeman and Brownlee had observers code each successive 15-second interval as either Unoccupied, Solitary, Together, Parallel, or Group. If two observers had each coded the same 100 intervals, the pattern of agreement might have been as depicted in Figure 4.1. In this case, the percentage of agreement would be 87 (87 divided by 87 + 13 times 100). However, as we show in the next section, an agreement of 22.5% would be expected, in this case, just by chance alone. The problem with agreement percentages is that they do not take into account the part of the observed agreement that is due just to chance.

Figure 4.1 is sometimes called a "confusion matrix," and it is useful for monitoring areas of disagreement that are systematic or unsystematic. After computing the frequencies of entries in the confusion matrix, the reliability checker should scan for clusters off the diagonal. These indicate



First Observer	Second Observer: Un.	Sol.	Tog.	Par	Gr.	Totals
Un.	7	0	1	0	0	8
Sol.	1	24	0	0	0	25
Tog.	1	1	17	2	2	23
Par.	0	0	3	25	1	29
Gr.	0	0	0	1	14	15
Totals	9	25	21	28	17	100

Figure 4.1. An agreement or "confusion" matrix. Tallies on the diagonal indicate agreement between the two observers, whereas tallies off the diagonal pinpoint disagreements.

confusions between specific codes. If many observers display the same confusion, it may suggest retraining, or clarification of the coding manual, or finding clear examples that sharpen distinctions between codes.

4.4 The advantages of Cohen's kappa

An agreement statistic that does correct for chance is Cohen's kappa (Cohen, 1960). As a result, it is almost always preferable to simple agreement percentages. It is defined as follows:

$$\kappa = \frac{P_{obs} - P_{exp}}{1 - P_{exp}}$$

where P_{obs} is the proportion of agreement actually observed and P_{exp} is the proportion expected by chance. P_{obs} is computed by summing up the tallies representing agreement (those on the upper-left, lower-right diagonal in Figure 4.1) and dividing by the total number of tallies. That is, it is analogous to P_A in the previous section, except that it is not multiplied by

100. Symbolically:

$$P_{obs} = \frac{\sum_{i=1}^{k} x_{ii}}{N}$$

where k is the number of codes (i.e., the order of the agreement matrix), x_{ii} is the number of tallies for the ith row and column (i.e., the diagonal cells), and N is the total number of tallies for the matrix. For the agreement portrayed in Figure 4.1, this is:

$$P_{obs} = \frac{7 + 24 + 17 + 25 + 14}{100} = .87$$

P_{exp} is computed by summing up the chance agreement probabilities for each category. For example, given the data in Figure 4.1, the probability that an interval would be coded Unoccupied was .08 for the first observer and .09 for the second. From basic probability theory, the probability of two events occurring jointly (in this case, both observers coding an interval Unoccupied), just due to chance, is the product of their simple probabilities. Thus the probability that both observers would code an interval Unoccupied just by chance is .0072 (.08 × .09). Similarly, the chance probability that both would code an interval Solitary is .0625 (.25 × .25), Together is .0483 (.21 × .23), Parallel is .0812 (.28 × .29), and Group is .0255 (.17 × .15). Summing the chance probabilities for each category gives the overall proportion of agreement expected by chance (P_{exp}), which in this case is .2247.

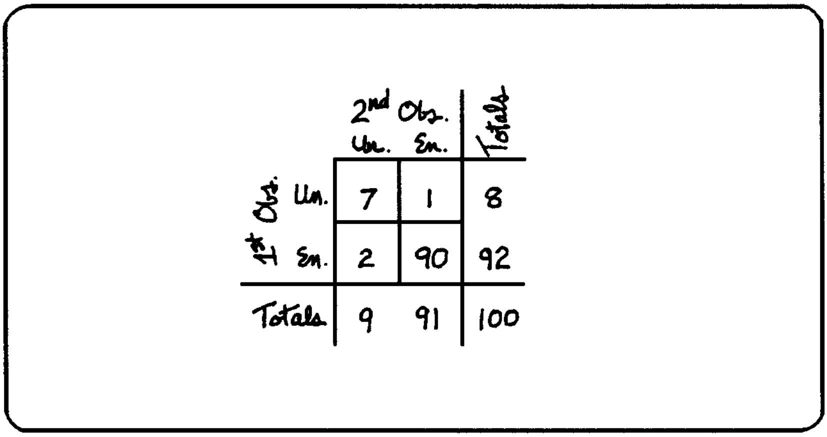
A bit of algebraic manipulation suggests a somewhat simpler way to compute P_{exp}. Multiply the first column by the first row total, add this to the second column total multiplied by the second row total, etc., and then divide the resulting sum of the column-row products by the total number of tallies squared. Symbolically:

$$P_{exp} = \frac{\sum_{i=1}^{k} x_{+i} x_{i+}}{N^2}$$

where x_{+i} and x_{i+} are the sums for the ith column and row, respectively (thus one row by column sum cross-product is computed for each diagonal cell).

For the agreement given in Figure 4.1, this is:

$$P_{exp} = \frac{9 \times 8 + 25 \times 25 + 21 \times 23 + 28 \times 29 + 17 \times 15}{100 \times 100}$$
$$= .2247$$



		2nd Obs.		
		Un.	En.	Totals
1st Obs.	Un.	7	1	8
	En.	2	90	92
Totals		9	91	100

Figure 4.2. An agreement matrix using a coding scheme with two codes. Although there is 97% agreement, 84% would be expected just by chance alone.

Now we can compute kappa for our example data.

$$\kappa = \frac{.87 - .2247}{1. - .2247} = .8323 \text{ (rounded)}$$

As the reader can see, the amount of agreement corrected for chance (about .83) is rather less than the uncorrected value (.87). In some cases, especially when there are few coding categories and when the frequency with which those codes occur is quite disproportionate, the difference can be quite dramatic. Imagine, for example, that instead of the five categories listed in Figure 4.1, only two had been used: Unengaged (meaning Unoccupied) and Engaged. In that case,the data from Figure 4.1 could reduce to the data shown in Figure 4.2. The proportion of agreement oberved here is quite high, .97 (7 + 90 divided by 100), but so is the proportion of chance agreement as well.

$$P_{exp} = \frac{9 \times 8 + 91 \times 92}{100 \times 100} = .8444$$

As a result, the value of kappa, although still respectable, is considerably lower than the level of agreement implied (misleadingly) by the .97 value:

$$\kappa = \frac{.97 - .8444}{1. - .8444} = .8072$$

The question now is, is a kappa of .8072 big enough? Fleiss, Cohen, and Everitt (1969) have described the sampling distribution of kappa, and so it is possible to determine if any given value of kappa differs significantly from zero (see also Hubert, 1977). The way it works is as follows: First, the population variance for kappa, assuming that kappa is zero, is estimated from the sample data. Then the value of kappa estimated from the sample data is divided by the square root of the estimated variance and the result compared to the normal distribution. If the result were 2.58 or bigger, for example, we would claim that kappa differed significantly from zero at the .01 level or better.

In this paragraph, we show how to compute the estimated variance for kappa, first defining the procedure generally and then illustrating it, using the data from Figure 4.2. The formula incorporates the number of tallies (N, in this case 100), the probability of chance agreement (P_{exp} in this case .8444), and the row and column marginals: p_{i+} is the probability that a tally will fall in the ith *row*, whereas p_{+j} is the probability that a tally will fall in the jth *column*. In the present case, $p_{1+} = .08$, $p_{2+} = .92$, $p_{+1} = .09$, and $p_{+2} = .91$. To estimate the variance of kappa, first compute

$$\sum_{i=1}^{k} = p_{i+} \times p_{+i} \times [1 - (p_{+i} + p_{i+})]^2$$

In the present case, this is:

$$\begin{aligned} &.08 \times .09 \times [1 - (.09 + .08)]^2 + .92 \times .91 \times [1 - (.91 + .92)]^2 \\ &\quad = .00496 + .57675 = .5817 \end{aligned}$$

Then add to it this sum:

$$\sum_{i=1}^{k} \sum_{\substack{j=1 \\ i \neq j}}^{k} = p_{i+} \times p_{+j} \times (p_{+i} + p_{j+})^2$$

In the present case, this is:

$$\begin{aligned} &.08 \times .91 \times (.09 + .92)^2 + .92 \times .09 \times (.91 + .08)^2 \\ &\quad = .07426 + .08115 = .1554 \end{aligned}$$

Next subtract P_{exp}^2. In the present case this is:

$$.8444^2 = .7130$$

and the result is:

$$.5817 + .1554 - .7130 = .0241$$

Finally, divide this result by $N \times (1 - P_{exp})^2$. This divisor is:

$$100 \times (1 - .8444)^2 = 2.421$$

and the final quotient is:

$$.0241/2.421 = .009955$$

This is the estimated variance for kappa, given the data in Figure 4.2. The z score is 8.091, the estimated kappa (.8072) divided by the square root of the variance (.09977). We would conclude that the agreement demonstrated in Figure 4.2 is significantly better than chance.

For many investigators, this will not be stringent enough. Just as correlation coefficients that account for little variance in absolute terms are often significant, so too, quite low values of kappa often turn out to be significant. This means only that the pattern of agreement observed is greater than would be expected if the observers were guessing and not looking. This can be unsatisfactory, however. We want from our observers not just better than chance agreement; we want *good* agreement. Our own inclination, based on using kappa with a number of different coding schemes, is to regard kappas less than .7, even when significant, with some concern, but this is only an informal rule of thumb. Fleiss (1981), for example, characterizes kappas of .40 to .60 as fair, .60 to .75 as good, and over .75 as excellent.

The computation of kappa can be refined in at least three ways. The first is fairly technical. Multiple observers may be used – not just two – and investigators may want a generalized method for computing kappa across the different pairs. In such cases, readers should consult Uebersax (1982); a BASIC program that computes Uebersax's generalized kappa coefficient has been written by Oud and Sattler (1984).

The second refinement is often useful, especially when codes are roughly ordinal. Investigators may regard some disagreements (confusing Unoccupied with Group Play, for example) as more serious than others (confusing Together with Parallel Play, for example). Cohen (1968) has specified a way of weighting different disagreements differently. Three $k \times k$ matrices are involved: one for observed frequencies, one for expected frequencies, and one for weights. Let x_{ij}, m_{ij}, and w_{ij} represent elements from these three matrices, respectively: then $m_{ij} = (x_{+j} \times x_{i+}) \div N$, and the w_{ij} indicate how seriously we choose to regard various disagreements. Usually the diagonal elements of the weight matrix are 0, indicating agreement (i.e., $w_{ii} = 0$ for $i = 1$ through k); cells just off the diagonal are 1, indicating some disagreement; cells farther off the diagonal are 2, indicating more serious disagreement; etc. For the present example, we might enter 4 in cells x_{15} and x_{51}, indicating that confusions between *unoccupied* and *group* are given more weight than other disagreements. Then weighted kappa is computed as follows:

$$\kappa_{wt} = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} = w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}}$$

If $w_{ii} = 0$ for $i = 1$ through k and 1 otherwise, then κ, as defined earlier, and κ_{wt} are identical. (To compute variance for weighted kappa, see Fleiss, Cohen, & Everitt, 1969).

A third way the computation of kappa can be refined involves P_{exp}. As usually presented, P_{exp} is computed from the row and column totals (the marginals), but in some cases investigators may believe they have better ways to estimate the expected distribution of the codes, and could use them to generate expected frequencies.

For example, imagine that an investigator has accumulated a considerable data archive using the coding scheme presented in Figure 4.1 and so has a good idea of how often intervals are coded Unoccupied, etc., in general. Imagine further that coders are subject to random agreement checks and that, as luck would have it, a tape selected for one such check shows an unusual child who spends almost all of his time Unoccupied with just a little additional Solitary play. In this case, the kappa would be quite low. What is really a 5-category scheme becomes for kappa computation a 2-category scheme with a skewed distribution, which usually results in low values. In effect, the observers have received no "credit" for knowing that Together, Parallel, and Group Play did not occur. In such a case, it may make sense to substitute the average values usually experienced for the marginals rather than use the actual ones from such a deviant case.

As the above example makes clear yet again, assessing observer agreement has more than one function. When our concern is to convince others (especially journal editors and reviewers) that our observers are accurate, then we might well pool tallies from several different agreement checks into one kappa table, computing and reporting a single kappa value. This has the merit of providing realistic marginals. When our concern is to calibrate and train observers, however, we would want to compute kappas separately for each agreement check, thus providing immediate feedback. At the same time, we should counsel our observers not to be discouraged when low kappas result simply from coding an unusual instance.

With respect to the training of observers, there is one final advantage of Cohen's kappa that we should mention. The kappa table itself provides a graphic display of disagreement. Casual inspection immediately reveals which codes are often confused and which almost never are. An excessive number of tallies in an off-diagonal cell would let us know, for example, that intervals one observer codes together are regarded as parallel by the other observer. Moreover, simple inspection can also reveal if one observer is more "sensitive" than the other. If so, a pattern of considerably more tallies above than below the diagonal (or vice versa) results. For example, if most of the disagreements in Figure 4.1 had been above the diagonal (meaning that what the first observer regarded as Unoccupied the second observer often regarded as something more engaged but not vice versa), this would indicate that the second observer was more sensitive, detecting engagement when the first observer saw only an unengaged child. Such patterns have clear implications for observer training. In the first case (code confusion), further training would attempt to establish a consensual definition for the two codes; and in the second (different sensitivities), further training would attempt to establish consensual thresholds for all codes.

4.5 Agreement about unitizing

Constructing a kappa table like that shown in Figure 4.1 is easy enough. For each "unit" coded, a tally is entered in the table. The codes assigned the unit by the two observers determine in which cell the tally is placed. This procedure, however, assumes previous agreement as to what constitutes a unit. Sometimes unit boundaries are clear-cut, and may even be determined by forces external to the observers. At other times, boundaries are not all that distinct. In fact, determining unit boundaries may be part of the work observers are asked to do. For example, observers may be asked to identify and code relatively homogeneous stretches of talk, or to identify a particular kind of episode (e.g., a negotiation or conflict episode).

In such cases, agreement needs to be demonstrated on two levels: first with respect to "unitizing" (i.e., identifying the homogeneous stretches or episodes), and second with respect to the codes assigned the homogeneous stretch or episode (or perhaps events embedded within the episode itself).

When the unit being coded is a time interval, as in the example just given, there is no problem. In that case, unit boundaries are determined by a clock and not by observers. When the unit being coded is an event, however, the matter becomes more difficult. For example, consider Gottman's study of friendship formation described in section 2.11. In such cases, there are two parts to the coding task. First, observers need to segment the stream of recorded talk into thought units, whereas second, they need to code the segmented thought units themselves.

Agreement with respect to the coding of thought units can be determined using Cohen's kappa, as described in the last section. However, how should agreement with respect to segmenting the stream of talk into thought units be demonstrated? In the older literature, a percentage score has often been used for this task.

If both observers worked from transcripts, marking thought unit boundaries on them, it should be a fairly simple matter to tally the boundaries claimed by both observers and the ones noted only by one observer. But in this case, there can be only omission disagreements. Moreover the percentage agreement would not correct for chance; still the score may have some descriptive value, albeit limited.

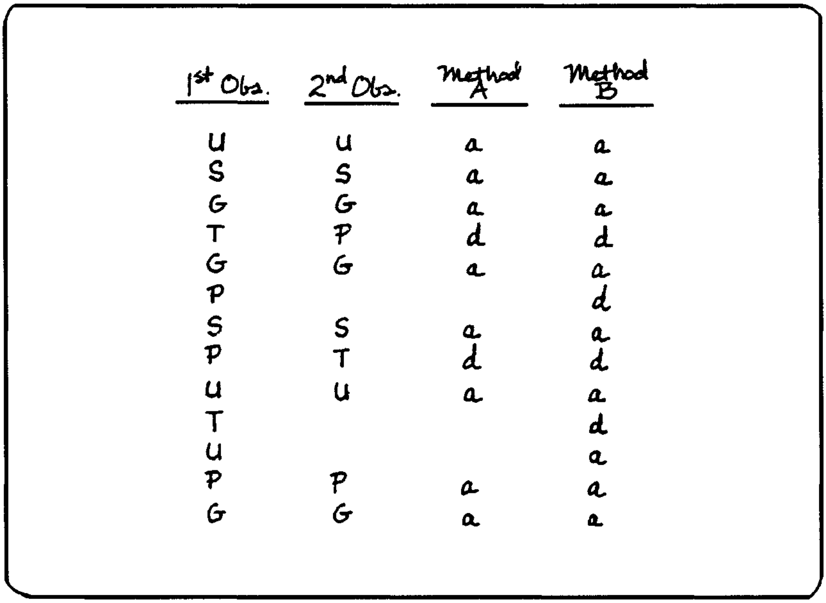
Alternatively, the investigator might regard every gap between adjacent words as a potential boundary. In this case, the initial coding unit would be the word gap. Each gap would contribute one tally to a 2 × 2 kappa table: (a) Both observers agree that this gap is a thought unit boundary; (b) the first observer thinks it is, but the second does not; (c) the second thinks it is, but the first disagrees; or (d) both agree that it is not a thought unit boundary. Although the chance agreement would likely be high (two coding categories, skewed distribution, assuming that most gaps would not be boundaries), still the kappa computed would correct for that level of chance agreement.

A better and more general procedure for determining agreement with respect to unitizing (i.e., identifying homogeneous stretches of talk or particular kinds of episodes) requires that onset and offset times for the episodes be available. Then the tallying unit for the kappa table becomes the unit used for recording time. For example, imagine that times are recorded to the nearest second and that observers are asked to identify conflict episodes, recording their onset and offset times. Then kappa is computed on the basis of a simple 2 × 2 table like that shown in Figure 4.2, except that now rows and columns are labeled yes/no, indicating whether or not a second was coded for conflict. One further refinement is possible. When tallying seconds, we could place a tally in the agreement (i.e., yes/yes) cell if one observer claimed conflict for the second and the other observer claimed conflict either for that second or an adjacent one, thereby counting 1-second disagreement as agreements as often seems reasonable.

In the previous chapter, we described five general strategies for recording observational data: (a) coding events, (b) timing onsets and offsets, (c) timing pattern changes, (d) coding intervals, and (e) cross-classifying events. Now we would like to mention each in turn, describing the particular problems each strategy presents for determining agreement about unitizing.

Coding events, without any time information, is in some ways the most problematic. If observers work from transcripts, marking event (thought unit) boundaries, then the procedures outlined in the preceding paragraphs can be applied. If observers note only the sequence of events, which means that the recorded data consist of a string of numbers or symbols, each representing a particular event or behavioral state, then determining agreement as to unit boundaries is more difficult. The two protocols would need to be aligned, which is relatively easy when agreement is high, and much more difficult when it is not, and which requires some judgment in any case. An example is presented in Figure 4.3.

When onset and offset or pattern-change times are recorded, however, the matter is easier. Imagine, for example, that times are recorded to the nearest second. Then the second can be the unit used for computing agreement both for unitizing (identifying homogeneous stretches or episodes) and for the individual codes themselves. Because second boundaries are determined



1st Obs.	2nd Obs.	Method A	Method B
U	U	a	a
S	S	a	a
G	G	a	a
T	P	d	d
G	G	a	a
P			d
S	S	a	a
P	T	d	d
U	U	a	a
T			d
U			a
P	P	a	a
G	G	a	a

Figure 4.3. Two methods for determining agreements ("a") and disagreements ("d") when two observers have independently coded the same sequence of events. Method A ignores errors of omission. Method B counts both errors of commission and errors of omission as disagreements.

by a clock external to the observers, there is no disagreement as to where these boundaries fall (the only practical requirement is that the clocks used by the two observers during an agreement check be synchronized in some way). An example showing how second-by-second agreement would be computed in such cases is given in the next section.

When time intervals are coded in the first place, the matter is similar. Again, the underlying unitization is done by clocks, not by observers. When cross-classifying events, however, we need to ask, to what extent are both observers detecting the same events? In older literature, often a percentage agreement was used. For example, in their study of social rules among preschool children, Bakeman and Brownlee (1982) asked two observers to cross-classify object struggles (see section 2.13). During an agreement check, one observer recorded 50 such struggles, the other 44; however, all 44 of the latter had also been noted by the first observer, and hence their percentage agreement was 88.0% (44 divided by 44 + 6). In a case like this, there seems to be no obvious way to correct for chance agreement.

Our recommendation is as follows: Report the agreement-disagreement tallies along with the percentage of agreement in such cases, but note their limitations. However, if at all possible, report time-based kappa statistics to establish that observers detected the same events to cross-classify. This is the same strategy we recommend to demonstrate that observers are identifying the same homogeneous stretches or episodes, which likewise are then subjected to further coding.

4.6 Agreement about codes: Examples using Cohen's kappa

In this section, we describe how Cohen's kappa can be used to document observer agreement about codes for each of the recording strategies listed in chapter 3. This is not the only way to determine observer agreement (or reliability), as we discuss in the next section, but it may be among the most stringent. This is because Cohen's kappa documents point-by-point agreement, whereas many writers would argue that agreement does not need to be determined for a level more detailed than that ultimately used for analysis. We think that this argument has merit, but that there are at least two reasons to favor a relatively stringent statistic like Cohen's kappa. First, as we argued earlier in this chapter, determining observer agreement has more than one function. For training observers and providing them feedback on their performance, we favor an approach that demands point-by-point agreement. We also like the graphic information about disagreement provided by the kappa table. Second, once agreement at a detailed level has been established, we can safely assume agreement at less detailed levels, and in any case a relatively detailed level is required for sequential analysis.

When observers code events, it is relatively straightforward to compute agreement about how units are coded, once the units are identified. For example, when thought units are coded from transcripts, the codes the two observers assign each successive thought unit would determine the cell of a 26 × 26 kappa table in which a tally would be placed (assuming 26 possible codes for thought units). What this example highlights is that kappa is a summary statistic, describing agreement with respect to how a coding scheme is used (not agreement about particular codes in the scheme), and that the codes that define each kappa table must be mutually exclusive and exhaustive.

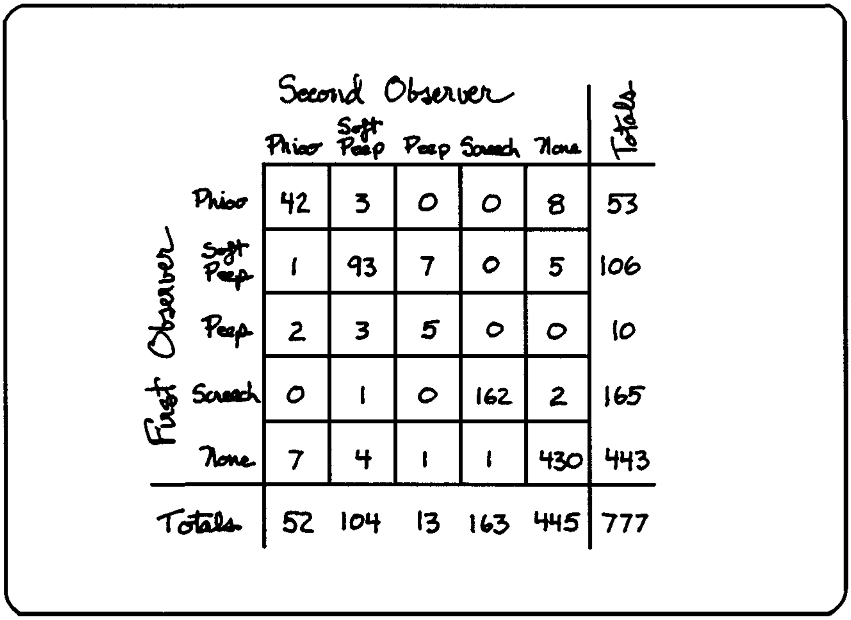
As a second example, still assuming that events are being coded, imagine that an investigator wants to segment the stream of behavior into the five play states used by Bakeman and Brownlee (Unoccupied, Solitary, Together, Parallel, Group). If observers are told where the segment boundaries are, their only task would be to code segments. In this case, it would be easy to construct a kappa table. Telling observers where the segment boundaries are, however, is not an easy matter. Other observers, working with videotaped material, would need to have determined those boundaries previously, and then the boundaries would need to be marked in some way, perhaps with a brief tone dubbed on the soundtrack.

If observers are not told where the segment boundaries are, however, but instead are asked to both segment and code at the same time, then protocols like those shown in Figure 4.3 would result. The question now is, how do we construct a kappa table from data like these? There are two choices. We could ignore those parts of the protocols where observers did not agree as to segment boundaries and tally only those segments whose boundaries were agreed upon, in effect ignoring errors of omission. This would result in eight agreements and two disagreements, as shown in Figure 4.3 (Method A). Or we could assume that there "really" was a segment boundary whenever one of the observers said there was. This would result in nine agreements and four disagreements (Method B). Neither of these choices seems completely satisfactory. The first probably overestimates agreement, whereas the second probably underestimates it. Our preference is for computing kappas using a time interval as the unit, but this requires timing onsets and offsets, timing pattern changes, or coding intervals directly.

As an example, consider the recording of onset and offset times for different kinds of embryonic distress calls done by Tuculescu and Griswold (section 2.10). Assume that these times were recorded to the nearest second. Then each second can be categorized as "containing" (a) a Phioo, (b) a Soft Peep, (c) a Peep, (d) a Screech, or (e) no distress call. If two observers both code the same audiotape, agreement data could be like that shown in Figure 4.4. Similarly, in addition to the kappa for embryonic distress calls, kappas could be computed for Tuculescu and Griswold's other superordinate categories as well (embryonic pleasure calls, maternal body movements, maternal head movements, and maternal vocalizations).

Occasionally, editors or colleagues ask for agreement statistics, not for a coding scheme, as kappa provides, but for individual codes. We usually think that kappa coupled with the agreement matrix is sufficient, but nonetheless kappas can be computed for individual codes by collapsing the table appropriately. Consider the agreement matrix for the five distress calls shown in Figure 4.4. From it we could derive five 2 × 2 matrices, first collapsing the tallies into Phioo/not Phioo, then Soft Peep/not Soft Peep, etc. Then a kappa could be computed separately for each table. In this case, the individual kappas would be .79, .87, .43, .98, and .93 for Phioo, Soft Peep, Peep, Screech, and None, respectively. Not surprisingly, the kappa associated with Peep is relatively low; of the 10 noted by the first observer and the 13 noted by the second observer, agreement occurred for only 5.

A second example of a time-based kappa is provided by Adamson and Bakeman (1985), who asked observers to record whenever infants displayed heightened affectivity. These displays were typically quite brief,

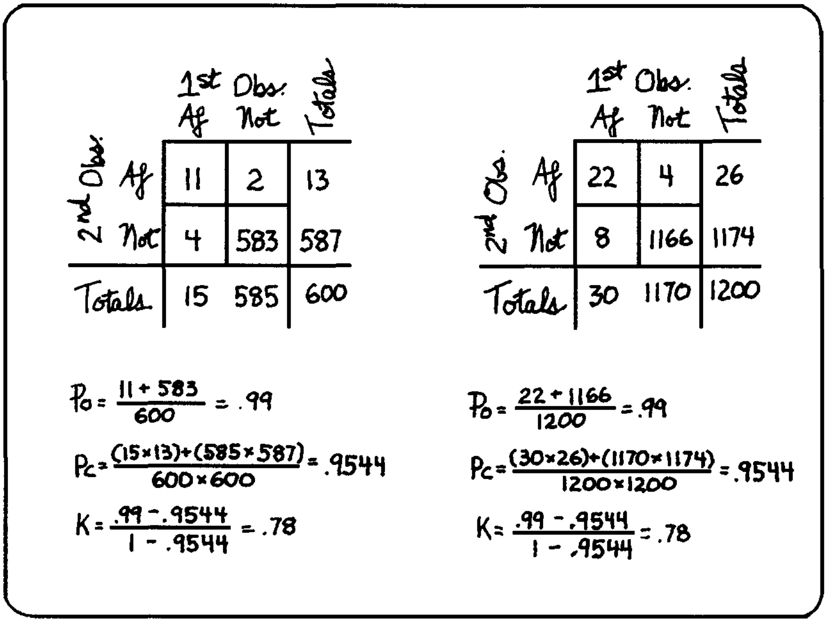


		Second Observer					
		Phioo	Soft Peep	Peep	Screech	None	Totals
First Observer	Phioo	42	3	0	0	8	53
	Soft Peep	1	93	7	0	5	106
	Peep	2	3	5	0	0	10
	Screech	0	1	0	162	2	165
	None	7	4	1	1	430	443
Totals		52	104	13	163	445	777

Figure 4.4. An agreement matrix for coding of chicken embryonic distress calls. Each tally represents a 1-second interval. The reader may want to verify that the percentage of agreement observed is 94.2%, the percentage expected by chance is 39.4%, the value of kappa is .904, its standard error is .0231, and the *z* score comparing kappa to its standard error is 39.2.

lasting just a few seconds, relatively infrequent, and consisted of such things as smiles, gleeful vocalizations, or excited arm waving. Assuming a unit of 1 second, a 10-minute agreement check could produce a kappa table like the one given on the left in Figure 4.5. What would happen, however, if a half-second unit had been used instead? The answer is essentially nothing, as is demonstrated by the table on the right in Figure 4.5. By and large, halving the length of the unit would result in tables that are roughly proportional, except that one would have twice as many tallies as the other. The kappa statistic (unlike chi-square) is not affected by this, however, as the kappa computations in Figure 4.5 demonstrate.

In the preceding two paragraphs (and in Figures 4.4 and 4.5), we have suggested how kappa can be computed when onset times for mutually exclusive and exhaustive codes are recorded. The principle is exactly the same when timing pattern changes. Again, for each set of mutually exclusive and exhaustive codes, a kappa table can be constructed, tallying agreement and disagreement for each second (or whatever time unit is used) coded. When timing pattern changes, codes are always parceled into



		1st Obs. Af	1st Obs. Not	Totals
2nd Obs.	Af	11	2	13
	Not	4	583	587
Totals		15	585	600

$$P_o = \frac{11 + 583}{600} = .99$$

$$P_c = \frac{(15 \times 13) + (585 \times 587)}{600 \times 600} = .9544$$

$$K = \frac{.99 - .9544}{1 - .9544} = .78$$

		1st Obs. Af	1st Obs. Not	Totals
2nd Obs.	Af	22	4	26
	Not	8	1166	1174
Totals		30	1170	1200

$$P_o = \frac{22 + 1166}{1200} = .99$$

$$P_c = \frac{(30 \times 26) + (1170 \times 1174)}{1200 \times 1200} = .9544$$

$$K = \frac{.99 - .9544}{1 - .9544} = .78$$

Figure 4.5. When time intervals are tallied, and a reasonable time interval is used, the value of kappa is not changed by halving the time interval (which doubles the tallies).

mutually exclusive and exhaustive sets. In other cases, constructing a set of mutually exclusive and exhaustive codes is no problem, as the examples presented in Figures 4.4 and 4.5 demonstrate. Adding the "none of the above" code to a code for affective display or to codes for distress calls makes the set mutually exclusive and exhaustive.

When the remaining two recording strategies – coding intervals and cross-classifying events – are used, computing agreement using kappa is straightforward. An example of what a kappa table might look like when intervals are coded was given earlier (see Figure 4.1). When events are cross-classified, the only difference is that there is one kappa table for each classification scheme (or dimension) and that events not detected by both observers cannot be entered into the table. For example, in their study of social rules, Bakeman and Brownlee (1982) reported kappas for each of their three dimensions: (a) prior possession, (b) resistance, and (c) success (see section 2.13). Kappa, however, is not the only agreement statistic there is. As we discuss in the next section, other statistics and approaches to observer agreement have their advantages.

4.7 Generalizability theory

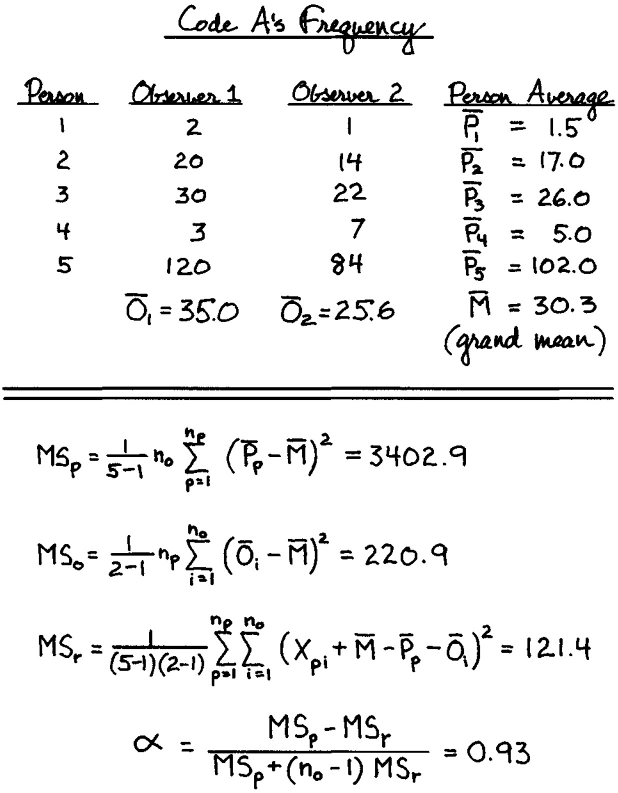
Cronbach, Gleser, Nanda, and Rajaratnam (1972; see also Brennan, 1983) presented what amounts to a conceptual breakthrough in thinking about both reliability and validity. To understand their notions, let us introduce the concept of the "work we want the measure to do." For example, we would like to be able to use our observations of the amount of negative-affect reciprocity to discriminate satisfied from dissatisfied marriages. Or, we might want our measure of the amount of negative affect to predict the husband's health in three years. This is the work our measure is to do. It is designed to discriminate or to predict something of interest.

Cronbach and colleagues' major point is that this work is always relative to our desire in measurement to generalize across some facet of our experimental design that we consider irrelevant to this work. For example, our test scores should generalize across items within a measurement domain. It should not matter much if we correlate math achievement and grade point average (GPA) using even or odd math achievement items to compute the correlation coefficient. We are generalizing across the irrelevant facet of odd/even items. The work the measure does is to discriminate high- from low-GPA students.

In a similar way, if we have a measure of negative affect, we expect it to discriminate among happily and unhappily married couples, and not to discriminate among coders (the irrelevant facet). We wish to generalize across coders. Let us briefly discuss the computations involved in this analysis. Figure 4.6 presents the results of one possible generalizability study. For five persons, each of two observers computed the frequency of code A for a randomly selected segment of videotape. The setup is the same as a simple repeated-measures experiment. The within-subject factor is observer (with two levels, i.e., data from two observers) and there is no between-subject factor as such; subjects represent total between-subject variability. The analysis of variance source table for the data shown in Figure 4.6, expanded to include R^2 as recommended by Bakeman (1992), is shown in Table 4.1. Given these data and the current question, an appropriate coefficient of generalizability, or reliability, is:

$$\alpha = \frac{MS_p - MS_r}{MS_p + (n_o - 1)MS_r} \tag{4.1}$$

where n_o is the number of observers (2 in this case) and MS_p and MS_r are the mean squares for persons and residual (or error), respectively (the first edition of this book omitted $n_o - 1$ in the denominator because it equaled 1, but this proved confusing). This is an intraclass correlation coefficient based on the classical assumption that observed scores can be divided into a true and an error component ($X = T + e$), so that the appropriate intraclass



Code A's Frequency

Person	Observer 1	Observer 2	Person Average
1	2	1	$\bar{P}_1 = 1.5$
2	20	14	$\bar{P}_2 = 17.0$
3	30	22	$\bar{P}_3 = 26.0$
4	3	7	$\bar{P}_4 = 5.0$
5	120	84	$\bar{P}_5 = 102.0$
	$\bar{O}_1 = 35.0$	$\bar{O}_2 = 25.6$	$\bar{M} = 30.3$ (grand mean)

$$MS_p = \frac{1}{5-1} n_o \sum_{p=1}^{n_p} (\bar{P}_p - \bar{M})^2 = 3402.9$$

$$MS_o = \frac{1}{2-1} n_p \sum_{i=1}^{n_o} (\bar{O}_i - \bar{M})^2 = 220.9$$

$$MS_r = \frac{1}{(5-1)(2-1)} \sum_{p=1}^{n_p} \sum_{i=1}^{n_o} (X_{pi} + \bar{M} - \bar{P}_p - \bar{O}_i)^2 = 121.4$$

$$\alpha = \frac{MS_p - MS_r}{MS_p + (n_o - 1)\, MS_r} = 0.93$$

Figure 4.6. A generalizability approach to observer agreement: n_o is the number of observers or 2, n_p is the number of persons or 5, MS_p is the mean square for persons, MS_o is the mean square for observers, and MS_r is the mean square for residual or, in this case, the $P \times O$ interaction; see also Table 4.1. The formulas for MS_p and MS_o were incorrect in the first edition of this book; they are also given incorrectly in Wiggins (1973, p. 289).

correlation is defined as

$$\alpha = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}. \tag{4.2}$$

(Equation 4.1 is derived from 4.2 by substitution and algebraic manipulation.) This statistic estimates the reliability of observations made by a randomly selected observer, selected from the pool that contained the two observers used here for the reliability study (Table 4.1), and further

Table 4.1. *Source Table for the Generalizability Study of Observer Reliability*

Source	R^2	ΔR^2	*SS*	*df*	*MS*
Person	0.95	0.95	13,611.6	4	3,402.9
Observer	0.97	0.02	220.9	1	220.9
P × O	1.00	0.03	485.6	4	121.4
Total			14,318.1	9	

assumes that data will be interpreted within what Suen (1988) terms a norm-referenced (i.e., values are meaningful only relatively; rank-order statistics like correlation coefficients are emphasized) as opposed to a criterion-referenced framework (i.e., interpretation of values references an absolute external standard; statistics like unstandardized regression coefficients are emphasized). Equation 4.1 is based on recommendations made by Hartmann (1982) and Wiggins (1973, p. 290). For other possible intraclass correlation coefficients (generalizability coefficients), based on other assumptions, see Fleiss (1986, chapter 1) and Suen (1988), although, as a practical matter, values may not be greatly different. For example, values for a criterion-referenced fixed-effect and random-effect model per Fleiss (1986) were .926 and .921, respectively, compared to the .931 of Figure 4.6. In contrast, assuming that observers were item scores and we wished to know the reliability of total scores based on these items, Cronbach's internal-consistency alpha (which is $MS_p - MS_r$ divided by MS_p; see Wiggins, 1973, p. 291) was .964.

This way of thinking has profound consequences. It means that reliability can be high even if interobserver agreement is moderate, or even low. How can this be? Suppose that for person #5 in Figure 4.6, Observer 1 detected code A 120 times, as shown but only 30 of these overlapped in time with Observer 2's 84 entries. Then the interobserver agreement would be only $30/120 = .25$. Nonetheless, the generalizability coefficient of equation 4.1 is .93. The reliability is high because either observer's data distinguishes equally well between persons. The agreement within person need not be high. The measure does the work it was intended to do, and either observer's data will do this work. This is an entirely different notion of reliability than the one we have been discussing.

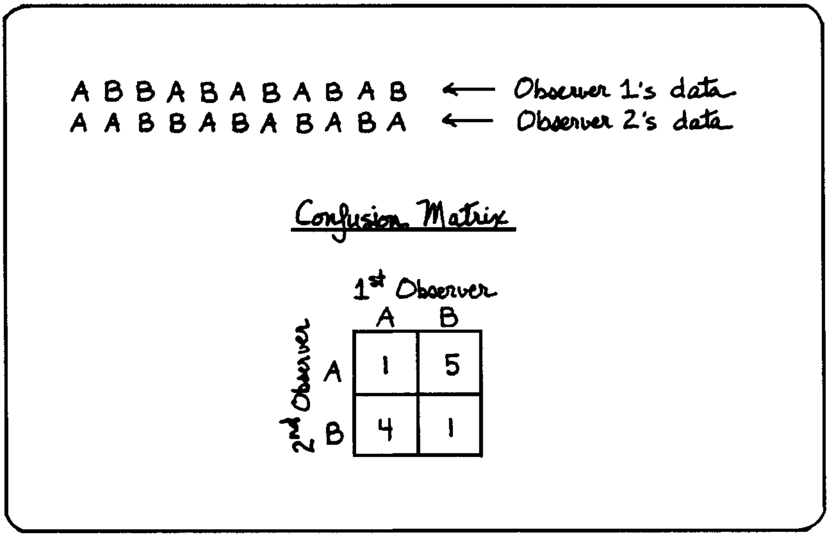
Note that the generalizability or reliability coefficient estimated by equation 4.1 is a specific measure of the relative variance accounted for by an interesting facet of the design (subjects) compared to an uninteresting one (coders). This is an explicit and specific proportion, but it does not tell us how large a number is acceptable, any more than does the proportion of variance accounted for in a dependent variable by an independent variable. The judgement must be made by the investigator. It is not automatic, just as a judgment of an adequate size for kappa is not an automatic procedure.

Note also that what makes the reliability high in the table in Figure 4.6 is having a wide range of people in the data, ranging widely with respect to code A. Jones, Reid, and Patterson (1975) presented the first application of Cronbach and colleagues' (1972) theory of measurement to observational data.

The reliability analysis just presented, although appropriate when scores are interval-scaled (e.g., number of events coded A by an observer), is inadequate for sequential analysis. The agreement required for sequential analysis cannot be collapsed over time, but must match point for point, as exemplified by the kappa tables presented in the previous section. Such matching is much more consistent with "classical" notions of reliability, i.e., before Cronbach et al. (1972).

Still, agreement point-for-point could be assessed in the same manner as in Figure 4.6. The two columns in Figure 4.6 would be replaced by sums from the confusion matrix. Specifically, the sums on the diagonal would replace Observer 1's scores, and the sums of diagonal plus off-diagonal cells (i.e., the row marginals) would replace Observer 2's scores. If the agreement point-for-point were perfect, all entries for code A in the confusion matrix would be on the diagonal and the two column entries would be the same. There would then be no variation across "observers," and alpha would be high. In this case, the "persons" of Figure 4.6 become codes, "Observer 1" becomes agreements and "Observer 2" becomes agreements plus disagreements.

This criterion is certainly sufficient for sequential analysis. However, it is quite stringent. Gottman (1980a) proposed the following: If independent observers produce similar indexes of sequential connection between codes in the generalizability sense, then reliability is established. For example, if two observers produced the data in Figure 4.7 (designed so that they are off by one time unit, but see the same sequences), their interobserver agreement would be low but indexes of sequential connection would be very similar across observers. Some investigators handle this simple problem by having a larger time window within which to calculate the entries in the confusion matrix. However, that is not a general solution because more complex configurations than that of Figure 4.7 are possible, in which both observers detect similar sequential structure in the codes but point-for-point agreement is low. Cronbach et al.'s (1972) theory implies that all we need to demonstrate is that observers are essentially interchangeable in doing the work that our measures need to do.



A B B A B A B A B A B ← Observer 1's data
A A B B A B A B A B A ← Observer 2's data

Confusion Matrix

2nd Observer	1st Observer A	1st Observer B
A	1	5
B	4	1

Figure 4.7. A confusion matrix when observers "see" the same sequence but one observer lags the other. In such cases, a "point-for-point" agreement approach may be too stringent.

4.8 Unreliability as a research variable

Raush (personal communication, 1974) once suggested that one source of unreliability is to be found in the nature of the social interaction itself. He referred to a message called a "probe" that a person might send to the receiver. The probe is designed to be taken one of two ways, depending on the state of the receiver. For example, in a potentially sexual interaction a sender may send a probe with a subtle sexual invitation. If it is ignored, the sender gains information that directs the interaction one way; if it is responded to, the interaction may proceed in another direction. Krokoff (1983) recently tested this notion in marital interaction. He reasoned that such probe messages would be more common during high-conflict interaction because of the delicate nature of the subject matter and the great danger that the conflict would escalate. If this were true, Krokoff reasoned, then reliability would be significantly reduced for those videotapes high in negative affect. This hypothesis was strongly supported by the data.

Patterson's (1982, p. 50) book quoted Reid as noting that "observer agreement is largely a function of the complexity of the interaction. By selecting *simple* interaction segments, one may obtain *very* high observer agreement." When complexity was defined as the number of different codes entered divided by the total entries for 5 minutes, this hypothesis was strongly supported; reliability was lower for more complex segments. This could partly be due to the increased task demands on the coder, but it could also be partly a property of the interaction itself, if in more complex interactions people sent more probe messages. The point of this section is to suggest that reliability can itself become a research variable of interest.

4.9 Summary

There are at least three major reasons for examining agreement among observers. The first is to assure ourselves and others that our observers are accurate and that our procedures are replicable. The second is to calibrate multiple observers with each other or with some assumed standard. This is important when the coding task is too large for one observer or when it requires more than a few weeks to complete. The third reason is to provide feedback when observers are being trained.

Depending on which reason is paramount, computation of agreement statistics may proceed in different ways. One general guiding principle is that agreement need be demonstrated only at the level of whatever scores are finally analyzed. Thus if conditional probabilities are analyzed, it is sufficient to show that data derived from two observers independently coding the same stream of behavior yielded similar conditional probabilities. Such an approach may not be adequate, however, when training of observers is the primary consideration. Then, point-by-point agreement may be demanded. Point-by-point agreement is also necessary when data derived from different observers making multiple coding passes through a videotape are to be merged later.

When an investigator has sequential concerns in mind, then, point-by-point agreement is necessary for observer training and is required for at least some uses of the data. Moreover, if point-by-point agreement is established, it can generally be assumed that scores derived from the raw sequential data (like conditional probabilities) will also agree. If agreement at a lower level is demonstrated, agreement at a higher level can be assumed. For these reasons, we have stressed Cohen's kappa in this chapter because it is a statistic that can be used to demonstrate point-by-point agreement. A Pascal program that computes kappa and weighted kappa is given in the Appendix. Kappa is also computed by Bakeman and Quera's Generalized Sequential Querier or GSEQ program (Bakeman & Quera, 1995a). At the same time, we have also mentioned other approaches to observer reliability (in section 4.7). This hardly exhausts what is a complex and far-ranging topic. Interested readers may want to consult, among others, Hollenbeck (1978) and Hartmann (1977, 1982).